Weakly Supervised Roadside Object Geolocation from Street-Level Imagery with Similarity Learning and Graph-Based Reasoning

Waqar Ahmad, B.Eng., M.Sc.

A dissertation submitted in fulfilment of the requirements for the award of Doctor of Philosophy (PhD)

to the



Dublin City University

School of Mathematical Sciences

Supervised by

Dr. Vladimir A.Krylov

August 2026

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy (PhD) is entirely my own work, and that I have exercised reasonable care to ensure that the work is original and have conformed to the regulations on the use and declaration of Generative AI, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Waqar Ahmad

ID No.: 22268538

Date: 07-August-2026

Appendix A - Reporting of Gen AI Tool Use in Research Thesis

Tools	Google Gemini, 2.0/2.5, used between February, 2025 and March, 2026.
Outputs	I have used Google Gemini to summarize key research papers related to my thesis topic. In particular, I used it to extract the main findings and provide an overall summary of each paper. I also used Gemini as a debugging aid for my Python code that I had written. The tool suggested possible fixes for errors encountered during execution, recommended clearer and easier-to-reproduce coding styles, and helped correct syntax-related issues. I treated these outputs as suggestions and reviewed them before incorporating any changes into my work.
Prompts	I asked Gemini to: "Identify the possible reasons for errors in my Python code and to suggest corrections without changing the original functionality of the code." "Summarize the work presented in the provided paper and to list the key findings." I also instructed the tool not to add unsupported suggestions, external interpretations, or information beyond the material provided, and to keep its response strictly aligned with the given task.
Iterations	I kept the number of iterations minimal, usually limited to two per task. I carefully reviewed each output before use and manually addressed any issues while considering the tool's suggestions. Suggestions that diverged from the specific task, altered the intended meaning, or went beyond the provided information were disregarded. Only relevant and understandable suggestions consistent with my own judgment were considered.
Critical Reflection	The use of the Gen AI tool in this project was minimal. It was helpful in identifying a few errors in basic Python code; however, when it suggested changes related to logical errors, I disregarded them if they appeared to change the intended functionality of the code. I used the tool only as limited support for coding assistance and research paper summarization. In many cases, particularly after repeated iterations, the responses diverged from the original request or provided suggestions that were not directly relevant. Therefore, the tool was not used for advanced algorithm development, methodological design, or any novel contributions presented in this thesis.

Acknowledgements

"In the name of ALLAH, the Most Gracious, the Most Merciful."

In the name of Allah, the Most Compassionate, the Most Merciful. All praise and gratitude belong to Allah Almighty, whose guidance, mercy, and blessings gave me the strength, patience, and perseverance to complete this doctoral journey. As stated in the Holy Qur'an, "My Lord, increase me in knowledge" (Surah Taha, Chapter 20, Verse 114). I remain deeply grateful to Allah for enabling me to pursue knowledge and complete this research.

I would like to express my deepest gratitude to my supervisor, Dr Vladimir Krylov, for his invaluable guidance, continuous encouragement, and unwavering support throughout my doctoral journey. His expertise, thoughtful feedback, patience, and commitment to my development as a researcher have played a fundamental role in shaping this thesis. I am sincerely grateful for the time and confidence he invested in me, particularly during the most demanding and challenging stages of this work. His mentorship has contributed not only to the completion of this research but also to my growth as an independent researcher. I also remember with profound gratitude my late co-supervisor, Dr Kevin McGuinness, whose knowledge, kindness, and contributions to my research remain deeply valued. I feel privileged to have benefited from his guidance, and his memory will continue to be a source of inspiration. I am also sincerely grateful to my external examiner, Professor Csaba Benedek, and internal examiner, Assist. Prof Marco Viola, for their time, careful review, and valuable assessment of this thesis, and to the Independent Chair, Assoc. Prof Eabhnat Ní Fhloinn, for overseeing the viva examination and ensuring its

smooth conduct.

I also extend my sincere gratitude to the ADAPT research centre for providing a supportive and stimulating research environment. I gratefully acknowledge the financial support provided by the SFI Research Centres Programme, co-funded by the European Regional Development Fund under Grant No. 13/RC/2106_P2.

I owe an immeasurable debt of gratitude to my parents and siblings for their unconditional love, continuous encouragement, sacrifices, and prayers. Their belief in me remained a constant source of strength and motivation throughout this journey, even when we were separated by distance. Their prayers sustained me during difficult moments, and their happiness gave meaning to every achievement. Whatever I have accomplished would not have been possible without the values, resilience, determination, and faith they instilled in me.

My heartfelt thanks go to my colleagues and friends for their valuable discussions, encouragement, and support throughout this journey. I am especially grateful to my coffee- and tea-time companions, whose humour, kindness, and friendship made the challenging moments lighter and more memorable.

Finally, and most importantly, I would like to express my deepest love and gratitude to my beloved wife. She endured the difficulties caused by my long absences and the considerable demands of my research, yet she remained patient, strong, understanding, and endlessly supportive. During the most challenging periods of this journey, she stood beside me with unwavering faith and encouragement, often carrying burdens and making sacrifices without complaint. Her love, resilience, patience, and constant support gave me the strength to continue when the path seemed most difficult. Words cannot fully express my gratitude for everything she has endured and contributed throughout this journey. This achievement belongs as much to her as it does to me.

To everyone who supported me, encouraged me, prayed for me, shared their knowledge, offered a kind word, or simply accompanied me during any part of this journey, I offer my most sincere and heartfelt thanks.

Contents

List of Figures **X**

List of Tables **XIV**

List of Abbreviations **XVII**

Abstract **XIX**

1 Introduction to Geolocation **1**

1.1 Problem Statement 2

1.2 Research Gaps and Contributions 4

1.3 Thesis Structure 5

2 Background and Related Work **8**

2.1 Historical Evolution of Geolocation 8

2.2 Fundamentals of Geographic Coordinate Systems 10

2.3 Categories of Geolocation 11

2.4 Object Detection and Scene Understanding in Images 13

2.5 Multi-view Geometry and Triangulation 18

2.6 Similarity Learning and Metric Learning 22

2.7 Graph Neural Networks for Structured Data 26

2.8 Datasets for Street-Level Object Geolocation 28

3 Roadside object geolocation from street-level images with reduced supervision **31**

3.1 Introduction . . . 32
3.2 Proposed geolocation pipeline . . . 35
3.3 Experiments . . . 38
3.3.1 Datasets and setup . . . 38
3.3.2 Traffic lights detection . . . 39
3.3.3 Public bins detection and geolocation . . . 41
3.4 Conclusion . . . 44
3.5 Appendix 3A: Supplementary Results . . . 46
3.5.1 Result Analysis . . . 46
3.5.2 Visual Results . . . 48

4 Similarity-based Outlier Detection for Noisy Object Re-Identification Using Beta Mixtures 51

4.1 Introduction . . . 52
4.2 Related work in object Re-ID . . . 55
4.3 Outlier Detection (OD) . . . 58
4.4 Beta mixtures identifiability . . . 64
4.5 Siamese architecture and Loss Functions . . . 66
4.6 Experiments . . . 68
4.6.1 Experimental Setup . . . 68
4.6.2 Results & Comparisons . . . 71
4.7 Conclusion . . . 81
4.8 Appendix 4A: Beta-SOD Supplementary Material . . . 82
4.9 Appendix 4B: Alternative Filtering . . . 84
4.9.1 Result Analysis . . . 87
4.10 Appendix 4C: Extension to Road-side Object Geolocation Pipeline . . . 90

5 Similarity-Augmented Roadside Object Triangulation from street level images 92

5.1 Introduction . . . 93

5.2 Methodology 94
5.2.1 Information fusion with MRF 97
5.3 Datasets and Experimentation 100
5.3.1 Dataset Description 100
5.3.2 Training Setup 100
5.3.3 Results Analysis 101
5.4 Conclusion 104

6 Edge-Aware Graph Neural Network for Roadside Object Geolocation 106

6.1 Introduction 107
6.2 Proposed Method 110
6.2.1 Triangulation problem 110
6.2.2 Graph Construction and GNN 113
6.3 Experimental Setup 116
6.3.1 Datasets 116
6.3.2 Experimental Setup 117
6.3.3 Model Training 119
6.3.4 Model Testing 119
6.4 Results Analysis 120
6.4.1 Noise setting for training 120
6.4.2 Benchmark comparisons 120
6.4.3 Computation Complexity 123
6.4.4 Ablation Study 123
6.5 Conclusion 124

7 Discussion 125

7.1 Conclusion 125
7.2 Limitations 127
7.3 Future Work 129

References **133**

List of Figures

3.1 Proposed geolocation pipeline. Step A: Image classification by Resnet34. Step B: The Grad-CAMs generation of the classified Image. Step C: Calculation of local maxima using Watershed. Step D: Depth estimation using pre-trained DL algorithm. Step E: Triangulation using Markov Random Fields. 36

3.2 Grad-CAMs of traffic light dataset. Row 1: Resnet34 results. Row 2: Resnet50 results. Column (a): Input Image, (b): True Traffic Light Position (c): No label noise, (d)-(f): 1%, 5%, and 10% label noise respectively. Black dots are the local maxima identified using watershed transform. 37

3.3 Segmentation from Resnet34 GradCAMs (blue ellipses), and Mask-RCNN (green rectangles). (1): Image, (2-3) Segmentations obtained by using: no label noise and 10% label noise datasets, respectively. . . 42

3.4 Grad-CAMs and local maxima identified via watershed (black dot) on public bins. Images from Mapillary (1-2), and the DCC dataset (3-4). 42

3.5 Public bins in central Dublin: DCC public records (green circles) vs. MRF (red squares) vs. manually identified ground truth (blue diamonds). 44

3.6 Example sequences of cropped detections from multi-view images, where each row corresponds to a single object detection across different views. False positive detections are highlighted in red, illustrating the presence of noisy and inconsistent observations within the sequences. 45

3.7 Multi-view image sequences used as input to the proposed model, showing six synchronized views of a single location captured by six cameras. Each row corresponds to a fixed camera viewpoint, and each column represents a consecutive frame. 49

3.8 Qualitative comparison under increasing label noise (0%–10%). Columns (a)–(e) show ground truth, no noise, and 1%, 5%, and 10% noise, respectively; rows correspond to ResNet34, ResNet50, and Mask R-CNN. ResNet34 preserves accurate localization, while Mask R-CNN produces false positives (red arrows) and ResNet50 misses detections (red circle) at higher noise levels. 50

4.1 Proposed Beta-SOD Architecture: (A) The Siamese architecture extracts features from paired images. (B) Feature vectors are flattened and passed to a fully connected neural network (FCNN) for pair classification. (C) Cosine similarity is computed from the extracted features and processed by the OD module after every K epochs. The identified outliers are are then filtered from the dataset. (D) The Siamese network and FCNN are trained jointly using the combination of three loss functions. 56

4.2 Histogram of cosine similarities for both similarity classes on the CUHK03 dataset with overlaid estimated Beta, Gaussian, and Gamma two-component mixtures: with (a) 0%, (b) 10% and (c) 20% random label noise. Unlike Gaussian and Gamma, the Beta distribution models the data in the target range $[0, 1]$. 64

4.3 Examples of incorrectly filtered image pairs by Beta-SOD on CUHK03 with 20% noise: false positives, i.e. true class 1 flagged as 0 (top row), and false negatives, i.e. true class 0 flagged as 1 (bottom row). The numbers are cosine similarities: below 0.8 for false positives and above 0.6 for false negatives. 79

4.4 Effect of the weighting parameter λ used to balance the multi-loss objective on validation accuracy, mAP, and Rank-1 accuracy on the CUHK03 dataset under 20% random label noise. The best overall performance is observed at $\lambda = 0.45$. 81

4.5 Examples of random noise in CUHK03 dataset. Row (a) shows dissimilar pairs (true Class 0) mislabeled as Class 1, while row (b) shows similar pairs (true Class 1) incorrectly assigned to Class 0. 83

4.6 Examples of Pattern noise in CUHK03 dataset. Row (a): dissimilar pairs with high similarity labeled as Class 1; row (b): similar pairs with low similarity labeled as Class 0. Similarity scores are shown for each pair. 83

4.7 Evolution of similarity distributions during filtering. Rows correspond to noise levels (10%, 20%, 30%), while columns represent successive filtering iterations. 84

5.1 Proposed training pipeline: (A) weakly supervised object classification and Grad-CAM–based localization, (B)MRF-based triangulation to generate multi-view image pairs, and (C) Siamese similarity learning with Beta-mixture filtering. 95

5.2 Inference pipeline: object detection and localization, triangulation-based view pairing, Siamese similarity estimation, and final geolocation using a similarity-augmented MRF. 98

5.3 Filtered image pairs after Beta-distribution-based denoising for Dublin (top) and Amsterdam (bottom) datasets 103

6.1 Overview of the EdgeAwareVH model. The input includes view, hypothesis and edge features. These inputs form a bipartite graph connecting views to their corresponding triangulated hypotheses. An encoder processes the node features, followed by stacked EdgeAwareVH convolution blocks (EdgeAwareConv + LayerNorm + ReLU). A final MLP head outputs a score for each hypothesis, allowing the model to select the most viable object locations. 111
6.2 GPU memory usage: linear with respect to the number of hypotheses for EdgeAwareVH, and quadratic for Set Transformer (Lee et al. 2019). 122

List of Tables

3.1 Classification performance on Traffic Lights Dataset 40

3.2 Instance-based IOU score for traffic lights dataset 41

3.3 Instance Based IOU score for the public bins Dataset 43

3.4 Classification performance under varying retrieval distances (d) and segmentation area thresholds (A_{max}). The model is trained and evaluated on the Mapillary dataset. 46

3.5 Instance Based IOU Score under varying retrieval distances (d) and segmentation area thresholds (A_{max}). The model is trained and evaluated on the Mapillary dataset. 47

3.6 Classification Performance under varying retrieval distances (d) and segmentation area thresholds (A_{max}). The model is trained on the Mapillary dataset and evaluated on the DCC dataset. 48

4.1 Siamese Backbone Network selection on CUHK03 Dataset with various random noise levels (5 runs average) 72

4.2 Noise identification performance with various random noise levels (5 runs average) 72

4.3 Noise identification performance of OD using Gaussian, Gamma, and Beta mixtures under varying noise levels on the CUHK03 dataset (5-run average) 73

4.4 Comparison of Person Re-ID performance of Beta-SOD against Noise-Aware Re-Id (Song et al. 2024), DistributionNet (Yu et al. 2019), PurifyNet (Ye and Yuen 2020), CORE (Ye et al. 2022a), URNet (Li et al. 2019), and TSNT (Liu et al. 2024) on the Market1501 and CUHK03 datasets with various random noise levels (5 runs Average and Standard Deviation) 75

4.5 Comparison of Person Re-ID performance of Beta-SOD against PurifyNet (Ye and Yuen 2020), CORE (Ye et al. 2022a), and TSNT (Liu et al. 2024) on the Market1501 and CUHK03 datasets under various **pattern noise** levels (5-run average and standard deviation) 76

4.6 Person Re-ID performance of the OD method using Gaussian, Gamma, and Beta mixture models on the CUHK03 dataset under varying noise levels (5-run average) 76

4.7 Comparison of Vehicle Re-ID performance on the VeRi-776 dataset with various random noise levels (5 runs Average) 80

4.8 Loss Function Selection on CUHK03 Dataset with random noise level of 0% and 20% (5 runs average) 80

4.9 Comparison of iterative filtering and Beta-SOD in terms of outlier removal, precision, and recall across different datasets and various noise levels. 89

5.1 Classification results on the Amsterdam dataset 102

5.2 Classification results with Resnet-50 architecture 102

5.3 Geolocation performance on the Amsterdam and Dublin datasets against MRF (Krylov et al. 2018) 103

5.4 Ablation study of the proposed similarity-augmented MRF, evaluating the impact of visual similarity and the proposed statistical noise filtering on geolocation performance. 105

6.1 Layer configuration of the EdgeAwareVH model, showing node encoding ($N = N_v + N_h$), geometric message passing, and final scoring stage. 116
6.2 EdgeAwareVH performance with different noise configurations on real D_{Real} and noisy simulated $D_{Sim+Noise}$ test sets. 121
6.3 Comparative performance of EdgeAwareVH against MRF (Krylov et al. 2018) and Set Transformer (Lee et al. 2019), evaluated on both real and simulated test sets. 121
6.4 Performance of Proposed method against Set Transformer (Lee et al. 2019) on D_{Real} at varying detection distance thresholds (in meters) . 122
6.5 Ablation comparisons for EdgeAwareVH on D_{real}. The evaluated variants include Graph Attention Networks (GAT) (Veličković et al. 2018), Edge-Conditioned Convolutions (ECC) (Simonovsky and Komodakis 2017), and the proposed EdgeAwareVH, each integrated with OSM-based priors 123

Acronyms

3D 3-Dimensional. 11, 15–19, 91, 105, XV

BCE Binary Cross-Entropy. 50, 54, 65, 80, XV

Beta-SOD Beta Mixture-based Similarity Outlier Detection. 6, 50, 51, 53, 69, 70, 75, 76, 79, 80, 82, 85, 86, 88, 89, 99, VI, XV

CAM Class Activation Map. 14, XV

CNN Convolutional Neural Network. 13, 52, 56, 61, 88, 94, 97, 105, 118, XV

DCC Dublin City Council. 37, 40, 42, 45–47, 98, XV

ECEF Earth-Centered, Earth-Fixed. 10, 110, XV

EM Expectation-Maximization. 53, 59–62, 69, XV

ENU East-North-Up. 10, 11, 108, 110, 111, 117, XV

GAT Graph Attention Network. 27, 106, XV

GCN Graph Convolutional Network. 27, 106, XV

GNN Graph Neural Network. 6, 26–28, 104, 106–108, 111, 114, 118, 122, VII, XV

GNSS Global Navigation Satellite System. 9, XV

GPS Global Positioning System. 9–12, 17, 18, 32, 42, 95, 97, XV

Grad-CAM Gradient-weighted Class Activation Mapping. 4, 6, 14, 30, 33, 34, 36, 38–41, 43, IX, XV

MOT Multiple Object Tracking. 22, 23, XV

MRF Markov Random Field. 6, 20–22, 26, 31, 34, 36, 42, 43, 90, 92–96, 98, 100–103, 105, 108, 118, 119, VII, XI, XV

OD outlier detection. 50, 53, 54, 61, 66, 69, 70, 72, 74–76, 79, 80, 82–84, XV

OSM OpenStreetMap. 7, 9, 98, 108, 111, 115, 119, 121, 122, XV

Re-ID Re-identification. 4, 6, 25, 50–54, 56–58, 61, 65, 67, 68, 70, 72, 74–76, 78–80, 88, VI, XV

SfM Structure from Motion. 11, 19, XV

UTM Universal Transverse Mercator. 10, 11, XV

Abstract

Waqar Ahmad

Weakly Supervised Roadside Object Geolocation from Street-Level Imagery with Similarity Learning and Graph-Based Reasoning

Mapping roadside infrastructure from street-level imagery is important for urban asset management and smart-city applications. Large collections of street-level images enable automated infrastructure monitoring. However, estimating the real-world locations of roadside objects remains challenging due to weak supervision, noisy detections, and uncertain geometric measurements across viewpoints. In practice, infrastructure records are often incomplete or outdated. Obtaining detailed annotations for large image datasets is costly and difficult to scale. This thesis investigates methods for detecting and geolocating roadside objects from street-level imagery under weak supervision. The aim is to enable reliable object localization under noisy and uncertain conditions. The proposed framework integrates visual localization, similarity learning, and geometric reasoning to address these challenges. Roadside objects are first localized in images using weak supervision. Multi-view geometric cues are then used to estimate their real-world locations from multiple camera observations. However, weak supervision and triangulation can introduce noisy observations and incorrect associations across views. Similarity learning is applied to estimate visual similarity between image pairs. A statistical outlier detection framework identifies and filters unreliable training pairs. This improves the reliability of the resulting geolocation estimates. To move beyond purely statistical inference, a learning-based graph model is introduced. This model improves the robustness and scalability of the geolocation process. The model represents relationships between observations and candidate object locations using a graph neural network. It learns to identify consistent multi-view evidence while suppressing

spurious geometric intersections from noisy measurements. Extensive experimental evaluation demonstrates that the proposed methods achieve competitive and robust roadside object geolocation, even when observations are noisy and geometric measurements are uncertain. Overall, this thesis develops and validates methods for detecting and geolocating roadside infrastructure from street-level imagery. This contributes to more accurate and reliable automated mapping of urban assets.

Chapter 1

Introduction to Geolocation

Accurate and up-to-date information about roadside infrastructure is essential for modern urban management (Chen et al. 2019; Loce et al. 2013). City authorities and infrastructure providers rely on detailed inventories of roadside assets, such as traffic signs, lighting poles, waste bins, and utility equipment, to support maintenance operations, urban planning, and public services. Traditionally, these inventories are created and maintained through manual surveys or dedicated mapping campaigns, which are often expensive, time-consuming, and difficult to update at a large scale. With the increasing availability of street-level imagery from platforms such as Google Street View (gsv), Microsoft Bing, Mapillary (Neuhold et al. 2017), and municipal imaging systems, there is growing interest in automatically extracting infrastructure information directly from images.

Computer vision and deep learning methods have significantly advanced the ability to detect and recognize objects in images (Ren et al. 2017; Redmon et al. 2016). However, detecting objects in images alone is not sufficient for many practical applications. For infrastructure management tasks, the geographic location of objects in the real world is often more important than their presence in individual images (Workman et al. 2015; Vo et al. 2017). Consequently, the problem extends beyond object detection and requires estimating the real-world geolocation of objects using multiple observations from different viewpoints (Hartley and Zisserman 2004).

Object geolocation from street-level images involves estimating the real-world position of a roadside object using visual observations captured from one or more viewpoints. Since the images are taken from different camera positions and viewing directions, the same object may appear in several images with different sizes, viewpoints, lighting conditions, or partial occlusions. Therefore, the camera location, viewing direction, and detected object are used together to estimate the object's actual location. In practice, this process can also be affected by noisy measurements, missed detections, incorrect detections, and nearby objects that look similar (Hartley and Zisserman 2004).

Recent work has used a combination of computer vision, multi-view geometry, and spatial reasoning for object geolocation (Snavely et al. 2006; Schönberger and Frahm 2016). In these methods, the same object is first detected across different images, and each observation is projected into the real world. The viewing rays from different images are then intersected or triangulated to form possible object locations (Hartley and Zisserman 2004). These possible locations are further checked using geometric consistency, probabilistic models, or learning-based methods to select the most likely object position (Krylov et al. 2018). In this way, roadside object geolocation can be seen as a process of detecting objects in images, forming possible locations from multiple views, and then selecting the final position from those locations.

1.1 Problem Statement

Estimating real-world object locations from street-level images introduces several challenges. In typical geolocation pipelines for roadside assets from street-level imagery, object detection is the initial step, identifying candidate objects within the image. A key challenge arises from the limited availability and reliability of object-level annotations in large-scale street-level imagery datasets.

Modern object detection models generally require a large amount of accurately labeled training data, where objects are annotated using high-level supervision such

as bounding boxes, object labels, and segmentation masks (Redmon et al. 2016; He et al. 2017). Producing such detailed annotations for street-level imagery, however, is both expensive and time-consuming. Consequently, large crowdsourced platforms such as Mapillary often rely on community-contributed or semi-automatically generated annotations. While these datasets provide valuable large-scale coverage, the annotations may contain inconsistencies, missing labels, or inaccurate object boundaries, which can introduce additional difficulties when training robust detection models for downstream geolocation tasks (Neuhold et al. 2017; Cordts et al. 2015).

Beyond detection and annotation challenges, estimating object locations requires reasoning across multiple views (Hartley and Zisserman 2004). Roadside objects are typically observed from different camera positions, and their appearance may vary due to changes in viewpoint, illumination, occlusion, and image quality. Accurate geolocation, therefore, depends on modeling the geometric relationship between camera positions, viewing directions, and object locations. Even small errors in camera metadata, object detection, or bearing estimation can propagate through the geometric estimation process, leading to significant inaccuracies in estimated object positions.

Another difficulty arises from noisy and incomplete observations in real-world datasets. Object detections may be missing, incorrectly localized, or associated with the wrong object. When triangulation or ray-intersection methods are used to estimate candidate object locations from multiple views, noisy measurements can generate a large number of spurious intersections (Chaabane et al. 2021). Identifying the correct object locations among these candidate intersections is particularly challenging in dense urban environments, where multiple objects may lie close together.

In addition to these challenges, existing methods often rely on either purely geometric approaches or purely learning-based approaches. Geometric methods can exploit multi-view relationships but are sensitive to noise and measurement errors.

On the other hand, learning-based approaches can model complex patterns in data but often require large amounts of accurate supervision and may not explicitly incorporate geometric constraints.

1.2 Research Gaps and Contributions

This thesis addresses the problem of automated detection and geolocation of roadside objects from street-level imagery under weak supervision. The key research gaps and the corresponding contributions are summarized as follows:

- The limited availability and reliability of object-level annotations in large-scale street-level imagery present a significant challenge for training accurate object detection models. To address this, a weakly supervised framework is proposed that leverages a deep learning-based classification model to extract discriminative features from image-level annotations. Gradient-weighted Class Activation Mapping (Grad-CAM) and unsupervised segmentation techniques are employed to localize objects within the image. These detections are combined with image metadata, including camera geolocation, object bearing, and estimated depth, within a statistical multi-view formulation to estimate real-world object positions.

- Weak supervision often leads to false positive detections, which can affect multi-view triangulation by introducing incorrect intersections and inconsistent view associations. This highlights the need to incorporate visual features as an additional source of information within the statistical triangulation framework. To address this issue, we introduce visual similarity between observations as a complementary cue to geometric reasoning. To achieve this, the problem is reformulated as an object Re-identification (Re-ID) task across multiple views. A similarity-based learning framework is introduced, employing a Siamese network to identify inconsistent views via learned similarity while explicitly accounting for noise. A probabilistic modeling strategy is further employed to

distinguish reliable pairs from noisy ones, enabling robust cross-view matching and reducing the impact of incorrect view associations.

- Building on this, the Re-ID framework and statistical noise filtering strategy are applied to multi-view street-level imagery, where faulty detections across views are systematically identified and filtered. The Siamese network, trained under noisy conditions, is further refined on multi-view street-level imagery to capture cross-view visual consistency, and similarity scores are computed across multiple observations. These learned visual similarities, together with image metadata such as camera location, object bearing, and estimated depth, are then integrated into a statistical triangulation framework. By combining visual and geometric information, the approach effectively suppresses unreliable observations, reduces spurious intersections, and improves the robustness and accuracy of object geolocation.

- Geometric triangulation alone is highly sensitive to noise and ambiguity in multi-view observations, limiting its reliability in real-world scenarios. To address this, a graph-based learning framework is proposed that models the relationships between camera views and candidate object locations. By leveraging graph neural networks, the approach reasons over multi-view observations and selects the most plausible object locations, effectively combining geometric constraints with learning-based inference for improved geolocation performance.

1.3 Thesis Structure

This thesis is structured to provide a clear methodological foundation and a consistent presentation of the individual research contributions. The background material first introduces the key concepts, methods, and literature that support the work developed in the thesis. Building on this foundation, each research chapter is presented as a self-contained contribution, beginning with an introduction and task-specific related work, followed by the proposed methodology, the datasets used for evaluation,

and the experimental setup. The results are presented and analyzed in depth, highlighting the approach's strengths and areas for further improvement. Finally, each research chapter concludes with a summary of findings, ensuring a consistent and comprehensive presentation across all contributions. The thesis ends with a concluding chapter that summarizes the overall findings, discusses the main limitations, and outlines possible directions for future work. A brief overview of the content covered in each chapter is provided below.

- **Chapter 2:** This chapter provides the theoretical and methodological background for the thesis. It reviews the main concepts, existing methods, and relevant literature that support the research contributions presented in the following chapters.

- **Chapter 3:** This chapter presents a low-supervision framework for detecting and geolocating roadside objects from street-level images using only binary image labels and existing geo-records. The method combines image classification, Grad-CAM-based localization, and Markov Random Field (MRF)-based spatial optimization, and is evaluated on traffic lights and public bins. Additional details related to this work, including preliminary experiments, extended results, and further observations, are provided in Appendix 3A 3.5.

- **Chapter 4:** This chapter presents a Siamese network-based Re-ID framework that learns reliable image similarity despite noisy labels. It introduces a Beta Mixture-based Similarity Outlier Detection (Beta-SOD) method for identifying noisy pairs, and evaluates its effectiveness on person and vehicle Re-ID datasets under different noise levels. Supplementary results for Beta-SOD are provided in Appendix 4A 4.8, preliminary experiments are presented in Appendix 4B 4.9, and its extension to object geolocation is described in Appendix 4C 4.10.

- **Chapter 5:** This chapter presents a similarity-augmented MRF framework for roadside object geolocation under weak supervision, where visual consistency is

combined with geometry-based multi-view triangulation. A Siamese network is used to estimate pairwise visual similarity between views, while a Beta-distribution-based filtering strategy is applied during training to reduce the effect of noisy labels.

- **Chapter 6:** This chapter presents Edge-Aware View-Hypothesis (EdgeAwareVH), a robust Graph Neural Network (GNN)- based framework for roadside object geolocation that learns from the geometry of the intersections of multi-view observations using edge-aware attention and geometric message passing. The method combines depth, bearing, camera pose, and OpenStreetMap (OSM)-based spatial cues, and is evaluated on noisy simulated and real-world waste-bin geolocation data.

- **Chapter 7:** This chapter summarizes the main findings of the thesis and highlights the key contributions established through the proposed methods. It also discusses the main areas for further improvement and outlines future directions for extending and strengthening the presented approaches.

Chapter 2

Background and Related Work

This chapter provides the background and literature review for the research presented in this thesis. It begins by introducing the historical development of geolocation, highlighting how the field has evolved from traditional positioning methods to modern vision-based and learning-based approaches. The chapter then discusses the main categories of geolocation, including image-level, camera-level, object-level, and multi-view geolocation, to establish the broader context of the problem addressed in this work. Finally, it reviews the state-of-the-art methods most relevant to the proposed frameworks, including weakly supervised object localization, similarity learning, multi-view geometric reasoning, and graph-based approaches for object geolocation. Together, these discussions provide the conceptual and technical foundation for the methods developed in the following chapters.

2.1 Historical Evolution of Geolocation

Manual Geolocation: The problem of geolocation has evolved gradually, moving from manual surveying practices to the highly automated, data-driven methods used today. In its early stages, cartography and land surveying depended heavily on geometric triangulation, astronomical observations, and careful manual measurements to determine positions on the Earth's surface. Techniques such as triangulation networks and plane-table mapping played a key role in establishing the mathematical

principles that underpin modern localization, particularly in relating angles, distances, and spatial coordinates (Wolf and Ghilani 2012). Although these methods could achieve high accuracy at the local level, they required significant human effort, were time-consuming to implement, and were not well-suited to large-scale mapping.

Global Positioning System: The introduction of satellite-based positioning systems marked a major turning point in geolocation. With the development of the Global Positioning System (GPS), it became possible to determine location accurately and in real time using signals from a network of satellites. GPS, along with other Global Navigation Satellite System (GNSS), provides near-global coverage with meter-level accuracy and has become a core component of modern navigation, mapping, and geospatial applications (Hofmann-Wellenhof et al. 2008). Despite these advances, satellite-based localization is not without limitations. In dense urban environments, signals are often obstructed or reflected by buildings, leading to multi-path effects and reduced accuracy, particularly in so-called urban canyons.

Visual Geolocation: At the same time, advances in computer vision and the growing availability of large-scale image datasets have led to the emergence of vision-based geolocation as a complementary approach. Early methods focused on matching ground-level images to geo-tagged databases using handcrafted features such as SIFT and SURF (Lowe 2004). More recently, deep learning has significantly improved performance by enabling models to learn robust visual representations for place recognition and image retrieval (Arandjelović et al. 2018). These approaches are particularly useful in scenarios where GPS signals are weak or unavailable, providing an alternative means of directly estimating location from visual data.

Building on these developments, multi-view and crowdsourced street-level mapping systems have further advanced the field. Platforms such as Mapillary (Mapillary 2024) and OSM (OpenStreetMap Contributors 2024) leverage large volumes of user-contributed geo-tagged images to create detailed, continuously updated maps. By combining observations from multiple viewpoints, multi-view geometry enables the triangulation of object locations and improves overall localization accuracy (Hartley

and Zisserman 2004). Together, these approaches have led to modern geolocation systems that integrate geometric reasoning, visual understanding, and large-scale data aggregation in a unified framework.

2.2 Fundamentals of Geographic Coordinate Systems

Accurate geolocation depends on representing positions on the Earth's surface consistently and precisely. The most commonly used global reference system is the World Geodetic System, particularly WGS-84, which expresses locations in terms of latitude, longitude, and altitude. This framework provides a standardized basis for positioning and is used by GPS and most modern mapping systems (Defense Mapping Agency 1987). While latitude and longitude are easy to interpret, they are not always ideal for geometric calculations, as their non-linear nature can make operations such as distance and angle computation less straightforward.

To address this, alternative coordinate systems are often used for spatial reasoning and geometric analysis. The GNSS coordinate system represents positions in a three-dimensional Cartesian space centered at the Earth's core, making it well suited for vector-based computations. For more localized applications, the East-North-Up (ENU) system provides a convenient tangent-plane approximation, which simplifies tasks such as multi-view geometry and triangulation. Similarly, the Universal Transverse Mercator (UTM) system projects the Earth onto a set of local planar grids, allowing distances to be measured in meters with relatively low distortion over smaller regions (Snyder 1987).

Converting between these coordinate systems relies on map projections and geospatial transformations. Because the Earth is curved, any projection onto a flat surface introduces some level of distortion, whether in area, distance, or angular relationships. Different projection methods are therefore chosen depending on which properties need to be preserved for a given application. For example, transfor-

mations from WGS-84 to ENU or UTM are particularly important in applications involving sensor fusion, triangulation, and spatial reasoning. They allow data from different sources, such as images, GPS measurements, and geographic databases, to be combined within a common and consistent computational framework.

2.3 Categories of Geolocation

Image-based Geolocation: Geolocation can be broadly categorized by level of abstraction and the nature of its output. At the most basic level, image-level geolocation estimates the geographic location of an entire image by matching it to a database of geo-tagged images. This problem is often formulated as a large-scale image retrieval task, where learned visual descriptors are used to identify the most likely location (Weyand et al. 2016). A more detailed formulation is camera pose estimation, which aims to determine the camera's position and orientation in 3-Dimensional (3D) space. This is typically addressed using Structure from Motion (SfM) or visual localization techniques that match image features to a known 3D model of the environment (Sattler et al. 2018). Camera pose estimation provides richer geometric information and is fundamental to applications such as augmented reality and robotics.

Object-level Geolocation: Moving beyond camera-level localization, object-level geolocation estimates the real-world positions of specific objects observed in images. A key challenge in object-level geolocation is the repeated appearance of a single object across multiple views. This task often requires combining visual detection with geometric reasoning, such as triangulation from multiple viewpoints. In more complex scenarios, multi-object geolocation aims to simultaneously localize multiple instances of objects, often requiring reasoning about associations between observations across views. This introduces challenges related to data association, noise, and ambiguity, particularly in weakly supervised settings. Object-level geolocation can be further divided into three subcategories, including tracker-based, triangulation-based, and re-identification-based geolocation (Wilson et al. 2023).

Tracker-based Geolocation: Tracker-based geolocation approaches typically begin by detecting object instances across multiple frames, after which an object-tracking mechanism, commonly referred to as a tracklet, is employed to track each object across successive views. By leveraging temporal consistency and motion cues, these methods can improve robustness and accuracy, particularly in dynamic environments. Tracking-based approaches are widely used in autonomous driving and surveillance, where continuous object localization over time is required (Milan et al. 2016). Recent work proposes an end-to-end learning framework that jointly estimates object pose, performs inter-frame association, and applies multi-object tracking to improve the geolocation of static objects from video sequences (Chaabane et al. 2021). Similarly, detection and tracking-based pipelines have been proposed to geo-localize dense, multi-class objects from street imagery, demonstrating robustness to missing observations and real-world conditions (Wilson et al. 2022). However, such approaches rely on dense temporal observations and consistent object tracking across consecutive frames, which may not be available in sparse street-level imagery settings. As a result, their applicability is limited in scenarios with unordered views, missing detections, and weak supervision.

Triangulation-based Geolocation: Triangulation-based methods localize objects by combining detections from multiple views and forming geometric relationships between camera positions and object observations. Using camera GPS and predicted object bearings, viewing rays are constructed and intersected to estimate the object location, often through trigonometric or least-squares formulations. In practice, due to noise and imperfect detections, these methods rely on approximate intersections and robust optimization techniques. Therefore, triangulation is often formulated as a least-squares optimization problem to obtain consistent estimates of object positions from noisy observations (Hartley and Sturm 1997). Furthermore, a range of triangulation-based and geometric computer vision methods have been proposed for object-level geolocation (Hartley and Zisserman 2004).

Re-Identification-based Geolocation: Another possible approach is to re-

identify objects across multiple views before estimating their geographic location. In this direction, recent work proposes an end-to-end learnable framework that jointly performs object detection, cross-view re-identification, and geolocation. The method leverages a graph neural network to model both visual appearance and relative pose across multiple views, enabling robust reasoning under occlusions, viewpoint changes, and visually similar neighboring objects. By integrating these components within a unified framework, the approach can handle an arbitrary number of input views and directly predict object locations (Nassar et al. 2020).

2.4 Object Detection and Scene Understanding in Images

Object Detection: Early approaches to object detection relied on hand-crafted features such as Histogram of Oriented Gradients (HOG) (Dalal and Triggs 2005) and deformable part-based models (DPM) (Felzenszwalb et al. 2010), which demonstrated strong performance on constrained datasets but struggled in complex urban environments. The advent of deep learning, particularly Convolutional Neural Network (CNN), significantly advanced the field. Region-based methods such as R-CNN (Girshick et al. 2014), Fast R-CNN (Girshick 2015), and Faster R-CNN (Ren et al. 2017) introduced a two-stage detection pipeline that first generates region proposals and then classifies them, achieving high accuracy on benchmark datasets.

Subsequent work focused on improving efficiency and scalability through single-stage detectors such as YOLO (Redmon et al. 2016) and SSD (Liu et al. 2016a), which perform detection in a unified framework without explicit region proposal steps. The YOLO family has continued to evolve, with later versions such as YOLOv7 (Wang et al. 2023), and the recently proposed YOLOv10 (Wang et al. 2024) significantly improving detection accuracy, computational efficiency, and real-time inference. These models are particularly suitable for large-scale street-level imagery due to their real-time performance and ability to handle dense scenes. More

recently, transformer-based architectures such as DETR (Carion et al. 2020) have reformulated object detection as a set-prediction problem, eliminating the need for handcrafted components like anchor boxes and non-maximum suppression. These models leverage global attention mechanisms to capture long-range dependencies within complex scenes.

Weakly supervised and semi-supervised approaches have also gained attention due to the high cost of manual annotation in large-scale datasets. Methods based on Class Activation Map (CAM) (Zhou et al. 2016) and Grad-CAM (Selvaraju et al. 2017) have been used to localize objects using only image-level labels, providing a practical alternative for large datasets with limited annotations. However, such approaches often produce coarse localization and are prone to false positives, especially in cluttered street scenes. In weakly supervised object detection, method such as One Point is All You Need (Zhang et al. 2025) demonstrate that reliable detectors can be learned using image-level or point-level annotations by exploiting vision-language alignment, object localization, and attention mechanisms. In semi-supervised object detection, transformer-based approaches including Sparse Semi-DETR (Shehzadi et al. 2024), Relational Matching (Wu et al. 2024), and the recent STEP-DETR (Shehzadi et al. 2025) leverage pseudo-label refinement, teacher-student learning, and query optimization to effectively utilize large quantities of unlabeled data. These methods have substantially reduced annotation requirements while achieving performance approaching fully supervised detectors on standard benchmarks.

Street-Level Imagery: Street-level imagery has emerged as a rich source of visual information for understanding urban environments and supporting applications such as infrastructure monitoring, autonomous mapping, and smart-city analytics. The rapid growth of large-scale image collections from platforms such as Google Street View and Mapillary has enabled the development of data-driven approaches for automatically extracting semantic information from urban scenes. Among these, object detection and scene understanding play a central role in identifying and localizing

relevant entities such as traffic signs, street furniture, and waste bins.

However, street-level imagery poses several challenges that distinguish it from traditional object-detection benchmarks. These include large variations in viewpoint, illumination, occlusions, and object scale, as well as cluttered backgrounds and dynamic urban settings. Furthermore, in many real-world scenarios, annotations are sparse or weakly supervised, limiting the applicability of fully supervised methods. As a result, robust detection and scene understanding in such environments require models that can generalize across diverse conditions while handling noisy and incomplete supervision.

In the context of street-level imagery, several works have explored domain-specific challenges, such as small-object detection, occlusions, and viewpoint variations(Aldubaikhi and Patel 2025). Multi-scale feature representations and Feature Pyramid Networks (FPN) (Lin et al. 2017) have been widely adopted to improve the detection of objects of varying sizes. Additionally, context-aware models have been proposed to incorporate scene-level information, enabling more robust reasoning about object relationships in urban environments. Recent work has significantly advanced small-object detection, multi-view perception, and roadside scene understanding through transformer architectures (Du et al. 2025), collaborative perception (Daryani et al. 2025), and large-scale datasets (Wang et al. 2025). However, these methods primarily focus on improving object detection or 3D perception accuracy under fully supervised settings. In contrast, this thesis addresses the complementary problem of weakly supervised roadside object geolocation, where only image classification labels are available for detector training and no ground-truth cross-view object associations or object segmentation masks are provided.

A natural extension of this line of work considers combining street-level imagery with aerial or satellite imagery, as the two provide complementary perspectives of the same environment. From a methodological perspective, probabilistic object-based frameworks have also been used to jointly model object extraction and change detection from image pairs. For example, (Benedek et al. 2012) introduced a multi-

temporal marked point process framework that integrates low-level change evidence, object-level appearance models, spatial interactions, and prior knowledge through global optimization. While street-level images capture fine-grained visual details of objects, overhead imagery offers broader spatial context and more complete coverage of urban areas. Several studies have explored this idea by fusing information from both views to improve urban scene understanding. For instance, previous work (Wegner et al. 2016) demonstrated that combining aerial and street-level imagery can improve the detection and mapping of urban objects such as trees. Similarly, (Hoffmann et al. 2019) and (Cao et al. 2018) showed that integrating features from both modalities leads to more robust classification and detection in complex urban environments, although challenges remain due to the significant viewpoint differences between ground-level and overhead images.

Closely related to this, cross-view geolocation methods have also highlighted the importance of learning correspondences between street-level and aerial imagery. Early works (Lin et al. 2013; Tian et al. 2017) focused on matching ground-level images to aerial views by learning viewpoint-invariant representations, demonstrating that meaningful relationships can be established despite drastic appearance changes. More recent surveys further emphasize that combining aerial and street-level data is becoming increasingly important for large-scale urban analysis and mapping applications (Biljecki and Ito 2021). Overall, these studies suggest that integrating multiple viewpoints, rather than relying on a single modality, can significantly enhance both detection and localization performance in real-world urban settings, which aligns closely with the motivations of this thesis.

3D LiDAR vs. Street-Level Imagery: While this thesis focuses on street-level imagery, roadside object geolocation can also be addressed using active sensing modalities such as terrestrial or mobile LiDAR. LiDAR offers several advantages over passive imaging, including accurate 3D geometry, direct depth measurements, robustness to illumination changes, and reduced sensitivity to shadows and appearance variations. Consequently, LiDAR-based approaches have become a cornerstone

of autonomous driving and high-definition mapping, with methods such as PointPillars (Lang et al. 2019), and VoxelNeXt (Chen et al. 2023) demonstrating highly accurate 3D object detection from point clouds. More recently, multimodal fusion frameworks have combined the complementary strengths of LiDAR and RGB imagery. Representative methods include BEVFusion (Liang et al. 2022), which unifies LiDAR and camera features in bird's-eye-view space, and more recent approaches such as GAFusion (Li et al. 2024) and IS-Fusion (Yin et al. 2024), which employ transformer-based feature fusion to jointly exploit geometric and semantic information for SOTA 3D object detection. These methods demonstrate that LiDAR and cameras provide highly complementary information for scene understanding.

Importantly, these advantages come at the cost of specialized sensors, precise calibration, dedicated data collection, and substantially higher storage and computational requirements for processing dense 3D point clouds. In contrast, street-level imagery is inexpensive, widely available through platforms such as Google Street View (gsv) and Mapillary for nearly two decades, and provides extensive geographic coverage together with essential metadata, including camera GPS location, heading, and acquisition time. Combined with recent advances in deep learning for object detection and visual similarity learning, these characteristics make image-based methods particularly suitable for scalable roadside object geolocation and deployment on resource-constrained edge devices.

Nevertheless, LiDAR and street-view imagery should not be viewed as competing sensing modalities, but rather as complementary sources of information. Recent camera-LiDAR fusion methods demonstrate that combining the precise geometric measurements of LiDAR with the rich semantic information of RGB imagery can substantially improve robustness under occlusion, adverse illumination, and complex urban environments. Similarly, the TUMTraf V2X Cooperative Perception Dataset (Zimmer et al. 2024) highlights the benefits of integrating roadside and vehicle-mounted camera–LiDAR observations. Extending the proposed framework to incorporate LiDAR within a multimodal fusion architecture therefore represents

a promising direction for future work.

Weak Supervision: Overall, while object detection and cross-view matching methods have advanced significantly, many existing approaches still rely on strong supervision or focus mainly on single-image analysis. In the context of this thesis, weak supervision means that the training data does not contain detailed object-level annotations such as object identities, bounding boxes, segmentation masks, or explicit correspondences across views. For object detection, instead of using fully annotated images with precise object locations, only image-level labels are available: images containing the target roadside object are labeled as class 1, while images without the object are labeled as class 0. This reduces the detection stage to a binary classification problem, from which approximate object locations are later inferred using explainability-based localization methods.

Weak supervision also affects the re-identification and multi-view association stage. For roadside objects, there are typically no large-scale datasets that provide object identities or associations of the same physical object across multiple street-level images. This is particularly challenging for geolocation, where reliable multi-view association is needed to determine whether detections from different images correspond to the same real-world object. Moreover, accurate roadside object geolocation requires not only visual annotations, but also spatial metadata such as camera GPS location, image heading, field of view, and other geometric information. Since such richly annotated datasets are rarely available, the problem must be addressed under noisy and weakly supervised conditions. This motivates the need for methods that combine visual similarity, geometric constraints, and structured reasoning, as explored in this thesis.

2.5 Multi-view Geometry and Triangulation

Estimating the real-world location of objects from images requires reasoning beyond single-view perception and involves understanding the geometric relationships between multiple observations. Multi-view geometry provides a mathematical frame-

work for reconstructing 3D structure from multiple images by exploiting known or estimated camera parameters and viewing directions. In the context of street-level imagery, where camera positions are typically geo-tagged, and object observations are associated with bearing and depth cues, triangulation becomes a natural approach for object geolocation.

Triangulation: Given multiple observations of the same object from different viewpoints, triangulation estimates the object's position as the intersection of corresponding viewing rays. In ideal conditions, where measurements are accurate and noise-free, these rays intersect at a single point. However, in real-world scenarios, uncertainties in camera pose, object detection, depth estimation, and bearing measurements result in imperfect intersections. Consequently, practical triangulation methods must account for noisy observations and often rely on probabilistic or optimization-based formulations to estimate the most likely object location.

Multi-view Geometry: The foundations of multi-view geometry are well established in the computer vision literature, with classical formulations based on projective geometry, epipolar constraints, and camera calibration (Hartley and Zisserman 2004; Szeliski 2010). Early approaches focused on reconstructing 3D scenes from calibrated image pairs using linear and non-linear triangulation methods, including midpoint and least-squares solutions (Hartley and Zisserman 2004). These methods assume accurate correspondences between image points and known camera parameters, which are often difficult to obtain in unconstrained environments such as street-level imagery.

Structure from Motion (SfM): In the context of large-scale urban reconstruction, SfM techniques have been widely used to jointly estimate camera poses and 3D scene structure from unordered image collections (Schönberger and Frahm 2016). While SfM provides dense and accurate reconstructions, it typically requires strong feature correspondences across images and is computationally intensive, making it less suitable for scenarios with sparse or object-specific observations.

Probabilistic graph models: Probabilistic graph models provide a principled

framework for modeling spatial relationships and uncertainty in multi-view geolocation tasks. These models represent dependencies between observations and latent object locations through graph structures, enabling the integration of geometric constraints, measurement uncertainty, and contextual priors. In the context of roadside object geolocation from street-level imagery, graphical models are particularly effective in resolving ambiguities caused by noisy detections, inaccurate depth estimates, and inconsistent multi-view associations.

MRFs are widely used for spatial reasoning because they provide a natural way to model dependencies between multiple observations. In geolocation tasks, particularly for roadside objects, MRFs are typically defined over intersections (candidate object locations) generated through triangulation from multiple views. Each intersection represents a possible object position, where unary potentials capture how well an individual observation supports that intersection, while pairwise potentials enforce geometric consistency across different views.

A related approach (Krylov et al. 2018) formulates roadside object geolocation as an energy minimization problem using a custom MRF. Object detections from street-level images are combined with estimated depth and bearing information to generate candidate ray intersections, where each intersection represents a possible object location. Each candidate intersection x_i is assigned a binary label $z_i \in \{0, 1\}$, where $z_i = 1$ indicates that the candidate is selected and $z_i = 0$ indicates that it is rejected. The objective is to minimize the following energy function:

$$U(z, \Theta) = \sum_{i=1}^{N_Z} \left[\alpha u_1(z_i) + \beta u_2(z_i) + (1 - \alpha - \beta) u_3(z_i)\right], \tag{2.1}$$

where u_1, u_2, and u_3 are custom energy terms designed to enforce geometric consistency.

The first term, u_1, measures depth consistency:

$$u_1(z_i) = z_i \sum_{j=1,2} \|\Delta_j - d_j\|, \tag{2.2}$$

where Δ_j is the estimated monocular depth from camera j, and d_j is the triangulated distance from camera j to the candidate intersection. This term penalizes candidates whose triangulated locations disagree with the estimated object depths.

The second term, u_2, encourages spatial consistency between selected neighboring intersections:

$$u_2(z_i) = z_i \sum_k z_k \|x_i - x_k\|, \tag{2.3}$$

where, k indexes the candidate intersections lying on the rays associated with x_i, and the summation is taken over all such intersections. It penalizes the simultaneous selection of neighboring candidates that are far apart, while allowing nearby intersections to support the same object hypothesis.

The third term, u_3, discourages the trivial solution in which all candidates associated with a detected ray are rejected:

$$u_3(z_i) = (1 - z_i) \prod_k (1 - z_k). \tag{2.4}$$

Thus, the MRFs balances depth agreement, geometric consistency, and sufficient support from candidate intersections. By minimizing the weighted energy, the model suppresses inconsistent intersections while retaining candidates supported by multi-view geometry. Conceptually, this custom MRF formulation is employed in Chapter 3 as a baseline geolocation pipeline that combines weakly supervised object detections with geometric reasoning to identify spatially consistent candidate object locations.

Subsequent works have explored variations of triangulation pipelines for different urban assets. Simplified triangulation without global optimization has been applied to tree detection (Lumnitz et al. 2021), while address-based alignment and heuristic filtering have been used for urban asset localization (Laumer et al. 2020). (Murphy et al. 2025) proposed a probabilistic framework for the precise geolocation of street furniture using geo-positioned energy maps that encode the spatial likelihood of object locations. The method integrates external geospatial information, such as

GIS layers, road maps, and placement constraints, and uses stochastic birth-and-death optimization to infer the most probable configuration of urban assets. Other approaches incorporate additional cues such as semantic context or temporal consistency to improve robustness.

Despite these advancements, triangulation-based geolocation remains highly sensitive to errors in input measurements, particularly in bearing and depth estimates derived from monocular images. In this thesis, the triangulation based method optimized by custom MRF model is initially evaluated with the weakly supervised object detection pipeline introduced in Chapter 3, and is later extended in Chapter 5 by incorporating an additional visual-similarity energy term to further support the selection of consistent object locations. However, since the energy terms are handcrafted, the model can struggle in more challenging scenarios involving severe noise, false detections, or incorrect associations across views. To address this limitation, Chapter 6 replaces the MRF with a learning-based approach based on a graph neural network, enabling the model to learn hypothesis selection under similarly noisy and realistic conditions.

2.6 Similarity Learning and Metric Learning

Weakly supervised object detection methods often introduce false positives and inconsistent detections across different views of the same scene. In the context of this thesis, such errors lead to incorrect associations between observations that are assumed to correspond to the same object, ultimately degrading the performance of the geolocation pipeline. In particular, when multiple images capture a similar region, weak supervision may yield spatially plausible but visually inconsistent detections, leading to unreliable triangulation and erroneous object localization. To address this limitation, it is necessary to incorporate visual cues that distinguish whether two observations correspond to the same physical object. This naturally reformulates the problem as one of object re-identification across multiple views, where the goal is to learn a representation that captures visual similarity despite

variations in viewpoint, scale, illumination, and occlusion.

Multiple Object Tracking: Object re-identification is closely related to Multiple Object Tracking (MOT), where the goal is to maintain consistent object identities as they move across frames. In practice, this is quite challenging, as objects may become occluded, disappear temporarily, or change appearance due to viewpoint and lighting variations. Early tracking methods relied heavily on motion cues, but more recent approaches have shown that incorporating appearance-based similarity is essential for reliable association. For example, DeepSORT (Wojke et al. 2017) combines motion with learned visual embeddings, allowing the tracker to better distinguish between similar-looking objects.

Building on this idea, several works have explored how to jointly use appearance, geometry, and temporal information to improve tracking performance. Graph-based approaches, for instance, model detections as nodes and learn how they should be connected over time (Braso and Leal-Taixe 2020), while other methods explicitly incorporate temporal consistency and spatial reasoning to handle more complex scenarios (Weng et al. 2020; Pang et al. 2023). More recent work also focuses on improving the robustness of learned features, particularly in situations where detections are noisy or ambiguous (Ma et al. 2024).

Although these methods are designed for tracking over time, the underlying problem is very similar to multi-view geolocation. In both cases, the key challenge is to correctly associate observations that belong to the same object despite noise and inconsistencies. This makes MOT-based re-identification approaches particularly relevant, as they highlight the importance of combining visual similarity with geometric and contextual cues for reliable association.

Vehicle Re-Identification: A large body of work in vehicle re-identification has explored the problem of matching the same vehicle across different cameras and viewpoints, particularly in urban surveillance settings. This is inherently challenging, as vehicles often appear quite similar and their visual characteristics can change significantly depending on viewpoint, lighting, and occlusions. Early work in this

area focused on large-scale matching and progressive search strategies to handle such variations (Liu et al. 2016b).

As deep learning became more prominent, many approaches shifted toward learning discriminative feature embeddings using metric-learning techniques, such as contrastive and triplet losses. (Zhang et al. 2017; Kuma et al. 2019). These methods aim to bring images of the same vehicle closer together in the feature space while pushing images of different vehicles apart. At the same time, researchers began incorporating additional cues, such as spatial-temporal information, to improve robustness in real-world scenarios (Shen et al. 2017; Wang et al. 2017; Wu et al. 2018a). Multi-modal approaches have also been explored, where visual features are combined with attributes or metadata to better handle challenging cases (Tang et al. 2017; Liu et al. 2018b).

Another important direction has been explicitly addressing viewpoint variation. Several works propose multi-view learning strategies that aggregate information from different perspectives (Zhouy and Shao 2018; Zhou et al. 2018), while attention mechanisms and region-aware models focus on the most informative parts of the vehicle to reduce the impact of background clutter (Liu et al. 2018a; Zhu et al. 2018). More recently, methods incorporating adversarial learning and re-ranking strategies have been introduced to refine similarity predictions and improve robustness to synthetic data (Lou et al. 2019a; Wu et al. 2018b).

Person Re-Identification: Person re-identification has been one of the most widely studied problems in similarity learning, aiming to match individuals across non-overlapping camera views. Similar to vehicle re-identification, this task is difficult due to variations in pose, illumination, background, and occlusion. Early deep learning approaches focused on learning pairwise similarity using Siamese networks and contrastive loss formulations (Li et al. 2014; Ahmed et al. 2015). These methods were later improved through triplet-based learning strategies, which enforce relative similarity constraints and have shown strong performance in practice (Yuan et al. 2020; Schroff et al. 2015).

As the field progressed, more sophisticated architectures were proposed to better capture fine-grained visual differences. For example, part-based models (Sun et al. 2018) learn localized features for different body regions, thereby improving robustness to misalignment and occlusion. Attention-based methods further enhance this by focusing on the most informative regions of the image (Li et al. 2018), while multi-scale and multi-branch networks capture features at different levels of detail (Chen et al. 2017). Additional cues such as pose estimation and semantic attributes have also been incorporated to improve discriminative power (Su et al. 2017; Sun et al. 2018).

More recently, there has been increasing interest in handling noisy labels and domain shifts, which are common in real-world data. Approaches based on unsupervised and semi-supervised learning (Ge et al. 2020; Zhong et al. 2019), as well as clustering and memory-based techniques (Fan et al. 2018), aim to learn robust representations without relying on perfectly clean annotations. These ideas are particularly relevant in weakly supervised settings, where incorrect associations can negatively impact performance.

Existing vehicle and person re-identification methods have demonstrated strong performance in learning discriminative feature representations that are robust to viewpoint, illumination, and occlusion variations. In particular, recent work in person MOT has begun to address the challenges posed by label noise, highlighting the importance of robust similarity learning in realistic settings where annotations are imperfect. However, most existing approaches are primarily designed for matching across camera networks and typically assume either clean supervision or do not explicitly account for the impact of noisy pairwise associations on downstream tasks. In the context of multi-view object geolocation, incorrect associations between observations can lead to substantial errors in triangulation, as even a small number of mismatched pairs can introduce significant geometric inconsistencies. This highlights the need for similarity learning methods that are both robust to noisy supervision and capable of integrating with geometric reasoning frameworks. Addressing

these limitations forms the basis of the approach developed in Chapter 4, where noise-aware similarity learning is proposed and subsequently applied in Chapter 5 to learn reliable cross-view similarities among roadside objects in the presence of noisy detections and weak supervision.

2.7 Graph Neural Networks for Structured Data

While probabilistic graphical models such as MRFs provide a structured way to model spatial relationships, they typically rely on hand-crafted energy functions and fixed assumptions about how observations interact. In practice, especially under weak supervision, these assumptions can be limiting. The relationships among views, detections, and hypotheses can be highly complex and difficult to capture with predefined potentials. This becomes even more challenging when the data is noisy or contains incorrect associations. In contrast, GNNs offer a more flexible alternative, as they learn these relationships directly from data rather than relying on manually designed terms. This allows them to adapt to complex patterns, better handle noise, and incorporate richer features such as visual similarity alongside geometric constraints, making them particularly well-suited for the type of multi-view reasoning required in this thesis.

Graph Neural Networks: GNNs are designed for problems where the data is naturally structured as a graph, i.e., a set of nodes connected by edges. Instead of treating each observation independently, a GNN learns by passing information between connected nodes, which allows it to model relationships as well as individual features. This makes GNNs especially relevant when the problem is not just about recognizing objects, but about understanding how different observations relate to one another. In this thesis, that is exactly the case: the final object location depends not only on the evidence from each view, but also on how multiple views, candidate hypotheses, and geometric constraints interact. For this reason, GNNs provide a natural extension beyond hand-crafted spatial models, since they can learn these dependencies directly from data (Kipf and Welling 2017; Veličković et al. 2018;

Simonovsky and Komodakis 2017).

At a general level, applying a GNN begins with defining a graph representation of the problem. The first step is to decide which items should be treated as nodes and which as edges. Depending on the application, nodes may represent objects, views, detections, or candidate locations, while edges encode relationships such as spatial proximity, geometric compatibility, or visual similarity. Once this graph is built, each node is assigned a feature vector, and sometimes edges are given attributes as well. The GNN then updates node representations through message passing, where each node aggregates information from its neighbors and combines it with its own features. Different architectures do this in different ways. Graph Convolutional Network (GCN) use neighborhood aggregation to learn locally consistent representations, Graph Attention Network (GAT) learn to assign different importance to different neighbors, and edge-conditioned models explicitly use edge features during message passing (Kipf and Welling 2017; Veličković et al. 2018; Simonovsky and Komodakis 2017). In practice, this is useful when the relationships between nodes are as important as the nodes themselves, which is often the case in spatial inference problems.

Graph Neural Networks in Geolocation: In geolocation more broadly, GNNs are attractive because location estimation is rarely an isolated prediction. It usually depends on combining multiple sources of evidence, such as image appearance, camera pose, spatial layout, and cross-view consistency. A recent survey of image and object geolocation notes that modern methods increasingly combine visual and geometric reasoning, especially for object-level localization from street imagery (Wilson et al. 2023). More generally, recent reviews of GNNs for multi-view learning also show that graph-based models are well-suited to settings where several related observations must be fused in a structured way rather than simply concatenated (Xiao et al. 2024). This is particularly relevant to street-level object geolocation, where different views may provide incomplete or noisy evidence, but their relationships still contain strong information about the true object position.

This motivation becomes even clearer in multi-view imaging. In such settings, the challenge is not only to detect an object in each image, but also to decide which detections belong together and how they should be combined spatially. Graph-based methods are well-suited to this because they can represent views, detections, and hypotheses as connected entities and reason over them jointly. For example, GeoGraph formulates multi-view object detection and geolocation as a graph problem, using a GNN to combine image evidence with relative pose information and to handle a variable number of views (Nassar et al. 2020). Taken together, these works suggest that GNNs are not just a convenient modeling choice for geolocation, but a particularly suitable one when the problem depends on structured interactions between views, geometry, and object correspondence.

2.8 Datasets for Street-Level Object Geolocation

Progress in multi-view object geolocation has been supported by several datasets that provide street-level imagery along with varying levels of geographic and object-level annotation.

Mapillary Vistas Dataset: Mapillary Vistas (Neuhold et al. 2017) is a large-scale street-level dataset designed for semantic scene understanding. It contains diverse urban imagery collected globally, with pixel-level annotations spanning a wide range of object categories, including vehicles, pedestrians, and road infrastructure. The dataset captures variations in weather, lighting, and geography, but does not provide explicit multi-view correspondences or object-level geolocation annotations.

ARTSv2 Dataset: The ARTS dataset (Wilson et al. 2022) focuses on traffic sign detection and geolocation from street-level imagery. It provides annotated images with sign categories and associated geographic information, enabling both recognition and spatial localization tasks. The dataset is structured around roadside objects, making it suitable for studying object-level geo-referencing, although it typically assumes relatively clean annotations and limited viewpoint variation.

Uber-TL Dataset: The Uber-TL (Chaabane et al. 2021) dataset is a large-scale

collection of street-level imagery designed for traffic light detection and geolocation. It consists of images captured by vehicle-mounted cameras in urban environments, along with annotations for traffic light instances, including their spatial locations, states, and orientations. The dataset provides geographic information that enables mapping detected objects to real-world coordinates, making it suitable for studying object-level geolocation. However, similar to other large-scale datasets, it does not explicitly provide multi-view correspondences between observations, and the data is collected under relatively structured driving conditions.

Pasadena Multi-View Dataset: The Pasadena Multi-View dataset (Braso and Leal-Taixe 2020; Nassar et al. 2020) is designed for multi-view analysis of urban trees. It contains street-level images capturing the same objects from different viewpoints, along with annotations that enable cross-view matching and object-level geolocation. This dataset is particularly useful for studying multi-view consistency and triangulation-based localization under controlled urban settings.

GeoSign Dataset: GeoSign (Nassar and Lefèvre 2019) is a large-scale dataset for traffic sign geolocation constructed from crowd-sourced street-level imagery, primarily sourced from platforms such as Mapillary. It provides annotations linking detected signs to geographic coordinates, enabling large-scale training and evaluation. However, the data is inherently noisy and unstructured, reflecting real-world acquisition conditions.

Auto Arborist Dataset: The Auto Arborist dataset (Beery et al. 2022) combines aerial and street-level imagery for urban tree analysis. It includes annotations for tree detection and classification across modalities, enabling cross-view reasoning between ground-level and overhead perspectives. While it supports multi-modal learning, its primary focus is on classification and inventory tasks rather than precise multi-view triangulation.

These datasets reflect a growing trend toward leveraging multi-modal data sources and large-scale image collections for urban mapping tasks. Despite these advances, existing datasets remain limited in several aspects. Many focus on specific

object categories, such as traffic signs or trees, and often rely on relatively clean annotations or structured data collection processes. Furthermore, explicit multi-view associations and precise ground-truth triangulation are rarely available. These limitations highlight the need for methods that can operate under weak supervision and noisy observations.

Chapter 3

Roadside object geolocation from street-level images with reduced supervision

We propose a method for automated detection and geolocation of roadside objects from street-level images by leveraging historical records of these objects. Such partial and/or noisy geo-records are often held by infrastructure owners and require frequent updating. We aim to reduce the amount of image-level supervision required for the deployment of deep learning methods to the geolocation problem from segmentation masks (very costly) to binary image labels (lower cost). Our proposed method integrates an image classification deep learning pipeline with Grad-CAMs and the watershed transform to identify the positions of roadside objects of interest in the images. The geolocation is performed by deploying the existing Markov Random Field-based optimization module. We analyze the robustness of the proposed low-supervision geolocation model to noisy records. We report experiments for the detection of traffic lights and public bins, with geolocation of the latter performed in central Dublin. A published version of this work and the corresponding implementation are publicly available.[1]

[1]Publication at:`https://ieeexplore.ieee.org/abstract/document/10715092` and implementation: `https://github.com/waqar3411/Roadside_object_geolocation`

3.1 Introduction

Recent advances in large-scale street-level imagery and data-driven methods have enabled new possibilities for automated mapping of roadside infrastructure. In this chapter, we focus on the problem of detecting and geolocating roadside assets such as traffic lights, utility poles, and public waste bins using street-level images.

Accurate and up-to-date inventories of such assets are essential for urban planning, infrastructure maintenance, and navigation systems. However, maintaining these records through manual surveys is costly and difficult to scale. The increasing availability of geo-tagged street-level imagery from platforms such as Google Street View and Mapillary provides an opportunity to automate this process using computer vision techniques.

Several prior works have addressed roadside object geolocation by combining object detection with geometric reasoning. In (Krylov et al. 2018), a two-stage pipeline is proposed where a segmentation network is used to detect traffic lights, followed by a monocular depth estimator to infer object distance. An MRF formulation is then used to refine triangulation and improve localization accuracy. Variants of this approach have been explored for other object categories, including street trees (Lumnitz et al. 2021; Laumer et al. 2020) and traffic signs (Pedersen and Torp 2020), while detection models such as RetinaNet have been applied to identify infrastructure elements in street-level imagery (Zhang et al. 2018; Wilson et al. 2022). Video-based approaches have also been proposed to exploit temporal consistency for object geolocation (Chaabane et al. 2021).

Beyond single-modality approaches, multi-sensor fusion methods have been investigated to improve robustness. For example, LiDAR data has been combined with street-level imagery to enhance the localization of traffic lights and utility poles (Krylov and Dahyot 2018), while other works integrate aerial imagery and airborne laser scanning data for urban asset mapping (Rodríguez-Puerta et al. 2022). Cross-view matching techniques based on Siamese networks have also been proposed to associate object observations across multiple images, and in some cases extended

to fuse aerial and ground-level imagery (Nassar et al. 2020).

Despite these advancements, most existing approaches rely on fully supervised detection or segmentation models, which require large amounts of finely annotated training data. This requirement poses a significant limitation in practice, particularly for large-scale deployments where annotations are expensive and often unavailable or outdated.

Although large-scale, fully annotated datasets and synthetic environments (e.g., CARLA) are available for object detection and instance segmentation, they do not address the requirements of the roadside object geolocation problem considered in this thesis. Object detection represents only the first stage of the geolocation pipeline. While conventional datasets can be used to train accurate object detectors, they typically lack the spatial metadata required for geolocation, including camera GPS coordinates, camera heading, field of view, and other geometric parameters needed to estimate object bearings and perform triangulation. Furthermore, they do not provide ground-truth associations of the same physical roadside object across multiple street-level views, nor do they contain the real-world geographic locations of individual objects required for evaluating geolocation accuracy. Consequently, although fully supervised detection datasets can improve the detection stage, they cannot be directly used to train or evaluate the complete roadside object geolocation pipeline. This motivates the use of weak supervision, where only image-level labels are required for detector training, while the subsequent stages exploit geometric reasoning and multi-view consistency to recover object locations without relying on expensive object-level and cross-view annotations.

Weakly supervised methods are not intended to replace fully supervised detectors when richly annotated geolocation datasets are available. Rather, they provide a practical and scalable alternative for applications where collecting object-level annotations, cross-view correspondences, and precise geospatial metadata is either expensive or simply unavailable, which is currently the case for most roadside object geolocation tasks.

To address these challenges, this work investigates a weakly supervised alternative to traditional detection pipelines. Instead of relying on pixel-level annotations or bounding boxes, we employ an image-level classification framework, which requires only a single label per image and is therefore significantly less expensive to obtain.

To recover spatial information from classification models, we leverage Grad-CAM (Selvaraju et al. 2017) to generate class-specific activation maps, highlighting regions of interest corresponding to the target objects. These activation maps are further refined using a watershed-based segmentation approach to estimate precise object locations within the image.

The resulting object detections are then integrated into a geometric triangulation framework, following the pipeline introduced in (Krylov et al. 2018), to estimate real-world object locations from multiple observations. Additionally, we explicitly investigate the robustness of the proposed approach under noisy and partially corrupted supervision, reflecting realistic conditions where historical records may be incomplete or inaccurate.

The main contributions of this work are as follows:

- Weakly Supervised Detection Framework: A classification-based approach for roadside object detection that eliminates the need for bounding box or segmentation annotations.

- Spatial Localization via Activation Maps: Integration of Grad-CAM with watershed segmentation to recover object locations from image-level supervision.

- Robustness to Noisy Supervision: A systematic analysis of the impact of noisy and outdated labels on detection and geolocation performance.

- Integration with Geometric Triangulation: A complete pipeline that combines weakly supervised detection with multi-view triangulation for real-world object geolocation.

3.2 Proposed geolocation pipeline

In this work, we propose an effective two-step method to detect and geolocate roadside objects. This includes object detection using DL algorithms, and object geolocation by resorting to triangulation. The overall proposed pipeline is shown in Fig. 3.1.

In our study, we consider the training data having only two labels, that are, 0 (no object) and 1 (object presence). There is a single label per image, and the binary mask or the bounding boxes are not assumed available. To achieve the object detection and localization we followed a five-step pipeline. In step A, an image is classified based on the presence or absence of the object using Resnet34 classification model. In step B, the Grad-CAM is calculated for the image that contains the object to roughly position the activations in the image. Once the area is highlighted, the watershed algorithm is applied to find the local maxima in the highlighted area constituting step C. In step D, the approximate distance from the camera position to the local maxima is determined using a pre-trained depth model (Farooq Bhat et al. 2021). The final configuration of the objects is then reported by performing step E, i.e. custom-grid MRF optimization for object triangulation (Krylov et al. 2018) based on detection from steps A-C and distances from step D.

Grad-CAM (Selvaraju et al. 2017) is a method developed for activations maps analysis in DL models. The last convolutional layer of the model contains detailed information. Grad-CAM uses this information and generates a heatmap, localizing the area that is relevant for the prediction of a specific class. Specifically, Grad-CAM is calculated as the gradient of the class-specific output y^c w.r.t. the feature map activations A^k of a convolutional layer. Then the average of all the gradients over width i and height j dimensions is taken:

$$\alpha_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^c}.$$

Finally, ReLU function is applied to the weighted combination of feature map

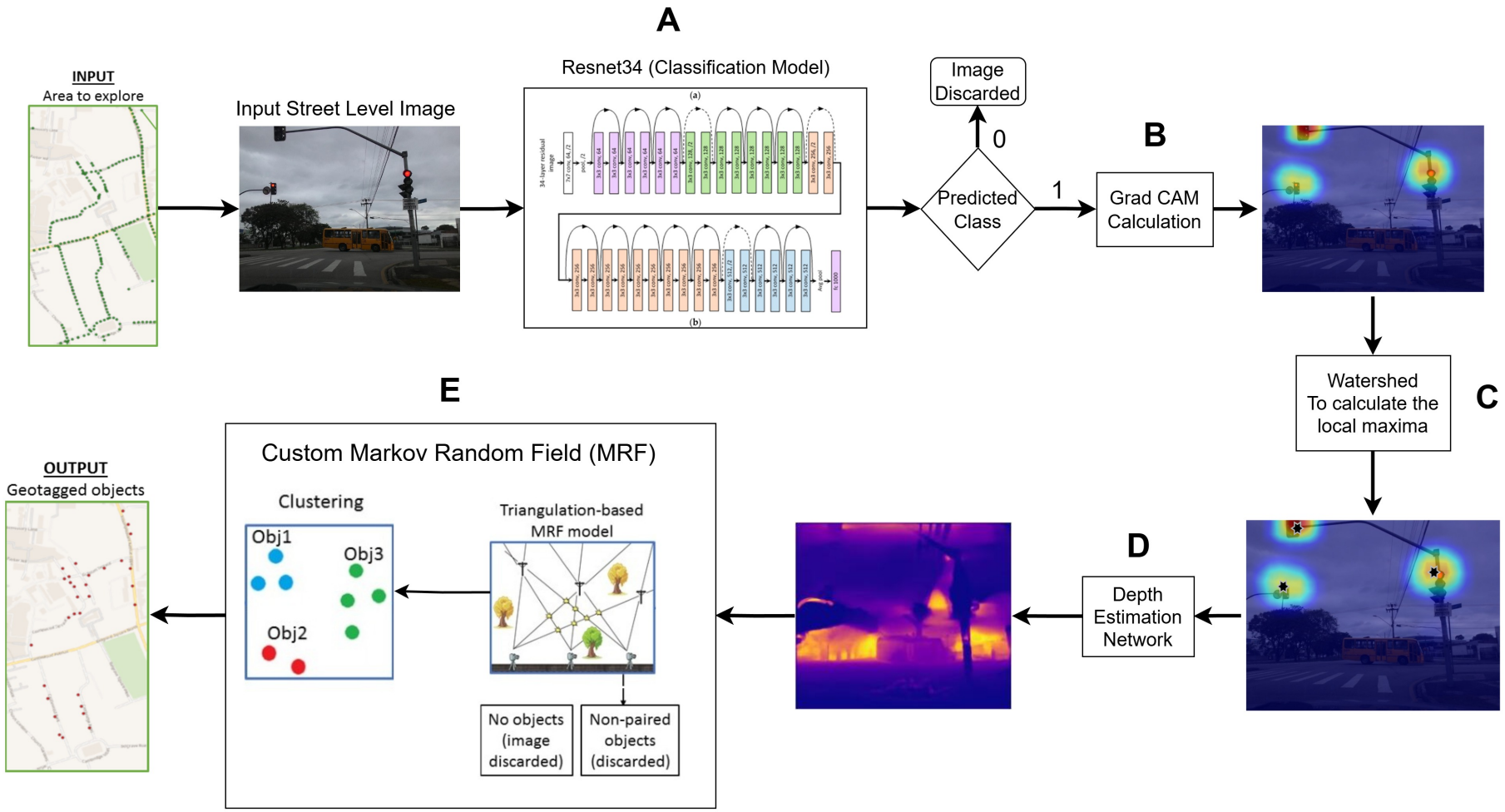


Figure 3.1: Proposed geolocation pipeline. Step A: Image classification by Resnet34. Step B: The Grad-CAMs generation of the classified Image. Step C: Calculation of local maxima using Watershed. Step D: Depth estimation using pre-trained DL algorithm. Step E: Triangulation using Markov Random Fields.

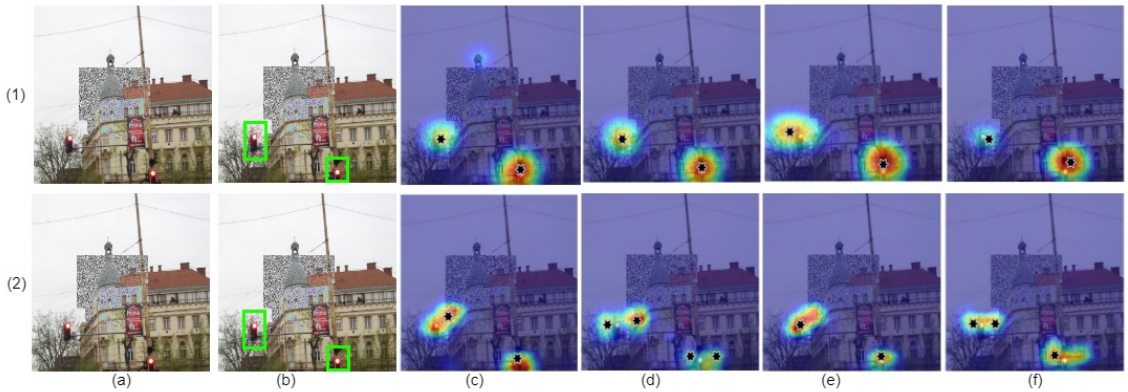


Figure 3.2: Grad-CAMs of traffic light dataset. Row 1: Resnet34 results. Row 2: Resnet50 results. Column (a): Input Image, (b): True Traffic Light Position (c): No label noise, (d)-(f): 1%, 5%, and 10% label noise respectively. Black dots are the local maxima identified using watershed transform.

activation A^k:

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k).$$

After performing Grad-CAM activation analysis on the classified images we then deploy the watershed algorithm to identify unique image positions for each activation blob, see Fig. 3.2.

The watershed algorithm thresholds the image, followed by applying a distance transform that identifies the foreground regions. A local maximum is then calculated for each foreground region.

The estimated depth paired with camera coordinates and image bearing are input to triangulation. Triangulation is performed using the custom-grid MRF (Krylov et al. 2018). This algorithm performs stochastic optimization to solve the numerical problem of finding optimal object positions as intersections of multiple view rays. Object information is contributing via energy terms that penalize discrepancies between depth estimation and intersection positions, detection false positives, and multiple objects simultaneously occupying the same position. For more details refer to (Krylov et al. 2018). The final geolocations returned by the MRF are fed through a local distance-based clustering to avoid double detections. The mean position of each cluster is the final output of our pipeline.

3.3 Experiments

3.3.1 Datasets and setup

In this work we consider two roadside objects: traffic lights and public bins. The images of traffic lights are extracted from the Mapillary Vistas dataset. Based on the presence of traffic light, 9,000 images, were extracted. The selection is done using the segmentation annotations provided. A total of 72,000 image crops of size 1200 x 1200 were generated, scaled 1/2 is each dimension for training. We thus obtained 24,000 images both in class 0 (no traffic light) and 1 (traffic light).

To build the public bins datasets we have extracted the images from the Mapillary API. We have extracted only those images that are flagged as the images containing the public bins. We also use the publicly available dataset of street public bins provided by the Dublin City Council (DCC). To enforce consistency with the DCC dataset we have selected only those Mapillary images that are within 15 meters distance of the locations of the DCC bins. A total of 9,827 images were obtained for class 1. Class 0 images are obtained by extracting the images that are 30-100 meters away from DCC bins. Each image is re-scaled to 1500 x 900 resolution for processing.

Importantly, due to the lack of high quality segmentation ground truth available for traffic lights (from Mapillary Vistas dataset), we cannot assess the segmentation quality obtained using the proposed method for the public bins data. We therefore rely on these two datasets for different purposes: traffic lights for assessing the segmentation performance, and public bins to assess the geolocation quality.

Implementation Details. The ResNet34 model was implemented and trained on an NVIDIA TITAN X GPU (12 GB). Training was performed for 50 epochs using a batch size of 8. The Adam optimizer was employed with an initial learning rate of 0.001. A step-based learning-rate scheduler was used to progressively reduce the learning rate during training, with the learning rate multiplied by a decay factor of 0.5 after every five epochs. The Mask RCNN model was trained for 50 epochs

with a batch size of 8. Following the standard Mask R-CNN training configuration, stochastic gradient descent was used with an initial learning rate of 0.005, a momentum value of 0.9, and a weight decay of 0.0005. A step-based scheduler reduced the learning rate by a factor of 0.1 after every 10 epochs.

3.3.2 Traffic lights detection

We employ the traffic lights dataset to compare the performance of the proposed object detection method with reduced supervision to the use of full segmentation pipeline. We have deployed Resnet34 and Resnet50 models, with the corresponding classification scores reported in Table 3.1. We observe that both models are performing comparably. Due to the model complexity, Resnet34 is preferred over Resnet50.

For the traffic lights dataset we then experiment with adding noisy labels. We consider three corrupted datasets, based on the class labels noise (1%, 5%, and 10%). Specifically, for 1% noisy dataset, the labels of random 1% of the total number of images in both classes were swapped: 0.05% images from class 1 were swapped with 0.05% images from class 0. The similar process is done for the 5% and 10% noisy datasets. The same amount of image labels were also manipulated for Mask-RCNN analysis. For example, for the 10% noisy dataset, 5% of true binary masks were replaced by complete black masks, and 5% images having no desired object were added to the dataset, and superpixel-based random binary masks were generated and added as the true labels for these images. In Table 3.1 we can see that both models demonstrate reasonable tolerance to noising, with a decrease in the performance by several percentage points when 10% of the labels are corrupted. These results confirm that Resnet34 can be applied to noisy realistic datasets (like public bins), without dramatic loss in terms of classification performance.

Post image classification, we calculated the Grad-CAMs and binarized the highlighted areas. We have calculated the IOU-based precision, recall, and detection rate for the test dataset, and compared the performance to the Mask-RCNN model.

Table 3.1: Classification performance on Traffic Lights Dataset

Model	Label Flipped	Classification Score (%)			
		Precision	Recall	F1-score	Accuracy
Resnet34	None	86.18	87.33	86.75	91.59
	1%	84.61	88.00	86.27	91.00
	5%	85.03	83.33	84.17	90.12
	10%	79.26	86.66	82.80	88.65
Resnet50	None	85.80	88.66	87.21	91.80
	1%	90.00	84.00	86.89	92.01
	5%	84.86	86.00	85.43	90.75
	10%	90.29	80.6	85.21	91.17

From Table 3.2, we can see that Resnet34 is lagging in detection rate performance by only around 2.5 percent, but reports higher precision compared to Mask-RCNN. Importantly, this result is achieved with significantly reduced requirement on image supervision (for object segmentation task) while Mask-RCNN uses full supervision (bounding boxes, binary mask, and class labels, etc). Compared to the Resnet50, Resnet34 is performing better in terms of IOU score. This is an additional reason to select Resnet34 as our main DL classifier.

The qualitative results of object detection and localization are shown in Fig. 3.2. The rows represent the results from Resnet34 and Resnet50 and the columns show the input image and true position of the traffic lights followed by the corresponding Grad-CAMs for no label noise, 1%, 5% and 10% label noise, respectively. The black dots on each image represent the local maximum returned by the watershed algorithm. In row 2 columns (c)-(e), we can see two distant point-locations identified a single object. This is due to the inaccurate gradient shape calculated from the Resnet50 model.

Sample segmentation results for the traffic lights with Resnet34 and Mask-RCNN are shown in Fig. 3.3. Rows (2)-(3) show the results for the original dataset (no label noise), and 10% label noise respectively. We observe that the shapes returned by the Resnet34 are less accurate. However, since we target object geolocation in this study, we are solely interested in the local maxima of the Grad-CAMs and not the full delineation of the object. In Fig. 3.3 column (a), false areas are identified by

Table 3.2: Instance-based IOU score for traffic lights dataset

Model	Labels Flipped	Instance Based IOU Score (%)		
		Precision	Recall	Detection Rate
Resnet34	None	92.35	78.00	79.39
	1%	95.30	75.96	75.04
	5%	92.33	77.63	74.29
	10%	96.78	74.01	75.51
Resnet50	None	85.53	65.20	62.24
	1%	81.63	61.02	58.35
	5%	73.82	58.80	55.75
	10%	73.51	63.41	62.07
Mask RCNN	None	83.03	83.72	81.40
	1%	83.11	81.85	79.45
	5%	83.85	80.90	78.79
	10%	80.63	81.99	79.45

Mask-RCNN. In Fig. 3.3 column (b), the traffic lights are near to each other, and thus both models perform weakly.

3.3.3 Public bins detection and geolocation

In this task we consider two test sets: extracted from Mapillary API and DCC images. The latter has been captured for DCC in central Dublin (Ireland) in 2020 by Murphy Geospatial (`https://murphygs.com`). This dataset covers approx. 2.7kms of roads, with the images sampled approx. every 3 meters. Each sample location has 6 views captured with field of view of 68.77 degrees, covering the entire panorama. Based on the above experimental analysis Resnet34 is trained on the Mapillary data and tested on both test sets. The classification scores obtained are reported in Table 3.3. From the table, we can see that the model is performing well on both test sets.

For object detection part, instance-based IOU score is calculated only on the Mapillary test set due to the availability of the ground truth segmentation masks. We achieved 65% precision, 92% recall, and 83% detection rate. The reason for this low precision is a substantial presence of inaccurate labels in the dataset. The qualitative results of the Grad-CAMs are shown in Fig. 3.4. Column 1 shows the input images followed by corresponding Grad-CAMs and watershed-identified local

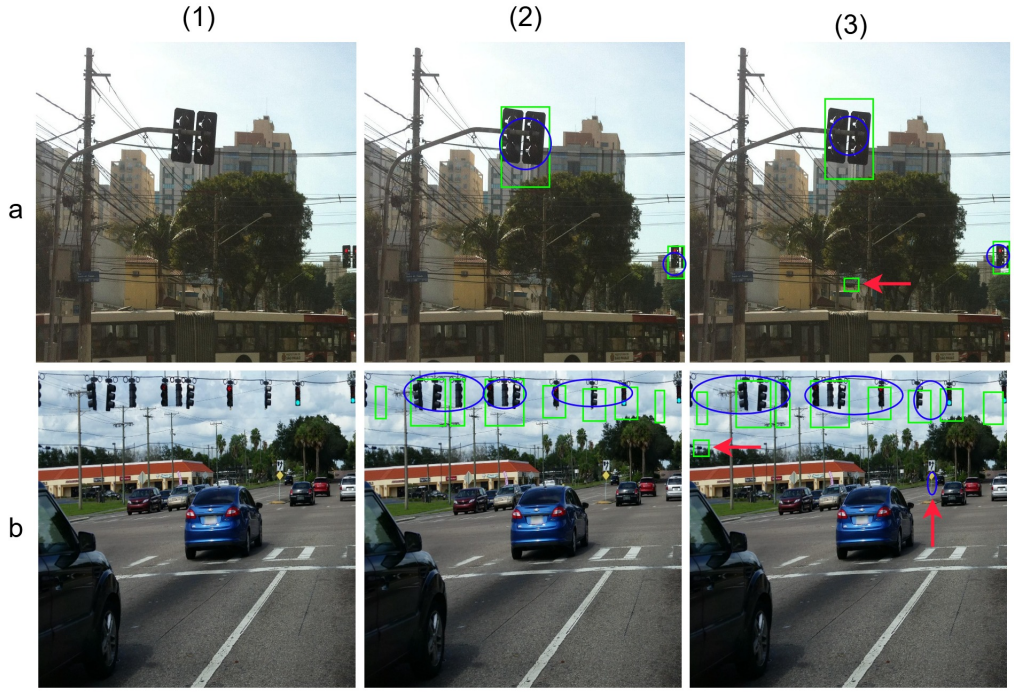


Figure 3.3: Segmentation from Resnet34 GradCAMs (blue ellipses), and Mask-RCNN (green rectangles). (1): Image, (2-3) Segmentations obtained by using: no label noise and 10% label noise datasets, respectively.

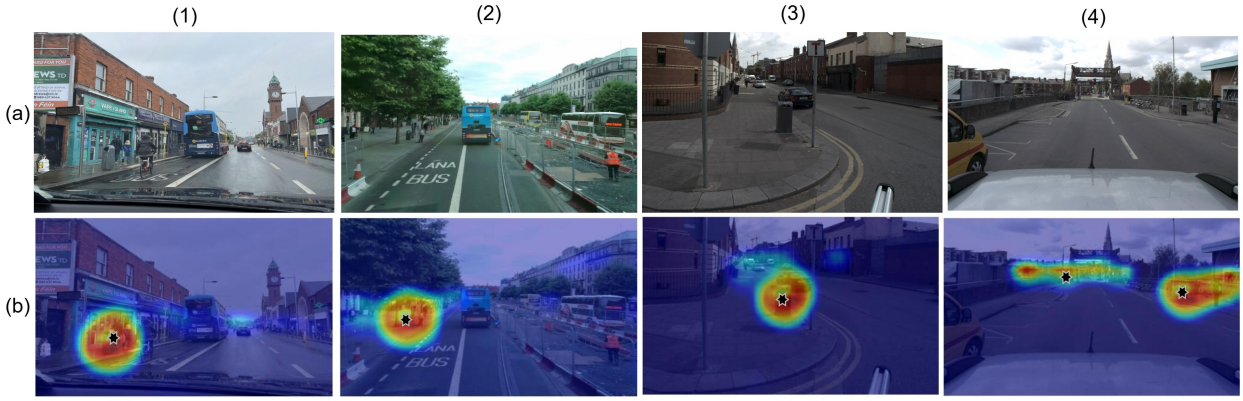


Figure 3.4: Grad-CAMs and local maxima identified via watershed (black dot) on public bins. Images from Mapillary (1-2), and the DCC dataset (3-4).

maxima. A false positive is highlighted at the centre of the image in the sub-figure of Fig. 3.4, specifically in row 2, column 4. Such false positives have been filtered out by the depth estimation model by considering only the objects within 25 meters of the camera position.

We now proceed to the evaluation of the complete roadside object geolocation pipeline. The geolocation experiments are performed only on stationary public bins, as their ground-truth geographic locations are available from Dublin City Coun-

Table 3.3: Instance Based IOU score for the public bins Dataset

Model	Test Set	Classification Score (%)			
		Precision	Recall	F1-Score	Accuracy
Resnet34	Mapillary	92.32	86	89.04	90.5
	DCC	95.69	89	92.22	92.5

cil (DCC), allowing the estimated object locations to be quantitatively validated. Although traffic lights from the Mapillary Vistas dataset are used to evaluate the proposed weakly supervised object detection framework, they are not included in the geolocation evaluation as described in Section 3.3.1. This is because the Mapillary Vistas dataset provides only object detection annotations (e.g., object classes, bounding boxes, and segmentation masks) and does not include the camera metadata required for geolocation, such as camera GPS coordinates, image heading, or camera field-of-view (fov). Furthermore, it does not provide the real-world geographic locations of the objects or ground-truth multi-view object associations necessary for evaluating a geolocation pipeline.

There are a total of 82 DCC stationary public bins within the area covered by the DCC imagery. The results produced by the proposed MRF-based geolocation framework are plotted together with the DCC bin locations on an OpenStreetMap layer in Fig. 3.5. Out of the 82 DCC bins, the presence of 63 bins is also identified by MRF. The average distance between the positions is 5.45 meters. The following comparative analysis is performed by manual inspection using the available Google Street View data in the area. The MRF-estimated locations of 40 bins are more accurate compared to the DCC bins, whereas 22 DCC dataset bin geolocations are more accurate. In addition, 43 genuine public bins are geolocated in this area which are absent from the DCC record. MRF has resulted in 46 false positives. This last number is severely impacted by the size of the training dataset, i.e. several types of objects were systematically confused for public bins (leading to over 2/3 of all the false positives), including bollards, bases of street lights, and cabinets. It is expected that these will be mitigated if the training data of class 'no-bin' includes more examples of these objects. Furthermore, the same types of training dataset

deficiency have resulted in 19 false negatives (i.e., public bins in the DCC record but not found by MRF).

These results are encouraging in suggesting the proposed geolocation pipeline has a strong capacity to validate and improve the existing geo-record by confirming the existing objects and identifying missing or erroneous records through automated processing of the most recent imagery available. The current limitations are due to the limited availability of imagery that has been available for this experimental study.

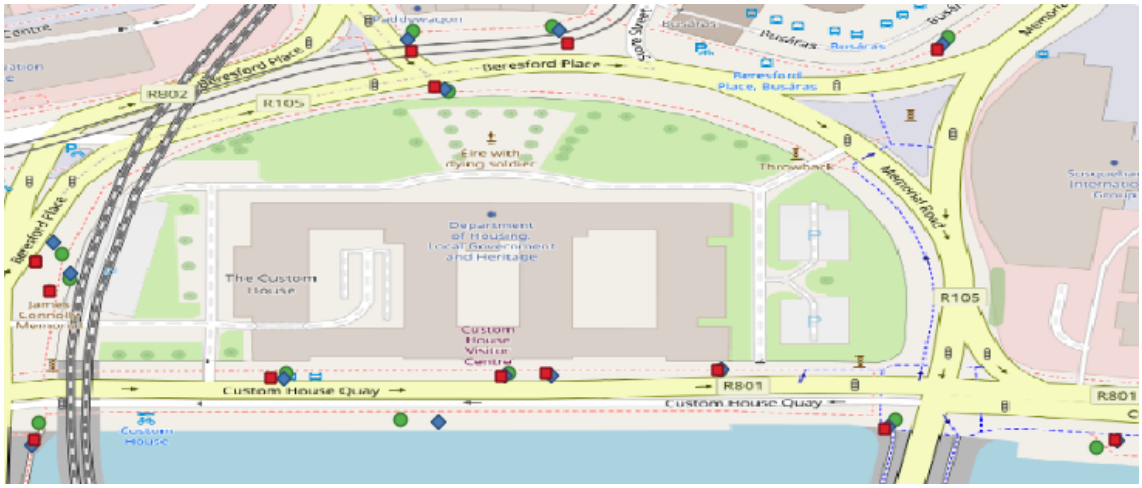


Figure 3.5: Public bins in central Dublin: DCC public records (green circles) vs. MRF (red squares) vs. manually identified ground truth (blue diamonds).

3.4 Conclusion

In this work, we have proposed an efficient method to achieve object detection using reduced supervision. The images are classified based on the binary labels, i.e. presence/absence of the roadside object of interest. Grad-CAMs followed by watershed transform have been employed to highlight the location of objects inside the images. We have also used monocular depth estimation and MRF-based triangulation procedure to geolocate the objects. We have experimentally observed that the reduced supervision procedure is capable of delivering object detection performance with minimal loss of accuracy and with reasonable tolerance to noise on traffic lights detection. The performance of the proposed geolocation pipeline has been validated on

public bins detection in central Dublin.These results highlight the strong potential of the proposed pipeline for automating the updating of infrastructure records.

At the same time, while the approach enables scalable training on large street-level datasets, it also introduces a higher number of false positive detections due to the lack of high level supervision. As illustrated in Fig. 3.6, sequences of cropped detections often contain incorrect or inconsistent object instances believed to be true positives. These noisy detections can significantly degrade the quality of downstream geo-localization. This limitation motivates the need for robust filtering mechanisms, leading to the development of a visual similarity-based approach presented in the following chapter.

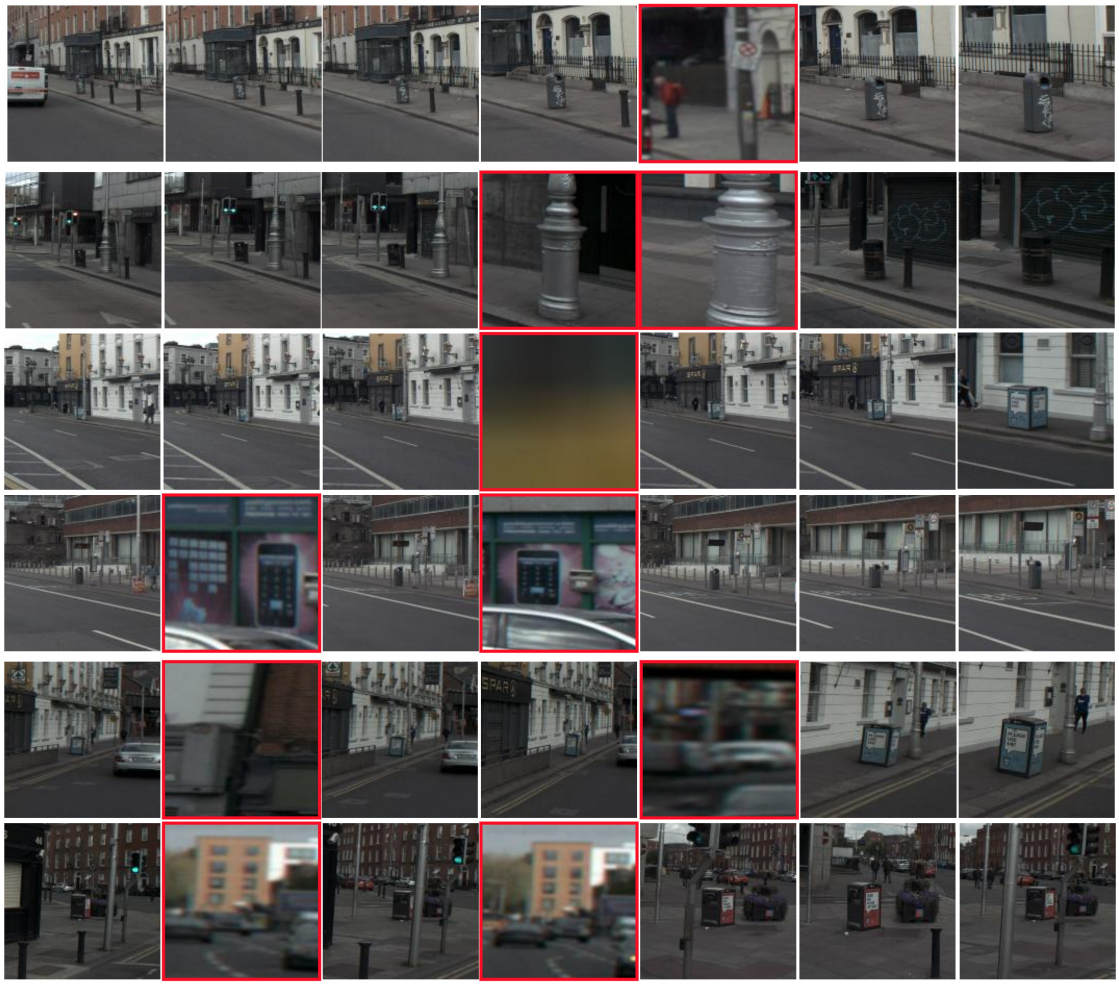


Figure 3.6: Example sequences of cropped detections from multi-view images, where each row corresponds to a single object detection across different views. False positive detections are highlighted in red, illustrating the presence of noisy and inconsistent observations within the sequences.

Table 3.4: Classification performance under varying retrieval distances (d) and segmentation area thresholds (A_{max}). The model is trained and evaluated on the Mapillary dataset.

Data		Classification Score (%)			
		Precision	Recall	F1-Score	Accuracy
$d = 10\,m$	$A_{max} \geq 1000$	87.64	78.00	82.40	83.50
	$A_{max} \geq 2000$	90.52	86.00	88.20	88.50
	$A_{max} \geq 5000$	89.41	76.00	82.16	83.50
$d = 15\,m$	$A_{max} \geq 1000$	**92.32**	**86.00**	**89.04**	**90.50**
	$A_{max} \geq 2000$	90.00	81.00	85.26	86.00
	$A_{max} \geq 5000$	87.20	82.00	84.53	85.00

3.5 Appendix 3A: Supplementary Results

This section presents preliminary experiments not previously discussed, conducted to inform key design choices for data preparation and model development.

During data collection, images for each DCC bin location were retrieved using the Mapillary API. At this stage, a suitable spatial distance d for image extraction was chosen. To assess this, images were extracted at $d = 10m$ and $d = 15m$ for each bin location, allowing evaluation of the balance between contextual information and object visibility.

As discussed in section 3.3.1, the Mapillary API provides high-resolution images, typically 4000×3000 pixels, which increase computational demands. Therefore, all images were resized to 1500×900 pixels. However, this reduction risks shrinking or losing segmentation masks of distant or small objects. Because the approach is semi-supervised, including weak or faint objects can lead to many false positives. To address this, an extra experiment defined inclusion criteria based on segmentation mask size: an image is positive only if its largest connected segmentation component area $A_{max} \geq 1000$. This ensures only clear object instances are used for training.

3.5.1 Result Analysis

Tables 3.4 and 3.5 present the classification and instance-level IoU-based segmentation performance under different values of (d) and A_{max}. The results show that both factors substantially affect model performance. Notably, the combination of

Table 3.5: Instance Based IOU Score under varying retrieval distances (d) and segmentation area thresholds (A_{max}). The model is trained and evaluated on the Mapillary dataset.

Data		Instance-based IoU Score (%)		
		Precision	Recall	Detection Rate
$d = 10\,m$	$A_{max} \geq 1000$	53.37	84.43	71.03
	$A_{max} \geq 2000$	52.00	82.54	71.72
	$A_{max} \geq 5000$	65.49	78.81	64.14
$d = 15\,m$	$A_{max} \geq 1000$	**64.90**	**92.00**	**82.69**
	$A_{max} \geq 2000$	62.59	78.38	60.00
	$A_{max} \geq 5000$	60.48	84.87	69.66

$d = 15m$ and $A_{max} \geq 1000$ yields the highest classification and segmentation scores, demonstrating the effectiveness of these choices. Additionally, testing the model trained on Mapillary images using the DCC dataset confirms its strong generalization ability under these optimal settings.

Table 3.4 shows that retrieving images within $d = 15m$ yields better classification results than at $d = 10m$. The best F1-score (89.04%) and accuracy (90.5%) occur when images within $d = 15m$ are selected, and segmentation blob areas A_{max} below 1000 are excluded. Conversely, raising the threshold to $A_{max} \geq 5000$ reduces precision (87.2%) and recall (82%), mainly because many valid bin-containing images are excluded. Thus, a less restrictive area threshold within $d = 15m$ best supports model generalization.

The segmentation results in table 3.5 support the same conclusion. The configuration with $d = 15m$ and $A_{max} \geq 1000$ achieves the best balance among IoU, precision, and recall, with 64.90% precision, 92.00% recall, and 82.69% detection rate. Thus, this setting is optimal for robust segmentation performance.

When evaluating the model trained on Mapillary images with the DCC dataset (as shown in table 3.6), the model achieves its highest performance with $d = 15m$ and $A_{max} \geq 1000$ (F1-score 92.22%, accuracy 92.5%), underscoring strong generalization. Increasing A_{max} improves precision at $d = 10m$ but degrades performance at $d = 15m$, indicating that including larger, noisier segments hampers generalization.

Overall, these results show that using images within $d = 15m$ and segmenta-

Table 3.6: Classification Performance under varying retrieval distances (d) and segmentation area thresholds (A_{max}). The model is trained on the Mapillary dataset and evaluated on the DCC dataset.

Data		Classification Score (%)			
		Precision	Recall	F1-Score	Accuracy
$d = 10\,m$	$A_{max} \geq 1000$	84.21	80	82.05	82.5
	$A_{max} \geq 2000$	83.49	86	84.72	84.5
	$A_{max} \geq 5000$	97.64	83	89.72	90.50
$d = 15\,m$	$A_{max} \geq 1000$	**95.69**	**89**	**92.22**	**92.5**
	$A_{max} \geq 2000$	95.50	85	89.94	90.5
	$A_{max} \geq 5000$	79.79	79.01	79.39	79.5

tion blob area $A_{max} \geq 1000$ yields the best balance of classification performance, segmentation accuracy, and robustness, outperforming other tested configurations. Accordingly, this configuration is selected for all subsequent experiments.

3.5.2 Visual Results

This section shows the visual results reported by the proposed method. Figure 3.7 depicts the multi-view input images from the DCC dataset, as described in Section 3.3.3. Six synchronized views, captured by six cameras at a single location, serve as input to the proposed model.

Figure 3.8 shows segmentation performance under rising label noise for three models: ResNet34, ResNet50, and Mask R-CNN, shown row-wise for each example. Column (a) shows the ground-truth object locations. Column (b) represents predictions without noise. Columns (c)-(e) correspond to 1%, 5%, and 10% noise.

The ResNet34 model consistently localizes the object accurately, even when predicted segmentation masks have irregular shapes, across all noise levels. This is tolerable, as the main goal is accurate object localization, not precise boundary reconstruction. The model is robust to noise, maintaining reliable localization even when mask quality drops.

ResNet50 and Mask R-CNN are more sensitive to increasing noise. In column (e) (10% noise), red arrows highlight false positives from Mask R-CNN, revealing overfitting to noisy patterns. In the second example, a red circle shows a missed

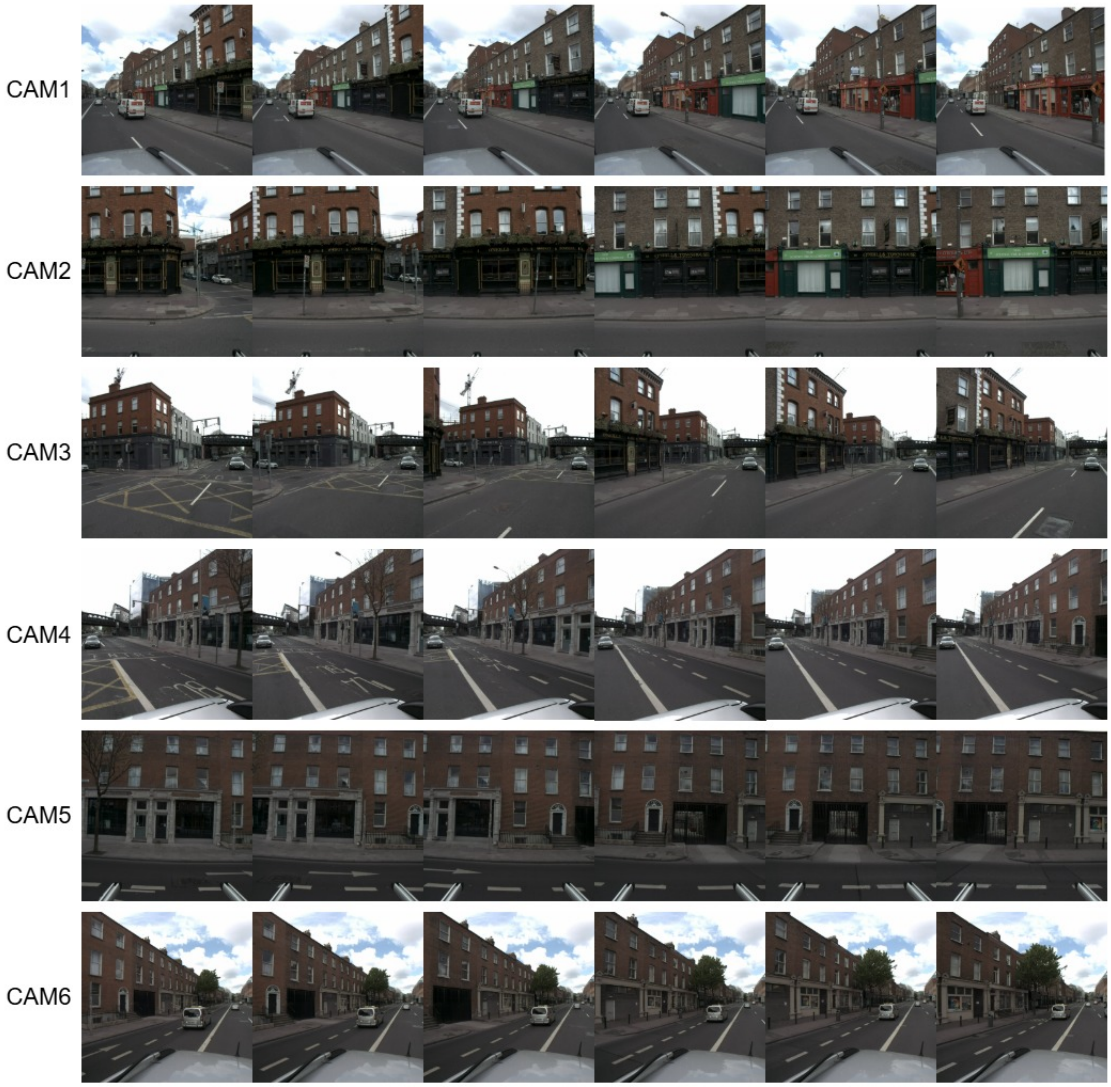


Figure 3.7: Multi-view image sequences used as input to the proposed model, showing six synchronized views of a single location captured by six cameras. Each row corresponds to a fixed camera viewpoint, and each column represents a consecutive frame.

detection by ResNet50, indicating reduced robustness in challenging conditions.

Overall, these observations indicate that while highly supervised models such as Mask R-CNN may aim to capture finer structural details, they can be more susceptible to noise. In comparison, ResNet34 provides more stable and reliable object localization under weak supervision and noisy training conditions, making it better suited for the proposed framework.

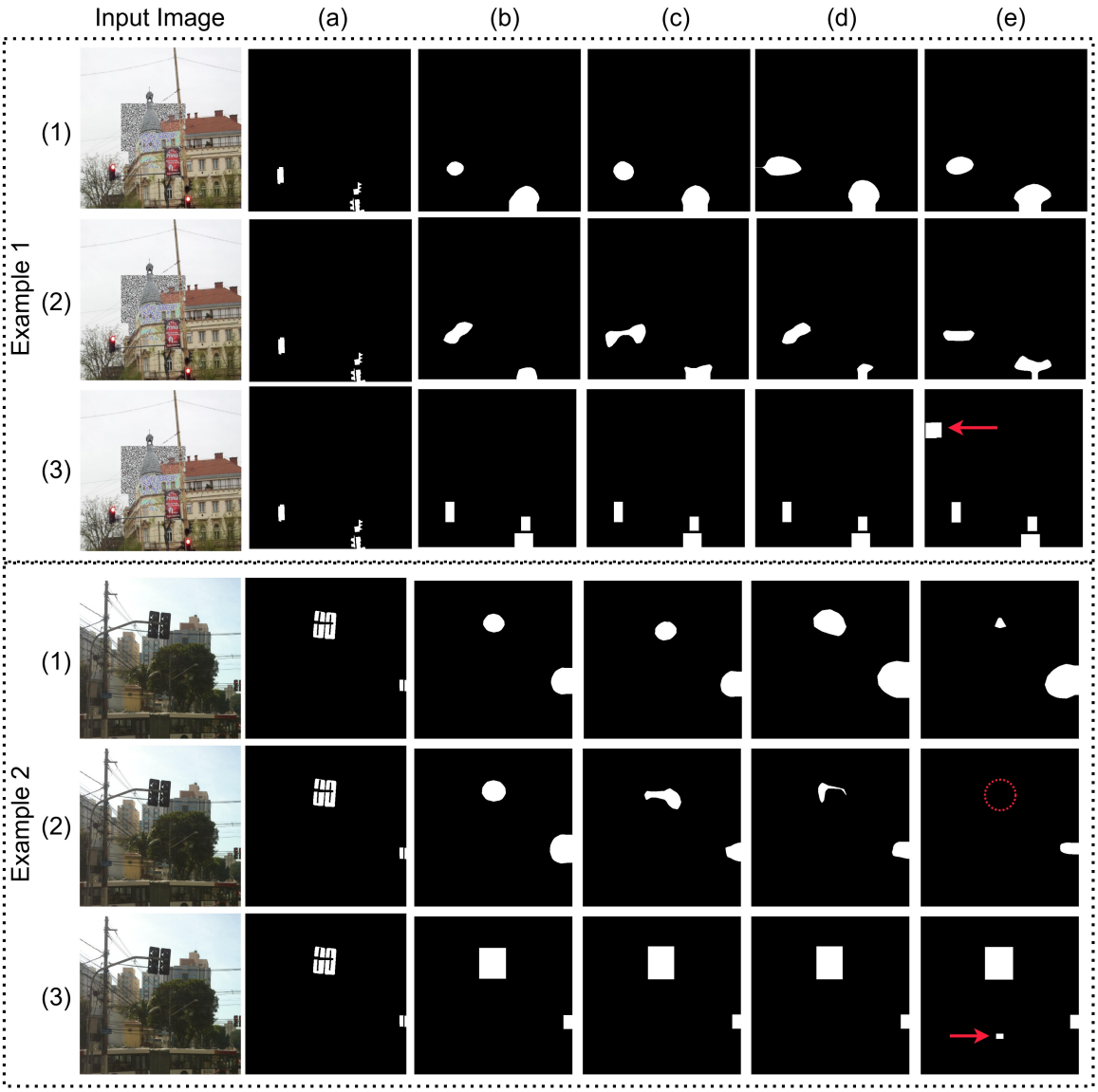


Figure 3.8: Qualitative comparison under increasing label noise (0%–10%). Columns (a)–(e) show ground truth, no noise, and 1%, 5%, and 10% noise, respectively; rows correspond to ResNet34, ResNet50, and Mask R-CNN. ResNet34 preserves accurate localization, while Mask R-CNN produces false positives (red arrows) and ResNet50 misses detections (red circle) at higher noise levels.

Chapter 4

Similarity-based Outlier Detection for Noisy Object Re-Identification Using Beta Mixtures

Object Re-ID methods are highly sensitive to label noise, which typically leads to significant performance degradation. We address this challenge by reframing Re-ID as a supervised image similarity task and adopting a Siamese network architecture trained to capture discriminative pairwise relationships. Central to our approach is a novel statistical outlier detection (OD) framework, termed Beta-SOD, which models the distribution of cosine similarities between embedding pairs using a two-component Beta distribution mixture model. We establish a novel identifiability result for mixtures of two Beta distributions, ensuring that our learning task is well-posed. The proposed OD step complements the Re-ID architecture combining Binary Cross-Entropy (BCE), contrastive, and cosine embedding losses that jointly optimize feature-level similarity learning. We demonstrate the effectiveness of Beta-SOD in de-noising and Re-ID tasks for person Re-ID, on CUHK03 and Market-1501 datasets, and vehicle Re-ID, on VeRi-776 dataset. Our method shows superior performance compared to the state-of-the-art methods across various noise levels (10-30%), demonstrating both robustness and broad applicability in noisy Re-

ID scenarios. A preprint of this work and the corresponding implementation are publicly available.[1]

The motivation behind evaluating Beta-SOD on person and vehicle re-identification datasets is to assess the robustness of the proposed method on large and realistic similarity-learning benchmarks before extending it to weakly supervised roadside object geolocation. Since weakly supervised object detectors often produce false positives and annotated datasets for roadside multi-view association are limited, Re-ID datasets are used as proxy sources for training cross-view similarity models. Controlled label noise is introduced to approximate false detections and ambiguous associations in realistic roadside scenarios. Beta-SOD further improves robustness by modeling the distribution of similarity scores and identifying unreliable matches, allowing spurious associations to be suppressed before they affect downstream geolocation. The detection of outliers in roadside multi-view object associations, together with additional numerical results obtained using Beta-SOD, is further presented in Chapter 5, as mentioned in Sec. 4.10. Overall, this provides a principled strategy for improving multi-view matching, which is essential for reliable triangulation and accurate roadside object localization under weak supervision.

The proposed work represents a combination of contributions. The concept of the method, its practical implementation, experimental design, and results analysis were carried out as the primary contribution of the main author. The theoretical identifiability proof presented in Section 4.4, which supports the choice of the beta distribution in the denoising method, was contributed by the co-authors indicated in the preprint version. [1]

4.1 Introduction

Object Re-ID is a fundamental task in computer vision that seeks to match instances of the same object across different cameras or viewpoints. Person Re-ID

[1] Preprint: `https://arxiv.org/abs/2509.08926` and implementation: `https://github.com/waqar3411/Beta-SOD`

has been extensively studied, tackling pose variation, occlusion, illumination, and background clutter (Ye et al. 2022b). Equally, vehicle Re-ID has gained prominence in intelligent transportation systems by confronting inter-class similarity, intra-class variability, and occlusion challenges (Wang et al. 2019a). Both subfields fundamentally aim to learn discriminative features that generalize across varied conditions. A further complication arises in practical systems from noisy labels: in object Re-ID, even small annotation errors due to mislabelled identities or detector misalignment can significantly distort learned representation boundaries. Indeed, recent work demonstrates that learning robust person Re-ID models under label noise poses a significant challenge, especially when each identity has limited samples, and naïve training can result in large degradation in accuracy and generalisability (Ye et al. 2022a; Liu et al. 2024).

To address the negative impact of label noise on deep Re-ID models, a variety of strategies have been developed in recent years. One line of work focuses on robust loss functions, since conventional objectives such as cross-entropy (CE) or mean squared error (MSE) are highly sensitive to mislabeled samples (Zhang and Sabuncu 2018; Wang et al. 2019c; Feng et al. 2021). A second line of research assigns adaptive importance weights to training examples, down-weighting the contribution of noisy labels while emphasizing reliable ones (Han et al. 2018; Zhang and Pfister 2021; Shu et al. 2019). Another family of approaches explicitly identifies mislabeled samples and either corrects them or removes them from training, thereby purifying the dataset (Karim et al. 2022; Pham et al. 2021; Wu et al. 2021). Collectively, these techniques aim to mitigate the effects of corrupted annotations, and improve the stability and generalization of Re-ID systems under realistic conditions, with loss-function design playing a particularly central role.

Among these strategies, the choice of loss function is especially critical, as it directly determines how the model learns from both clean and noisy samples. Standard CE loss, while common, is highly sensitive to label noise because of its strong penalties on misclassifications, often leading CNNs to overfit corrupted labels (Zhang

et al. 2021; Ghosh et al. 2017). Similarly, contrastive loss (CL), commonly applied in Re-ID tasks to enforce distance-based similarity constraints, is vulnerable to noisy annotations since mislabeled pairs are penalized even when the underlying match is correct (Wang et al. 2019b; Roth et al. 2020). Alternatives such as cosine-embedding loss replace Euclidean distance with angular similarity, offering improved robustness in high-dimensional spaces, yet they too remain affected by label noise due to their dependence on angular separation (Song et al. 2022; Schroff et al. 2015). To overcome these limitations, several robust loss formulations have been proposed, including generalized cross-entropy, bootstrapped loss, and noise-corrected contrastive objectives, which aim to reduce sensitivity to corrupted labels while preserving discriminative feature learning (Zhang and Sabuncu 2018; Reed et al. 2014; Yi et al. 2022). Nonetheless, loss-based remedies address the complexities of noisy Re-ID data only partially, motivating complementary approaches that filter or adjust corrupted samples through statistical modeling.

In our work we reformulate the Re-ID task as a binary image similarity problem by pairing images into similar and dissimilar pairs, and processing them through a Siamese network (Melekhov et al. 2016). Cosine similarity scores generated by the network are modeled using an explicit statistical model defined as a two-component Beta distribution mixture, where one component represents clean image pairs similarity scores, and the other captures noisy or mislabeled ones. An Expectation-Maximization (EM)-based OD algorithm identifies and filters noisy pairs. This OD step is alternated with network training every few epochs, allowing the model to iteratively refine the training data and improve task-specific representation learning. Multi-view pairing ensures that enough data is retained even after filtering, which is particularly useful in limited-data scenarios. To enhance learning, we employ a multi-loss strategy that enables the network to capture features beneficial for both similarity classification and Re-ID. The proposed architecture, termed Beta-SOD, is outlined in Fig 4.1.

At the heart of the proposed methodology lies statistical modeling for outlier de-

tection. To express both the signal and noise, or *contamination*, components underlying cosine similarity scores, we introduce the use of the Beta distribution (Gupta and Nadarajah 2004). This flexible two-parametric family is defined to model quantities restricted to the unit interval $[0, 1]$. To establish the theoretical soundness of this formulation and justify the use of EM, we provide, to the best of our knowledge, the first proof of *identifiability* for the two-component Beta mixtures.

The main contributions of this work are as follows:

- We reformulate the object Re-ID task as pairwise image similarity classification and propose an explicit statistical OD technique, based on modeling the cosine similarities as two-component Beta distribution mixtures, which is particularly relevant in the presence of noisy labels.

- We establish a novel identifiability result for mixtures of two Beta distributions. This ensures that our proposed filtering framework is rigorously well-posed.

- We employ three loss functions in our Siamese-based architecture, including BCE, contrastive and cosine embedding losses, enabling the model to learn robust features suitable for both similarity classification and object Re-ID tasks.

- Extensive experiments on benchmark Re-ID datasets, including Market1501, CUHK03, and VeRi-776, under varying noise levels 0-30% demonstrate that our method consistently achieves state-of-the-art performance.

4.2 Related work in object Re-ID

Supervised deep learning methods typically rely on large volumes of accurately labeled data, yet in real-world settings collecting such fine-grained person Re-ID annotations is costly and often impractical. Annotations are frequently corrupted by human error, limitations of labeling tools, or dataset re-purposing, where tracking

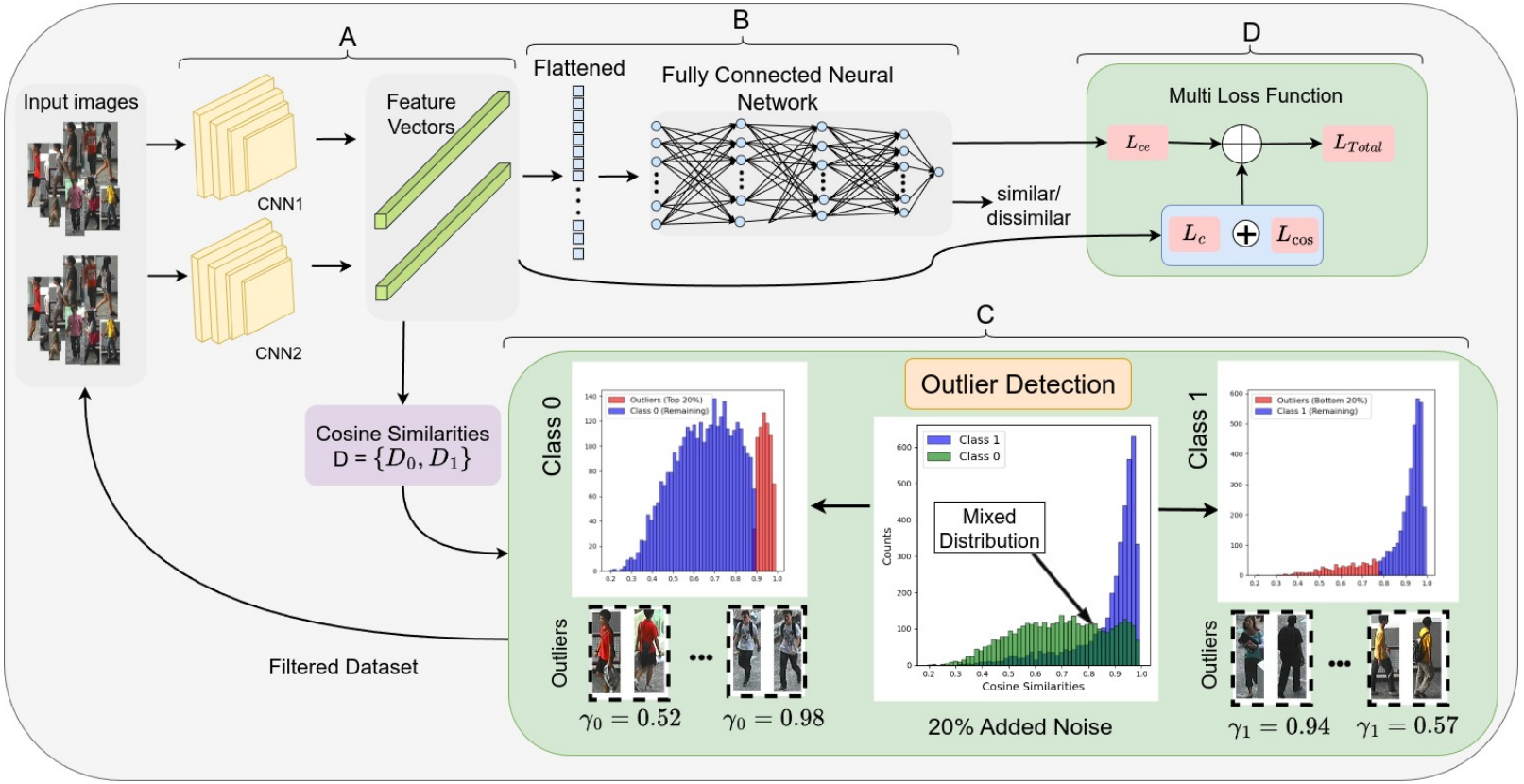


Figure 4.1: Proposed Beta-SOD Architecture: (A) The Siamese architecture extracts features from paired images. (B) Feature vectors are flattened and passed to a fully connected neural network (FCNN) for pair classification. (C) Cosine similarity is computed from the extracted features and processed by the OD module after every K epochs. The identified outliers are are then filtered from the dataset. (D) The Siamese network and FCNN are trained jointly using the combination of three loss functions.

outputs or estimated cross-camera view matches are reused across tasks, introducing identity inconsistencies and mislabeled similarity pairs (Yu et al. 2024; Ahmad and Krylov 2024). This phenomenon has been empirically documented to degrade performance even under moderate noise levels (Ye et al. 2022a). In unsupervised or domain-adaptive Re-ID pipelines, clustering-based pseudo-labels are widespread, but errors in these labels can mislead representation learning unless confidence or sample uncertainty is explicitly modeled (Miao et al. 2024). While CNNs can still learn under moderate noise, even low rates of identity-level corruption severely impair feature discriminability and generalization in Re-ID tasks (Ye et al. 2022a; Liu et al. 2024).

In this work, we focus on the noisy label setting in person Re-ID, where annotation errors caused by mislabeling, detector misalignments, or identity ambiguities can severely degrade model performance, particularly in datasets with few samples per identity. To combat this, TSNT (Liu et al. 2024) introduces a two-stage noise-tolerant paradigm that first self-refines presumably clean samples and then employs co-training with rectified cross-entropy and robust triplet losses to jointly suppress noisy labels. URNet (Li et al. 2019), though originally proposed for large-scale noisy web image classification, is relevant in this context as it adaptively reweights training instances to mitigate the effects of class imbalance, label noise, and dataset bias. PurifyNet (Ye and Yuen 2020) leverages a hard-aware instance reweighting strategy that gradually purifies noisy annotations while emphasizing challenging but reliable samples. DistributionNet (Yu et al. 2019) models feature uncertainty via per-image distributions to mitigate the impact of outliers, though its reliance on richer data per identity may limit generalization in low-sample settings. More recently, CORE (Ye et al. 2022a) applies collaborative refining via online mutual learning between peer networks with selective consistency to correct labels dynamically during training. Complementing this, Noise-Aware Re-ID (Song et al. 2024) estimates local uncertainty per sample, adaptively down-weighing unreliable data while preserving discriminative features. Additionally, KNN-based filtering techniques (Galanakis et al.

2024) detect and remove suspicious samples by examining neighborhood consistency in feature space, offering a lightweight noise-robust mechanism.

While these methods target noisy annotations in person Re-ID, similar challenges arise in vehicle Re-ID, where viewpoint changes, domain variations, and annotation inconsistencies require specialized solutions. STE-VReID (Kumar et al. 2019) proposes deep triplet embedding combined with optimized sampling strategies and achieves competitive performance with minimal complexity. To further strengthen feature learning, GS-TRE (Bai et al. 2018) introduces a group-sensitive triplet embedding that explicitly models intra-class variance, improving discrimination under diverse viewpoints. Moving beyond supervised paradigms, SGFD (Li et al. 2021) proposes a self-supervised geometric feature discovery framework with interpretable attention, removing reliance on fine-grained labels while still achieving state-of-the-art results. More recently, GiT (Shen et al. 2023) integrates graph-based local correction with transformer modules to capture both discriminative details and contextual relations, while GLSD (Chen et al. 2024) employs a multi-branch structure with global-local self-distillation to enhance representation quality through adaptive fusion. Extending this line of attention-driven methods, LSKA-ReID (Xiang et al. 2025) combines large separable kernel attention with hybrid channel attention to jointly model global structure and fine-grained details, achieving top performance on benchmarks such as VeRi-776 (Lou et al. 2019b).

4.3 Outlier Detection (OD)

Noisy labels degrade the learning capability of Re-ID models. From a statistical perspective, image pairs with incorrect or corrupted labels can be regarded as outliers within the dataset. In a typical Re-ID setup, image pairs are constructed and categorized as similar (class 1) if they belong to the same identity, or dissimilar (class 0) otherwise. Given a suitably trained Siamese architecture, similarity is quantified by computing the cosine similarity between the feature vectors extracted from the backbone architecture: $\text{sim}(x, y) = x \cdot y/(\|x\| \, \|y\|)$. Cosine similarity generally lies in the

range $[-1, 1]$, where negative values indicate opposite directions and positive values indicate similar directions in the embedding space. However, in our case, the feature embeddings are obtained after ReLU activation, which produces non-negative feature representations. As a result, the cosine similarity scores are restricted to the $[0, 1]$ range. These bounded similarity scores can therefore be effectively modeled using a two-parameter Beta distribution, defined by its density function:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \; x \in [0, 1], \tag{4.1}$$

where $\Gamma(\cdot)$ is the Gamma function, and $\alpha, \beta > 0$ are model parameters. The Beta distribution (Gupta and Nadarajah 2004) is widely used for modeling unknown quantities in the unit interval due to its flexibility in capturing diverse shapes, including uniform, skewed, U-shaped, or unimodal, thus making it well-suited for modeling the similarities. Importantly, this model allows us to avoid imposing unrealistic distribution shape assumptions, like symmetry (Gaussian), or exponential tail decay (Gamma).

In a clean consistent dataset, the histograms of the two classes are well-separated, with minimal overlap and peaks at the opposite ends of the $[0, 1]$ interval. However, as can be observed in Fig. 4.1, the noisy label case presents a mixed pattern. The outliers skew the underlying distribution, causing its mean to shift towards the center of the unit interval.

We propose to model a noisy distribution as a finite mixture of two Beta distribution components. Note that due to the similarity-based reframing of the Re-ID problem any image pair is coming from either similar or dissimilar class and hence any class histogram is a mix of the pure component plus a contamination component. We model the noisy data as

$$F_{\text{observed}}(x) = \omega_0 f(x; \alpha_0, \beta_0) + \omega_1 f(x; \alpha_1, \beta_1) \tag{4.2}$$

where (α_0, β_0) and (α_1, β_1) are the parameters of the Beta distribution for class 0

(dissimilar) and class 1 (similar), respectively, and the mixing proportions $\omega_0, \omega_1 > 0$: $\omega_0 + \omega_1 = 1$. Note that if Eq. 4.2 models class 0 [class 1] then ω_1 [ω_0] represents the fraction of outliers.

To estimate such mixtures, we adopt the EM algorithm (Dempster et al. 1977). The EM algorithm is particularly well-suited for estimating the parameters of mixture distributions, where the component memberships of observations are unknown (latent). Direct maximum likelihood estimation is intractable due to the coupling of mixture weights and the Beta distribution parameters, but EM efficiently addresses this by iteratively estimating the posterior probabilities of component assignments (E-step) and updating the mixture weights and Beta parameters (M-step). This allows for stable convergence even in the presence of noisy or overlapping components, making EM a standard and principled approach for mixture modeling in such cases. Alternative approaches include simulation-based Bayesian approaches (Diebolt and Robert 1994) and Stochastic EM, which are more costly computationally and unnecessary in the considered case, due to the observed convergence of the classical EM.

EM consists of iterating the following two steps applied to a Beta mixture model estimated on dataset D:

E-Step. The posterior probabilities for each sample x_i are computed:

$$\begin{aligned} \gamma_0(x_i) &= P(x_i \in \text{class } 0 \mid x_i; \omega, \alpha, \beta) \\ &= \frac{\omega_0 \, f(x_i; \alpha_0, \beta_0)}{\omega_0 \, f(x_i; \alpha_0, \beta_0) + \omega_1 \, f(x_i; \alpha_1, \beta_1)} \\ \gamma_1(x_i) &= P(x_i \in \text{class } 1 \mid x_i; \omega, \alpha, \beta) = 1 - \gamma_0(x_i) \end{aligned} \tag{4.3}$$

where $\omega_1 = 1 - \omega_0$ and $f(x; \alpha, \beta)$ is the density function defined in Eq. 4.1.

M-Step. The mixture parameters are estimated from the posterior probabilities γ_k. To this end, we first assign the mixture component k to each sample x_i such that

$\gamma_k(x_i) \geq 0.5$. The mixture model parameters are estimated as follows:

$$(\alpha_k, \beta_k) = \text{MLE}_{\text{Beta}}(\{x_i \mid \gamma_k(x_i) \geq 0.5\}) \quad k \in \{0, 1\} \tag{4.4}$$

where Maximum Likelihood Estimation (MLE_{Beta}) is used to estimate the parameters α_k and β_k by maximizing the log-likelihood $l(\alpha_k, \beta_k)$ based on the observed data. The maximization involves the digamma function, which does not allow for a closed-form solution, so the optimization is done numerically using the Newton-Raphson method. Specifically, starting from the trivial initialization (or that from the method of moments using sample mean and variance) we iteratively update the estimates of α_k and β_k using the first and second derivatives of the log-likelihood until convergence is achieved. The updated class prior ω_0 shown in equation 4.5, reflects the proportion of the dataset that is currently believed to the outlier class, and is computed as:

$$\omega_k = \frac{|\{x_i \mid \gamma_k(x_i) \geq 0.5\}|}{|D|}, \quad k \in \{0, 1\} \tag{4.5}$$

The noise is assumed to be uniformly distributed across both classes, with roughly 50% of the noise affecting class 1 and 50% affecting class 0. Therefore, the first run of EM is indented to identify the two large components in the whole dataset comprising the merged contaminated classes 0 and 1. Note that the noise partition does not need to be equal between the classes, it is necessary only that a representative amount of samples is present in both similar and dissimilar classes. This step allows us to use all the available data to assess the statistical behavior of the components, each presented by a substantial number of samples. To initialize, ω is set to 0.5 and $\alpha_0, \beta_0, \alpha_1, \beta_1$ are given to present a right-skewed (1,5) and a left-skewed (5,1) components, respectively. This setup helps to automatically have component 0 associated with the dissimilar class since it's peak is lower, than that of component 1, describing similar pairs and located closer to similarity 1. Beyond that, EM is not sensitive to initialization in the two-component Beta mixture case.

The EM algorithm is subsequently run separately for each of the classes. The estimated parameters from the first EM run are used as initial values. These two runs are performed on data which is dominated by the samples from one component: for instance in case of class 1, the data modeled is mostly similar pairs described by $f(x; \alpha_1, \beta_1)$, and a smaller fraction ω_0 of noisy (dissimilar) pairs. In this case $\omega_0 << \omega_1$ and hence a pure EM would typically result in a collapsed performance to $\omega_0 = 0$. This happens because the noise observations are few and do not present a consistent statistical pattern. We address this undesirable performance by fixing the parameters of the dominant component and estimating solely ω and the Beta parameters of the noise component. After running EM on class 0 and 1 we use the resulting estimates of ω_1 and ω_0 to threshold the tails of the two distributions, and eliminate the corresponding image pairs from the training data. This constitutes our OD strategy.

Integration with CNN. We alternate between Siamese training and the OD step, consisting of three EM runs (see Alg. 2). The initial fine-tuning (prior to the first round of OD) and the later fine-tuning epochs help ensure that hard inliers are learned by the model and are not prematurely filtered out along with the noisy samples. The training process is terminated after a fixed number of cycles C. The overwhelming amount of filtering occurs after cycle 1, with the filtered fraction dropping by a factor of 2-5 with each additional cycle of training. Fig. 4.1 illustrates the impact of label noise on the similarity score histograms, along with the corresponding detected outliers. The image pairs shown represent samples identified as outliers by the EM-based filtering method, with their respective probabilities indicating likelihood of belonging to each class. This visualization highlights how noise skews the score distribution and how the proposed method effectively isolates mislabeled pairs.

Distribution choice for Cosine Similarities. The proposed method is based on the Beta distribution, since it provides the most flexible natural fit to the cosine similarity quantities in range $[0, 1]$. Fig. 4.2 shows the quality of fit of this distribution to both classes similarities on the CUHK03 dataset under 0-20% levels of random noise.

Algorithm 1: Outlier Detection (OD) Algorithm for Noisy Image Pairs Removal

Function `EM(`D, α_k, β_k, $Comp$`)`**:**
- **Initialize:** $\omega = \{0.5, 0.5\}$; $\Delta\omega = 1$;
- **while** $\Delta\omega \geqslant 0$ **do**
  - **E-Step:** Compute γ_0, γ_1 in D using Eqs. 4.1, 4.3;
  - **M-Step:** Compute ω_k^{new}, α_k^{new}, β_k^{new} in D using Eqs. 4.4, 4.5;
  - `// Update both mixture weights`
  - $\Delta\omega = \|\omega_k^{new} - \omega_k\|$; $\omega_k \leftarrow \omega_k^{new}$;
  - **if** $0 \in Comp$ **then**
    - `// Update component 0`
    - $\alpha_0 \leftarrow \alpha_0^{new}$; $\beta_0 \leftarrow \beta_0^{new}$;
  - **if** $1 \in Comp$ **then**
    - `// Update component 1`
    - $\alpha_1 \leftarrow \alpha_1^{new}$; $\beta_1 \leftarrow \beta_1^{new}$;
- **return** ω_k, α_k, β_k;

Input: Datasets $D_s = \{x_i, x_j, s\}_{i,j}^{N_s}$, with $s = 0$ dissimilar, and $s = 1$ similar image pairs x_i, x_j;

Import pre-trained model;

for $cycle = 1$ *to* C **do**
- Train Siamese model on $D = \{D_0, D_1\}$ for K epochs;
- Compute cosine similarities from the Siamese model;
- `// Learn the outline of components from all data`
- ω_k^*, α_k, $\beta_k \leftarrow$ `EM` $(\{D_0, D_1\}, \alpha = \{1, 5\}, \beta = \{5, 1\}, \{0, 1\})$;
- `// Filter the outliers separately in` D_0 `and` D_1
- `// Learn only component 1 (component 0 frozen)`
- ω_k, α_k^*, $\beta_k^* \leftarrow$ `EM` $(\{D_0\}, \alpha_k, \beta_k, \{1\})$;
- Update D_0 by filtering out highest $\omega_1|D_0|$ samples (outliers) based on cosine similarities;
- `// Learn only component 0 (component 1 frozen)`
- ω_k, α_k^*, $\beta_k^* \leftarrow$ `EM` $(\{D_1\}, \alpha_k, \beta_k, \{0\})$;
- Update D_1 by filtering out lowest $\omega_0|D_1|$ samples (outliers) based on cosine similarities;

Output 1: Filtered $D = \{D_0, D_1\}$ cleaned of similarity-based outliers;

Output 2: Trained Siamese architecture (CK epochs) for Object Re-ID.

We also overlay the fitted probability density functions for Gaussian, and Gamma distributions to evaluate their suitability for modeling the data. The fitting has been achieved following a similar EM procedure relying on MLE to obtain corresponding distribution parameter estimates in the M-step. These two classical families of two-parametric distributions provide worse quality fit largely due to unbounded support: $x \in \mathbb{R}$ for Gaussian, and $x \in \mathbb{R}_{>0}$ for Gamma. As can be observed in Fig. 4.2, the increasing noise levels push the two similarity distributions closer together, with the Beta distribution consistently providing better fit, highlighting its robustness for modeling pairwise similarity distributions in noisy settings. Importantly, Gaussian and Gamma mixtures are known to allow the mixing distributions' parameters to be identified uniquely from the observed mixtures (Atienza et al. 2006), and in Sec. 4.4 below we prove for the first time that the same can also be said for the Beta mixtures considered in this chapter.

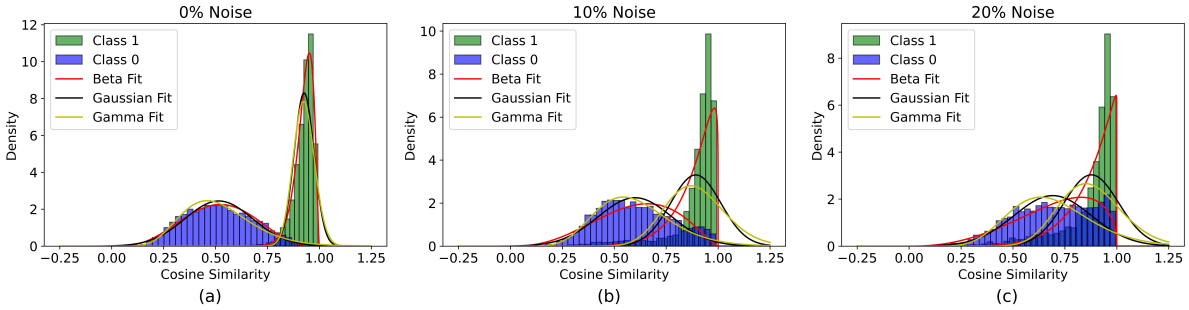


Figure 4.2: Histogram of cosine similarities for both similarity classes on the CUHK03 dataset with overlaid estimated Beta, Gaussian, and Gamma two-component mixtures: with (a) 0%, (b) 10% and (c) 20% random label noise. Unlike Gaussian and Gamma, the Beta distribution models the data in the target range $[0, 1]$.

4.4 Beta mixtures identifiability

We will demonstrate here that the identification of mixture components designed in Sec. 4.3 is a correctly posed problem in that a unique composition of two Beta distribution components can be recovered. This can be achieved in Theorem 1 below:

Theorem 1. For $k \geq 1$, let $(\alpha, \beta, \pi) \in \Theta_k$, and define the mixture density as

$$p(\cdot) = \sum_{i=1}^{k} \pi_i \, f(\cdot; \alpha_i, \beta_i). \tag{4.6}$$

If $\alpha_i - \alpha_j \notin \mathbb{Z}$ for all $i \neq j$, or $\beta_i - \beta_j \notin \mathbb{Z}$ for all $i \neq j$, then the mixture p is identifiable.

Proof (outline).

Without loss of generality, assume that $\alpha_i - \alpha_j \notin \mathbb{Z}$ for all $i \neq j$. We begin by considering the Mellin transform of the mixture density p, and study the poles of its analytic continuation along the real axis. Each parameter α_i gives rise to a sequence of poles that are spaced at unit intervals. Because of the condition $\alpha_i - \alpha_j \notin \mathbb{Z}$, the pole sequences corresponding to different components do not overlap, and therefore remain disjoint. This separation of poles plays a key role. Since each sequence can be uniquely associated with a specific α_i, the parameters $\alpha_1, \ldots, \alpha_k$ can be directly identified from the locations of these poles.

Once the α_i parameters are determined, we proceed by examining the complex residues of the Mellin transform at the first two poles of each sequence, namely $\mathrm{Res}(p; 1 - \alpha_i)$ and $\mathrm{Res}(p; -\alpha_i)$. These residues can be explicitly computed and provide a system of equations associated with each component. Solving this system yields the corresponding parameters β_i and π_i. Since both the pole locations and the associated residues uniquely determine all the parameters $\{\alpha_i, \beta_i, \pi_i\}_{i=1}^{k}$, it follows that the mixture density p is identifiable.

The proof shows that the individual components of a two-component Beta mixture can be uniquely recovered from the observed mixture with probability 1. This result is novel and departs from established identifiability proofs for more classical Gaussian, Gamma (Atienza et al. 2006), and Generalized Gamma mixtures (Li et al. 2015). In the case of Beta distributions, the required property may fail only on subsets of Lebesgue measure zero, which necessitates the distinct line of argument introduced here. Since these pathological cases occur with probability 0 under continuous parameter sampling, they do not affect the practical identifiability of

the model. This statement follows directly from Theorem 1, as the non-identifiable parameter sets have lower dimensionality than the full parameter space, and there are only countably many such lower-dimensional subspaces. Therefore, under any absolutely continuous measure, the total probability of the non-identifiable settings is 0. Equivalently, the set of uniquely identifiable parameter settings has probability 1. Further details are provided in the preprint version of this work.[2]

4.5 Siamese architecture and Loss Functions

When dealing with Siamese architectures, contrastive loss is commonly employed, where pairs yielding lower loss values are treated as similar, while those with higher loss values are treated as dissimilar (Jin et al. 2021; Shorfuzzaman and Hossain 2021; Wang et al. 2021). Alternatively, similarity can be assessed using cosine similarity between feature vectors, followed by classification using either a predefined threshold or a learnable linear layer (Chen et al. 2020).

In this work, rather than relying solely on contrastive loss or cosine similarity-based loss, we employ a combination of loss functions to optimize the model parameters more effectively. The combination includes BCE loss ($\mathcal{L}_{CE}$), cosine-embedding loss ($\mathcal{L}_{COS}$) and contrastive loss ($\mathcal{L}_{C}$). The cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{4.7}$$

is obtained from a fully connected neural net (FCNN). The FCNN takes the 1280-dimensional vectors as an input and includes four fully connected layers with a binary output layer.

The contrastive loss:

$$\mathcal{L}_{C} = \frac{1}{N}\sum_{i=1}^{N}\left[y_i d_i^2 + (1-y_i)\max(0, m-d_i)^2\right], \tag{4.8}$$

[2]Preprint: `https://arxiv.org/abs/2509.08926`

where d_i is the euclidean distance and m is the margin parameter, is effective when class boundaries are well separated. However, in applications such as person Re-ID, visually similar individuals, for instance, two people viewed from behind wearing the same clothing color, may produce low Euclidean distances, leading to potential misidentification. This arises because contrastive loss alone fails to capture subtle appearance differences that are not well represented by Euclidean distance.

To address this, we combine contrastive loss with cosine embedding loss, which is more effective at distinguishing such instances by focusing on angular similarity rather than just distance.

$$\mathcal{L}_{\text{COS}} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 1-\cos(\mathbf{x}_i,\mathbf{y}_i), & \text{if similar} \\ \max(0,\cos(\mathbf{x}_i,\mathbf{y}_i)-m), & \text{otherwise} \end{cases} \tag{4.9}$$

where $\mathbf{x}_i, \mathbf{y}_i$ represent the feature vectors of a similar/dissimilar image pair. $\cos(\mathbf{x}_i, \mathbf{y}_i)$ is the cosine similarity between feature vectors and m is the margin value, controlling the penalty for high similarity. $\mathcal{L}_{\text{COS}}$ is beneficial when comparing images that are similar in terms of appearance but not in spatial orientation / pose.

The total loss $\mathcal{L}_{total}$ is the weighted sum of three loss functions:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{COS} + (1-\lambda)\mathcal{L}_{C} \tag{4.10}$$

where λ is the weighting hyper-parameters that balances the contribution of the loss terms.

Here, $\mathcal{L}_{CE}$ is retained as the primary supervised classification loss, while $\mathcal{L}_{COS}$ and $\mathcal{L}_{C}$ act as complementary metric-learning losses on the learned embedding space. During training, it was observed that the cross-entropy loss converged relatively quickly, whereas the cosine embedding and contrastive losses remained higher and converged more slowly. Therefore, instead of assigning separate weights to all three losses, the cross-entropy term is kept fixed, and the weighting is applied only between the two metric-learning terms.

Using a single parameter λ provides a controlled trade-off between angular similarity and Euclidean-distance-based separation. This is appropriate because both $\mathcal{L}_{COS}$ and $\mathcal{L}_{C}$ influence the same feature embedding space and are therefore not fully independent. Increasing the contribution of the cosine embedding loss naturally reduces the relative contribution of the contrastive loss, and vice versa. In this way, λ directly controls the balance between orientation-based similarity and distance-based separation without introducing an additional hyperparameter.

The single-parameter formulation is restrictive, but it avoids overparameterization. If two independent weighting parameters were used, the optimization would require tuning both the absolute scale and the relative contribution of the two metric losses, which could make training less stable and harder to interpret. By using λ and $1 - \lambda$, the combined contribution of the metric-learning losses remains normalized, while the model can still adjust whether cosine similarity or contrastive separation should have greater influence during training.

The overall architecture of the Siamese network with the proposed losses and the OD module is shown in Fig. 4.1. The OD module is performing filtering of the two datasets with dissimilar and similar image pairs after every K epochs such that the outliers are being identified and filtered out in a task-specific manner.

4.6 Experiments

4.6.1 Experimental Setup

Evaluation Metrics. In our work, we focus on noisy-label object Re-ID but also address noise detection. The proposed architecture consists of two components—MobileNetV2 for feature extraction and a FCNN for classifying image pairs similarity. To evaluate the performance of the model, we employ distinct metrics for each task. For classification, we report standard evaluation metrics including test accuracy, precision, and recall. For object Re-ID, we utilize the rank-k matching accuracy and the mean Average Precision (mAP).

The mAP metric provides a comprehensive evaluation of Re-ID performance by averaging the precision values across all queries. The Average Precision (AP) for each query, and hence the overall mAP are obtained as follows:

$$AP = \frac{\sum_{k=1}^{n} P(k)gt(k)}{N_{gt}}, \qquad mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q},$$

where, k denotes the rank position among the n retrieved objects, N_{gt} represents the overall number of relevant (ground-truth) objects, and Q is the total number of query images. $P(k)$ denotes the precision at rank k in the retrieval list, while $gt(k)$ indicates if the k-th retrieved image is a correct match or not.

Rank-k (Rk) measures the proportion of queries for which the correct match appears within the top-k retrieved results:

$$Rk = \frac{\sum_{q=1}^{Q} gt(q,k)}{Q},$$

where $gt(q,k)$ equals 1 when ground truth of q image appears in the top k retrieved samples.

Datasets. The proposed methodology is evaluated on three multi-view Re-ID datasets: CUHK03 (Li et al. 2014) and Market1501 (Zheng et al. 2015) for person Re-ID, and VeRI-776 (Lou et al. 2019b) for cars Re-ID.

CUHK03 dataset contains 14,097 images of 1,467 identities captured by 5 disjoint camera views. The images are divided into training set with 7,365 images, gallery set with 5,332 images and query set with 1400 image samples. Two type of annotation is provided i.e., manually labeled and detected bounding boxes. We use the latter to enable comparisons with the benchmark methods.

Market1501 dataset contains 32,668 images captured by 6 disjoint camera views. The images set is further divided into training+gallery sets containing 12,936+19,732 images of 751 identities, and the query set containing 3,368 images of another 750 identities.

VeRi-776 dataset contains 51,035 images of 776 vehicles. Each vehicle is captured

by up to 18 disjoint camera views. The dataset is provided with different types of annotations including bounding boxes, vehicle types, colors, brands and cross-camera relations. The dataset is further divided into training set with 37,778 images, gallery set with 11,579 images and query set with 1678 images.

As the proposed method classifies image pairs into similar and dissimilar categories, we begin by generating image pairs across all identities and assigning them to similar class (label 1). An equal number of dissimilar pairs (label 0) are created by randomly selecting images from different identities and pairing them together.

Label Noise Generation. We evaluate our methodology under the following two core noise settings, each applying the equal target amount of corrupt labels to both image pairs classes, similar and dissimilar. Firstly, we generate *Random Noise* datasets. Similarly to (Yu et al. 2019; Ye and Yuen 2020; Liu et al. 2024) we introduce noise by randomly (uniformly) selecting images from the dataset and reassigning their labels to different (random) identities. Specifically, class 0 (dissimilar) label noise is introduced by pairing images from the same identity while label noise in class 1 is generated by pairing images from distinct identities. Secondly, we consider *Pattern Noise*, where the datasets is curated to render denoising more challenging (Song et al. 2024; Liu et al. 2024). To do so, we trained a ResNet-50 model on a clean dataset. This network was then used to extract feature representations of all image samples. To generate pattern noise labels, we paired samples from different identities that exhibited highest cosine similarity and labeled them as similar pairs (class 1). Conversely, we paired samples of the same identity with lowest cosine similarity (originally belonging to class 1) and mislabeled these as class 0.

Implementation Details. The proposed methodology is implemented in PyTorch (Python) using a Siamese network with a pre-trained MobileNetV2 (Sandler et al. 2018) as the backbone. The similarity FCNN head consists of four fully connected layers with 512, 512, 256, and 128 neurons, respectively, each followed by Batch Normalization and ReLU activation. Dropout with a rate of 0.3 is applied after the final

two fully connected layers to reduce overfitting. Data augmentation includes random cropping and horizontal flipping, with image resolutions of 300×150, 128×64, and 150×150 for the CUHK03, Market1501, and VeRi-776 datasets, respectively. Training is performed using the Adam optimizer with an initial learning rate of 0.001, a batch size of 32, and learning-rate decay applied every 5 epochs. All experiments are conducted on an NVIDIA Titan X GPU with 12 GB of memory. The number of epochs K is determined based on validation accuracy; once the validation accuracy stabilizes, the OD process is initiated. We have observed little sensitivity to this parameter having considered a range of 2-8 epochs. Parameter C depends on w_k, and OD is terminated when w_k drops below 0.01%. In general, the selection of K and C should allow for a similar overall number of training epochs CK as the original architecture suggests. The combined loss is used to update model parameters, with $\lambda = 0.45$ tuned via grid search.

From a computational perspective, the cost of running Beta-SOD is essentially equivalent to training the underlying Siamese backbone and subsequent FCNN. The additional overhead of the EM step for noise identification and filtering is minimal, requiring less than 10 milliseconds on a CPU. Since the OD module is relevant only for the training phase, the deployment consists in simply running the Siamese backbone + FCNN.

4.6.2 Results & Comparisons

In this section we present the outcomes of the noise detection, as well as the person and vehicle Re-ID task. We also report comparisons with alternative statistical models for the OD module and ablation studies.

The technique proposed in this work relies on explicit statistical modeling under the assumption that noise represents a contamination to the clean signal component. For this reason, we focus on Beta-SOD evaluation with noise levels at 0-30%, where from statistical point of view, the "true" distribution is dominant. At a 50% noise level, the histograms of the two classes are practically identical (in the random

Table 4.1: Siamese Backbone Network selection on CUHK03 Dataset with various random noise levels (5 runs average)

Backbone	Parameters	Noise	mAP	R1	R5	R10
Resnet18	∼4M	0%	50.94	52.71	65.18	75.2
		10%	38.1	43.31	56.18	70.92
		30%	31.8	35.4	54.92	68.8
Resnet50	∼23.9M	0%	58.9	**58.7**	**74.61**	79.7
		10%	**48.91**	52.1	70.9	77.7
		30%	45.51	50.14	70.21	**77.1**
MobileNet-V2	∼3.4M	0%	**59.2**	58.56	74.58	**79.82**
		10%	48.33	**52.61**	**71.82**	**77.75**
		30%	**46.8**	**50.2**	**71.01**	76.9

noise setting), making it unfeasible for the proposed statistical method to perform effectively without additional supervision. We hence choose not to consider such scenarios of noise at 50% or higher, which require structural modifications and result in relatively poor performance, e.g., under 20% mAP on the CUHK03 dataset (Liu et al. 2024). Furthermore, we consider such extreme contamination levels to be uncommon in practice, and thus of less interest.

Siamese backbone selection. In Tab. 4.1 we present comparisons between several typical Siamese architecture backbones: ResNet-18, ResNet-50 (He et al. 2016), and MobileNetV2 (Sandler et al. 2018). Given its size and very competitive accuracy reported, we employ MobileNetV2 as the primary architecture for our model.

Table 4.2: Noise identification performance with various random noise levels (5 runs average)

Dataset	Noise	Outliers	Precision	Recall	Accuracy
Market1501	0%	0.5	-	-	94.8
	10%	10.31	85.79	76.87	92.1
	20%	21.26	82.93	80.56	91.8
	30%	32.97	75.73	81.67	90.7
CUHK03	0%	1.62	-	-	89.5
	10%	11.32	82.09	73.63	87.21
	20%	21.56	80.47	76.08	85.1
	30%	32.28	75.75	81.18	84.2
VeRi-776	0%	1.3	-	-	93.89
	10%	11.34	87.46	84.8	92.1
	20%	22.76	85.66	88.39	90.83
	30%	33.05	84.47	89.86	89.76

Table 4.3: Noise identification performance of OD using Gaussian, Gamma, and Beta mixtures under varying noise levels on the CUHK03 dataset (5-run average)

Model	Noise	Outliers	Precision	Recall	Accuracy
Gaussian mixture	0%	2.13	-	-	87.71
	10%	13.18	76.01	71.99	85.62
	20%	24.14	72.15	75.7	81.55
Gamma mixture	0%	4.51	-	-	87.3
	10%	13.09	72.4	70.95	83.92
	20%	25.06	64.42	75.93	77.6
Beta mixture	0%	1.62	-	-	89.5
	10%	11.33	82.28	73.88	87.12
	20%	21.56	80.32	76.19	85.4

Noise filtering performance. We first assess the OD module's performance in identifying the contamination, i.e. incorrectly labeled (noisy) image pairs, in the dataset of cosine similarities. As reported in Tab. 4.2, the proposed model achieves 94.8%, 89.5%, and 93.89% similarity classification accuracy on clean Market1501, CUHK03, and VeRi-776 datasets. At 30% noise level, these accuracies decline by 4.1%, 5.3%, and 4.13%. Increasing the noise level from 10% to 30% results in slightly more detected outliers, indicating that the proposed denoising method effectively filters noisy pairs with a tendency towards overshooting. At higher noise levels (30%), precision decreases slightly due to more false positives, while recall improves as more true positives are retained and fewer false negatives occur. Nevertheless, precision remains around 76% for Market1501 and CUHK03, and 84% for VeRi-776, demonstrating the robustness of the proposed method against label noise. This degradation can be primarily attributed to the removal of challenging positive or negative examples during the denoising process, which somewhat reduces the model's ability to generalize. A small subset of challenging noisy samples persist after filtering, further contributing to the similarity classification decline.

Distribution modeling. To further evaluate the suitability of the Beta distribution for modeling cosine similarities, we compare the proposed Beta model with Gamma and Gaussian distributions for the OD on the CUHK03 dataset, which is summarized in Table 4.3.The Beta distribution achieves the highest classification accuracy of 89.5% on the clean dataset, whereas the relatively lower scores of the

Gaussian and Gamma distributions. Moreover, as the noise level increases to 20%, the amount of over-filtering reaches 4–5% for Gamma and Gaussian distributions, whereas the Beta model shows substantially lower overshooting of only 1–2%, confirming its robustness in noisy conditions. Qualitatively, in Fig. 4.2 we can observe that both Gaussian and Gamma distributions do not fit the data as effectively as the Beta distribution. This over-filtering of challenging pairs negatively affects dataset composition and, consequently, reduces model generalization. Precision and recall scores in Tab. 4.3 further highlight these differences: at 20% noise level, the Gamma distribution filters a large number of false positives, resulting in a lower precision of around 65%, whereas the Beta distribution maintains a higher precision of over 80% compared to both alternatives. This empirically supports our assumption that the Beta distribution is a more appropriate model due to both its flexibility and native unit range support.

Person Re-ID Results. We compare our method with several state-of-the-art approaches: Noise-Aware Person Re-IDentification (Song et al. 2024), DistributionNet (Yu et al. 2019), PurifyNet (Ye and Yuen 2020), CORE (Ye et al. 2022a), URNet (Li et al. 2019), and TSTN (Liu et al. 2024). Tab. 4.4 presents the comparison under three different noise ratios on CUHK03 and Market1501 datasets. The results demonstrate that the proposed method consistently outperforms the other state-of-the-art methods across the three noise levels. As shown in the table, while the performance of Re-ID deteriorates significantly as the noise level increases from to 30%, the proposed method maintains comparatively superior performance. We can also observe the relative contribution of the OD module by comparing with the same Siamese architecture without OD. In particular, on the clean dataset the performance is very similar (especially taking into account the standard deviation measured across the 5 re-runs). With added noise, the use of the OD module results in a very significant performance gain.

DistributionNet, originally designed for noisy image classification, filters out noisy samples but can discard significant data when noise exceeds 20%, harming per-

Table 4.4: Comparison of Person Re-ID performance of Beta-SOD against Noise-Aware Re-Id (Song et al. 2024), DistributionNet (Yu et al. 2019), PurifyNet (Ye and Yuen 2020), CORE (Ye et al. 2022a), URNet (Li et al. 2019), and TSNT (Liu et al. 2024) on the Market1501 and CUHK03 datasets with various random noise levels (5 runs Average and Standard Deviation)

Noise	Methods	Market1501				CUHK03			
		mAP	R1	R5	R10	mAP	R1	R5	R10
0%	DistributionNet	70.8	87.3	94.7	96.7	38.5	39.4	58.9	67.9
	URNet	71.2	87.9	95.7	97.4	43.2	45.7	64.4	74.6
	PurifyNet	72.1	88.4	95.8	97.6	44.8	45.8	65.2	74.2
	CORE	74.6	89.6	96.4	98.1	46.2	46.3	65.5	75.2
	TSNT	80	92.7	97.1	98.2	54.3	55	73.4	81.3
	Noise-Aware Re-Id	84.4	94.4	98.1	99	50.6	51.7	72.1	81
	Siamese without OD	**85.5** ±0.9	94.44 ±0.62	**98.81** ±0.32	98.57 ±0.02	**57.1** ±1.32	**59.57** ±0.93	74.11 ±0.81	**81.52** ±0.64
	Beta-SOD	84.5 ±1.2	**94.51** ±0.71	98.35 ±0.32	**98.83** ±0.02	56.56 ±1.5	59.2 ±1.1	**74.58** ±0.91	79.82 ±0.72
10%	DistributionNet	63.36	89.85	96.25	98.02	31.8	32.29	51.81	61.67
	PurifyNet	64.3	84.2	93.7	95.6	32.8	32.8	54.8	65
	CORE	67.7	85.5	94	96.4	39.6	40.4	60.8	70.9
	Noise-Aware Re-Id	69.9	88.8	95.4	97.4	42.4	43.5	64.8	74
	Siamese without OD	59.21 ±1.12	75.92 ±0.91	89.34 ±0.51	92.11 ±0.05	29.01 ±2.2	32.44 ±1.2	48.1 ±1.1	60.91 ±1
	Beta-SOD	**80.36** ±1.43	**92.26** ±0.92	**97** ±0.49	**98** ±0.03	**48.33** ±1.9	**52.61** ±1.3	**71.82** ±0.95	**77.75** ±0.77
20%	DistributionNet	53.4	77	90.6	94	24.2	24.32	43.1	53.1
	URNet	52.9	76.0	90.0	93.4	21.9	23.1	41.6	52.1
	PurifyNet	63.1	83.1	93.3	95.9	29.2	30.3	49.8	58.4
	CORE	66.2	84.1	93.1	95.5	35	34.4	55	64.1
	TSNT	76	90.3	96.4	**97.7**	**48.3**	49.4	67.9	76.4
	Noise-Aware Re-Id	65.5	86.2	95.7	97.6	29.2	30.9	50.6	62
	Siamese without OD	50.1 ±1.51	72.71 ±1.1	85.55 ±1	92 ±1	23.21 ±2.3	25.3 ±1.5	40.58 ±1.25	50.01 ±1.21
	Beta-SOD	**78.32** ±1.62	**91** ±1.39	**97.11** ±0.61	97.61 ±0.10	48.12 ±2.1	**51.25** ±1.32	**71.11** ±1	**77.75** ±0.78
30%	PurifyNet	60.2	81.4	92.3	94.9	26.4	26.6	44.6	54.2
	CORE	60.6	81.2	92.2	94.5	34.7	34.6	56.4	64.6
	Noise-Aware Re-Id	54.5	79.1	92.5	95.2	17.3	18.1	34.8	44.1
	Siamese without OD	44.66 ±1.97	65.15 ±1.5	78.83 ±1.1	82 ±1.13	15.71 ±2.5	19.19 ±1.51	30.84 ±1.31	42.4 ±1.27
	Beta-SOD	**66.01** ±1.91	**85.32** ±1.52	**93.26** ±1.1	**95.64** ±0.27	**46.8** ±2.1	**49.88** ±1.29	**65.71** ±1.1	**76.9** ±0.78

*We show only the available benchmark results as reported in the corresponding papers

Table 4.5: Comparison of Person Re-ID performance of Beta-SOD against PurifyNet (Ye and Yuen 2020), CORE (Ye et al. 2022a), and TSNT (Liu et al. 2024) on the Market1501 and CUHK03 datasets under various **pattern noise** levels (5-run average and standard deviation)

Noise	Dataset	Method	mAP	R1	R5	R10
10%	Market1501	PurifyNet	63.2	81.8	92.6	95.4
		CORE	66.4	84	93.8	96.1
		Beta-SOD	**68.57 ± 0.6**	**87.21 ± 0.07**	**95.23 ± 0.08**	**97.22 ± 0.05**
	CUHK03	PurifyNet	32.9	33.6	56.4	66
		CORE	32.2	38.2	54.9	64.8
		Beta-SOD	**33.92 ± 0.53**	**41.65 ± 0.06**	**59.29 ± 0.07**	**69.66 ± 0.06**
20%	Market1501	PurifyNet	56.2	77.8	91.1	93.8
		CORE	62.8	81.9	92.7	95.5
		TSNT	74.8	89	95.3	97.2
		Beta-SOD	**76.89 ± 0.92**	**89.58 ± 0.08**	**96.82 ± 0.05**	**97.67 ± 0.04**
	CUHK03	PurifyNet	27.9	28.8	47.3	58.2
		CORE	33.2	34.6	54.6	64.9
		TSNT	37.2	38	57.4	68.8
		Beta-SOD	**38.1 ± 0.82**	**40.31 ± 0.07**	**59.55 ± 0.07**	**71.08 ± 0.05**

*We show only the available benchmark results as reported in the corresponding papers

Table 4.6: Person Re-ID performance of the OD method using Gaussian, Gamma, and Beta mixture models on the CUHK03 dataset under varying noise levels (5-run average)

Model	Noise	mAP	R1	R5	R10
Gaussian mixtures	0%	50.2	57.33	71.21	78.36
	10%	44.74	47.32	63.39	74.9
	20%	43.84	45.68	63	73.7
Gamma mixtures	0%	46.92	54.41	65.22	71.29
	10%	38.21	50.1	60.2	68.5
	20%	34.41	42.4	54.1	60.85
Beta mixtures	0%	56.56	59.2	74.58	79.82
	10%	48.33	52.61	71.82	77.75
	20%	48.12	51.25	71.11	77.75

formance, especially on CUHK03 with limited images per identity. In contrast, Re-ID-specific methods like PurifyNet, CORE and TSNT down-weight noisy samples instead of removing them. However, this still suppresses useful data and degrades performance as noise increases. Similarly to DistributionNet, our approach adopts a (hard) filtering strategy, adapted for a pair-based setting. Since each image forms multiple cross-view pairs, filtering out noisy pairs still retains sufficient representative samples per identity. This redundancy preserves data diversity and maintains model generalization.

Tab. 4.6 demonstrates how the choice of the OD mixture distribution impacts the Re-ID task performance on the CUHK03 dataset at varying random noise levels 0-20%. Across all noise settings, the Beta distribution consistently outperforms both Gaussian and Gamma models in terms of mAP and Rank-k accuracy. As the noise level increases to 20%, the Beta maintains its superiority achieving an mAP of 48.12%, compared to 34-44% for Gamma and Gaussian mixtures. These results confirm the earlier observation based on Tab. 4.3 that the Beta model provides a more appropriate fit for cosine similarities.

Tab. 4.5 compares the Re-ID performance of Beta-SOD on the Market1501 and CUHK03 datasets under pattern noise at 10-20%. At 20% noise level, the relative improvement achieved by Beta-SOD becomes more pronounced. On Market1501, our approach attains 76.89% mAP and 89.58% Rank-1 accuracy, surpassing the second best, TSNT, by 2% mAP and 0.5% Rank-1, while also maintaining consistently higher Rank-5 and Rank-10 results. Similarly, on the more challenging CUHK03 dateset, the relative improvement over TSNT is of 0.9% mAP and 2.3% Rank-1. To provide a measure of statistical relevance, we also report the standard deviations available for our experiments (over 5 re-runs), which confirm the meaningful nature of these improvements. Overall, the results confirm the robustness of our OD-based framework, which performs well in presence of both random (realistic setting) and pattern (adversarial setting) noise.

Mislabeled examples. Fig. 4.3 presents examples of image pairs incorrectly handled by the OD method during denoising. The first row shows false positives, i.e., similar image pairs that were wrongly filtered out by the OD method, with cosine similarity scores below 0.8. The second row shows false negatives, i.e., noisy image pairs that were mistakenly retained as clean, with cosine similarities above 0.6. These examples illustrate the difficulty of distinguishing visually similar pairs, where both under-filtering (false negatives) and over-filtering (false positives) compromise the quality of the denoised dataset. These and conceptually similar challenging examples are treated incorrectly by the OD module even at 0% noise, see in Tab 4.2.

At the same time, the overall impact achieved by the OD-based filtering in the person Re-ID task ranges from a modest performance reduction of 0.5-1% mAP at 0% noise to substantial improvement of 21-31% mAP at 30% noise levels, see Tab 4.4.

Vehicle Re-ID Results. We now evaluate Beta-SOD on vehicle Re-ID using VeRi-776 dataset. As shown in Tab. 4.7, our method achieves high mAP and R1/R5 scores on clean dataset (0% noise), only behind that of specialized (Chen et al. 2024; Xiang et al. 2025). Interestingly, the configuration without the OD module attains the level not far behind the top performer without attention mechanism incorporated. As established in the de-noising comparisons, without the presence of noise, the OD module removes more challenging training examples which reduces the test performance. This results in a drop of mAP by 1.2%.

To evaluate the model's performance under noisy conditions, random label noise is introduced into the vehicle Re-ID dataset. Tab. 4.7 reports performance with and without the Beta-SOD. On the clean dataset, the model achieves 83% mAP, 97.22% Rank-1, and 98.1% Rank-5 accuracy. With added noise, performance without the Beta-SOD filtering drops significantly, but is effectively handled with filtering. No benchmark methods report their performance in the presence of noisy labels.

These results confirm the ability of the proposed technique to generalize to another object Re-ID task without modifications.

Ablation Analysis. Finally, we investigate the impact of the considered loss functions on the model performance when using the CUHK03 dataset. This involves a comparative analysis between a basic classification loss and similarity-based constraints. The experimental results are presented in Tab. 4.8.

We establish a baseline Siamese architecture comprising the FCNN trained with the ($\mathcal{L}_{CE}$) loss. Using only $\mathcal{L}_{CE}$ our model achieves a mAP of 47.41% on the clean dataset and 36.12% in the presence of 20% label noise. Adding $\mathcal{L}_{cos}$ to $\mathcal{L}_{CE}$ improves the mAP by just over 4% on both the clean and the 20% noisy datasets. Similarly, combining $\mathcal{L}_c$ with $\mathcal{L}_{CE}$ results in a mAP improvement of about 3-5% across both



Figure 4.3: Examples of incorrectly filtered image pairs by Beta-SOD on CUHK03 with 20% noise: false positives, i.e. true class 1 flagged as 0 (top row), and false negatives, i.e. true class 0 flagged as 1 (bottom row). The numbers are cosine similarities: below 0.8 for false positives and above 0.6 for false negatives.

Table 4.7: Comparison of Vehicle Re-ID performance on the VeRi-776 dataset with various random noise levels (5 runs Average)

Noise	Methods	mAP	R1	R5
0%	GS-TRE (Bai et al. 2018)	59.47	96.24	98.97
	STE-VReID (Kumar et al. 2019)	67.55	90.23	96.42
	GiT (Shen et al. 2023)	80.3	96.9	-
	SGFD (Li et al. 2021)	81	96.7	98.6
	GLSD (Chen et al. 2024)	83.5	97.3	98.6
	LSKA-ReID (Xiang et al. 2025)	**85.64**	97.97	**99.04**
	Siamese without OD	84.2	**98.13**	98.8
	Beta-SOD	83	97.22	98.1
10%	Siamese without OD	56.25	88.12	91.45
	Beta-SOD	**78**	**97**	**98**
20%	Siamese without OD	52.22	84.92	90.11
	Beta-SOD	**76.26**	**96.13**	**97.62**
30%	Siamese without OD	49.41	83.52	88.21
	Beta-SOD	**74.21**	**95.93**	**97.54**

Table 4.8: Loss Function Selection on CUHK03 Dataset with random noise level of 0% and 20% (5 runs average)

Loss Functions			0% noise			20% noise		
$\mathcal{L}_{CE}$	$\mathcal{L}_{COS}$	$\mathcal{L}_{C}$	mAP	R1	R5	mAP	R1	R5
✓			47.41	50.9	63.13	36.12	40.3	48.21
✓	✓		51.82	55.3	70.22	40.2	45.01	58.3
✓		✓	54.72	57.1	73.1	45.8	50.13	68.83
✓	✓	✓	56.52	59.2	74.58	48.12	51.25	71.11

settings. Finally, by combining all three loss functions, as described in Eq. 4.10, the mAP improves by an additional 2%.

The effect of the weighting parameter λ in Eq. 4.10 was evaluated by varying its value between 0 and 1, as shown in Fig. 4.4. The best overall performance was obtained at $\lambda = 0.45$, achieving a validation accuracy of 89.0%, an mAP of 48.12%, and a Rank-1 accuracy of 51.25%. Performance decreased when λ was either too small or too large, indicating that over-emphasizing either the contrastive loss or the cosine embedding loss reduces the effectiveness of the learned representation. This supports the choice of $\lambda = 0.45$, where the model achieves a balanced contribution from angular similarity and Euclidean-distance-based separation.

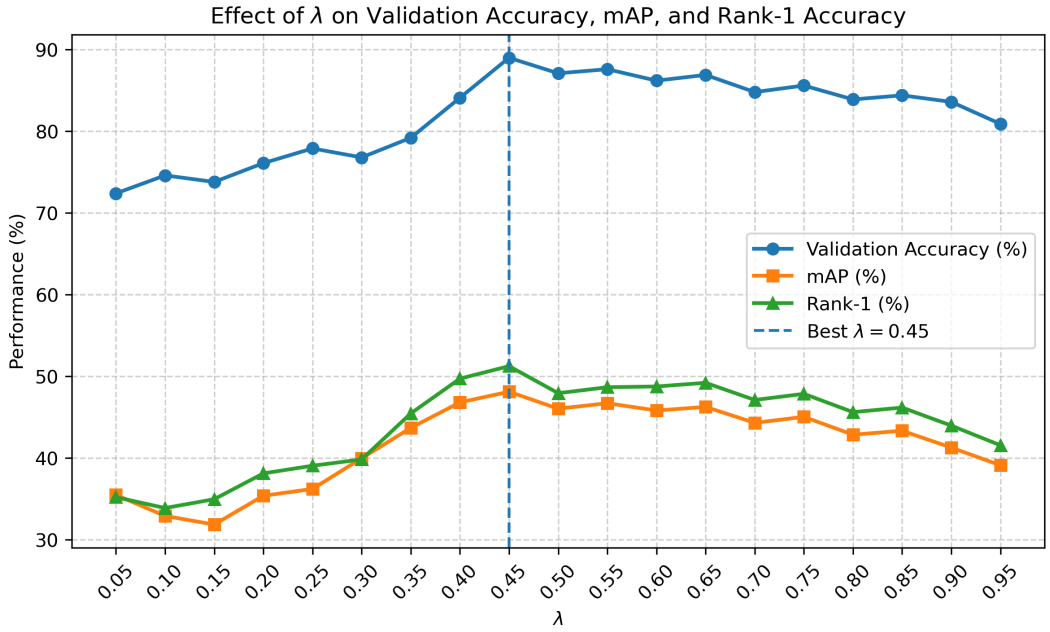


Figure 4.4: Effect of the weighting parameter λ used to balance the multi-loss objective on validation accuracy, mAP, and Rank-1 accuracy on the CUHK03 dataset under 20% random label noise. The best overall performance is observed at $\lambda = 0.45$.

4.7 Conclusion

We propose a robust object Re-ID approach that integrates a statistical OD framework into training. Our Beta-SOD framework models cosine similarity distributions with a two-component Beta mixture to iteratively filter noisy labels, enabling unsupervised task-specific dataset de-noising. A Siamese network is then trained using combined BCE, contrastive, and cosine embedding losses to enhance both classification and feature similarity. Experiments show improved robustness and accuracy across Re-ID benchmarks under varying noise levels. The observed performance gains strongly suggest the effectiveness and generalization capacity of Beta-SOD in real-world noisy label environments.

We present a novel theoretical result demonstrating that the unique identifiability of two-component Beta mixtures is attainable with probability 1. Based on the strong performance observed in Re-ID tasks, we expect Beta-mixture modeling to be a versatile and effective tool for de-noising cosine similarity scores produced by Siamese architectures, with potential applications extending to face recognition, speaker verification, and sentence embedding models.

While our statistical OD strategy proved effective as a standalone denoising module, future work will explore integrating this mechanism more tightly into the training neural network pipeline. This can be achieved by embedding the EM-based outlier scores directly into the loss function by re-weighting each training pair's similarity contribution based on its clean class probability γ (soft-filtering). An additional extension would be to integrate an attention mechanism into the Siamese architecture, allowing the model to emphasize the most discriminative features within paired images. This could lead to more accurate estimation of cosine similarities, thereby enhancing the effectiveness and efficiency of the similarity-based OD.

4.8 Appendix 4A: Beta-SOD Supplementary Material

The examples in Fig. 4.5 and Fig. 4.6 show the two types of label noise added to the CUHK03 dataset. Fig. 4.5 illustrates the random noise setting. In row (a), pairs that are actually dissimilar (true Class 0) are purposely mislabeled as Class 1. Row (b) shows pairs from the same identity (true Class 1) that are incorrectly assigned to Class 0. This form of noise does not follow a structure and appears randomly.

Fig. 4.6 shows a structured noise setting. Here, "noise" refers to incorrect label assignments in the dataset. Noise is introduced using similarity scores generated by the ResNet50 model, a deep neural network used for image feature extraction (see section 4.6). In row (a), we add pairs from different identities (true Class 0), meaning images of different individuals, with high cosine similarity (a measure of how similar two vectors are) to Class 1. Row (b) shows pairs from the same identity (true Class 1), that is, images of the same individual, with low similarity scores, which are assigned to Class 0. This noise follows a clear pattern, unlike random noise. It reflects errors arising from the model's limitations in learned features. These figures present the two noise models and form the basis for evaluating the robustness of the proposed filtering approach.



Figure 4.5: Examples of random noise in CUHK03 dataset. Row (a) shows dissimilar pairs (true Class 0) mislabeled as Class 1, while row (b) shows similar pairs (true Class 1) incorrectly assigned to Class 0.



Figure 4.6: Examples of Pattern noise in CUHK03 dataset. Row (a): dissimilar pairs with high similarity labeled as Class 1; row (b): similar pairs with low similarity labeled as Class 0. Similarity scores are shown for each pair.

Figure 4.7 shows how similarity distributions evolve during iterative filtering at different noise levels. Each row is a specific noise setting (10%, 20%, 30%), and each column shows a filtering step. In the first column, Class 1 (clean pairs) and

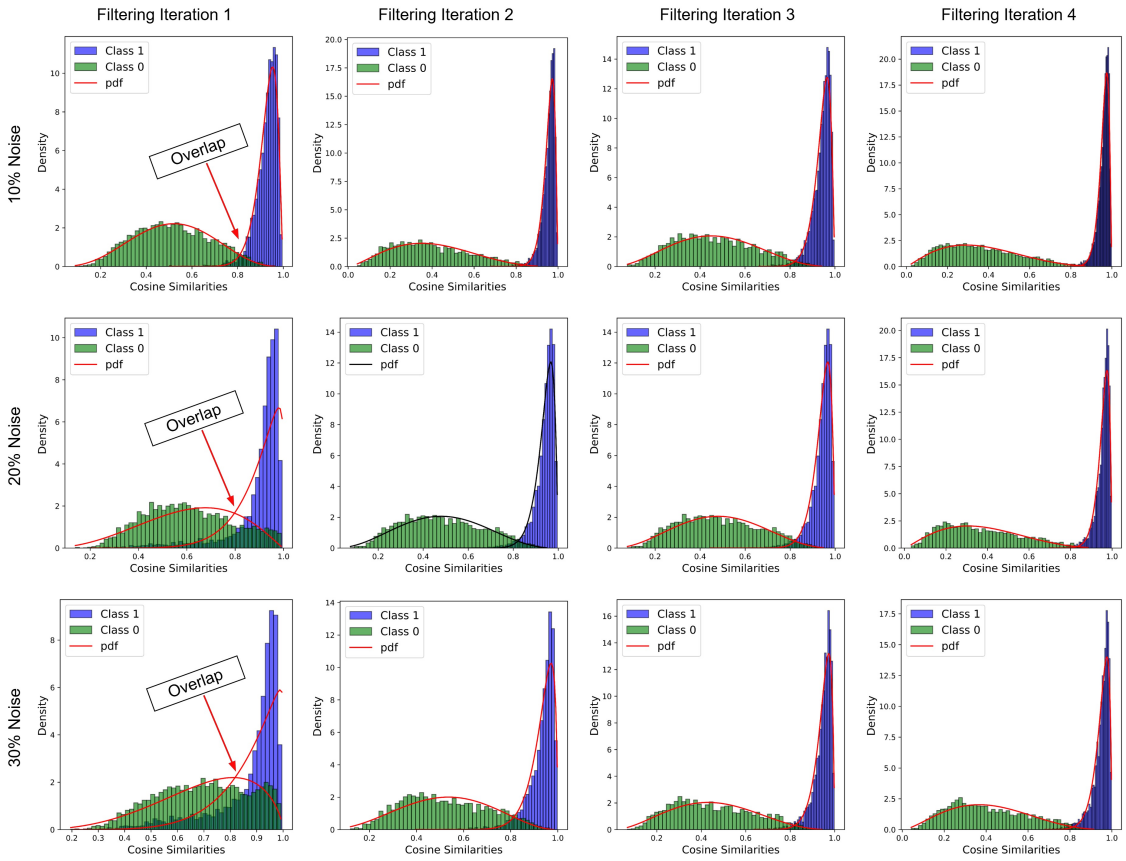


Figure 4.7: Evolution of similarity distributions during filtering. Rows correspond to noise levels (10%, 20%, 30%), while columns represent successive filtering iterations.

Class 0 (noisy pairs) distributions overlap significantly, indicating substantial noise. After the first filtering step (second column), the overlap decreases, suggesting that much of the noise was removed. As filtering continues (third and fourth columns), distributions separate further and eventually show minimal overlap. This growing separation shows improved distinction between clean and noisy pairs. The figure demonstrates that the proposed filtering method helps separate the data, leading to clearer and more homogeneous similarity representations.

4.9 Appendix 4B: Alternative Filtering

To study how noisy labels affect similarity learning, we first use an OD framework that iteratively removes low-similarity samples. This method helps us understand noisy pair distributions and motivates the more principled Beta-SOD approach de-

scribed earlier.

The proposed framework operates on image pairs labeled as either similar (class 1) or dissimilar (class 0). A MobileNetV2 network, pre-trained on ImageNet and fine-tuned on the target dataset, is used to extract high-dimensional feature representations for each image, summarizing image attributes. The cosine similarity between the feature vectors of each pair is then computed and used as a measure of visual similarity. Cosine similarity is a value between -1 and 1 indicating the angle between two vectors, with higher values indicating greater similarity.

We assume correctly labeled similar pairs show higher cosine similarity, while mislabeled pairs show lower similarity. The OD process removes class 1 samples with the lowest similarity scores in each iteration.

At first, the model fine-tunes for several epochs. The validation accuracy is monitored to decide when to start filtering. Filtering starts when the feature extractor's validation accuracy stabilizes, showing consistent learned representations. At each step, cosine similarity is calculated for all pairs. Class 1 samples are ranked in ascending order. The 2% with the lowest similarity scores are removed because they are likely to have noisy or incorrect labels. To keep class balance, the same number is randomly removed from class 0. The removal rate is capped at 2% per step. This prevents underfitting and ensures the model continues to learn from hard examples. Aggressive early filtering can hinder generalization.

To monitor the evolution of the dataset during filtering, we compute a statistical summary of the similarity distributions at each iteration. Specifically, a one-dimensional feature vector of length six,

$$\mathbf{v} = \{m_1, m_2, m_3, v_1, v_2, v_3\},$$

is constructed using the median and variance of three subsets: (i) the filtered class 1 samples,

$$(m_1, v_1) = \left(\text{Median}\left(C_1^{\text{filtered}}\right),\ \text{Var}\left(C_1^{\text{filtered}}\right)\right),$$

(ii) the remaining class 1 samples,

$$(m_2, v_2) = \left(\text{Median}\left(C_1^{\text{remaining}}\right),\ \text{Var}\left(C_1^{\text{remaining}}\right)\right),$$

and (iii) the remaining class 0 samples,

$$(m_3, v_3) = \left(\text{Median}\left(C_0^{\text{remaining}}\right),\ \text{Var}\left(C_0^{\text{remaining}}\right)\right).$$

This feature vector captures the properties of the filtered samples and how the remaining data changes.

We repeat filtering for five iterations and concatenate each iteration's statistical descriptors into a higher-dimensional vector, $\mathrm{V}^{5\times 6}$. This vector tracks how the dataset changes as noise is removed.

A key challenge is knowing when to stop filtering and how much to remove each time. To address this issue, we introduce a learning-based control mechanism using a Random Forest (RF) classifier. The RF is trained on synthetic noisy datasets created by adding dissimilar pairs to the set of similar pairs. For each set, the OD procedure creates a sequence of statistical descriptors used as input features for the RF.

Each sequence is labeled by the convergence of its similarity distribution. If the median and variance change significantly across steps, the sequence is labeled class 1. This means more filtering is helpful. If the statistics remain stable, the sequence is class 0, indicating that the dataset is clean enough and that further filtering is not needed.

During inference, the RF classifier predicts whether to continue filtering. If further filtering is suggested, we adapt the removal rate using the RF's confidence p via $\mathrm{R_n} = 3p - 1$. This method limits filtering to 2% per iteration and reduces aggressiveness as the dataset becomes cleaner. It prevents excessive removal of valid samples. The overall iterative filtering procedure is summarized in Algorithm 2.

Limitation. Although this approach shows that similarity-based filtering and basic

statistics can reduce label noise, it relies on heuristic assumptions rather than modeling the distribution of similarity scores directly. The separation between clean and noisy samples is inferred from summary statistics rather than modeled explicitly.

This procedure also adds computational overhead. Each evaluation needs five iterations to create the high-dimensional $V^{5\times6}$. For example, with 4% noise, the first five steps remove 2% each time, forming $V^{5\times6}$. The RF classifier then evaluates this result, usually assigning a positive label when variation in the median and variance is seen across the sequence.

Consequently, the filtering process continues with additional iterations. Using a sliding window mechanism, an updated $V^{5\times6}$ is constructed and re-evaluated by the RF model. However, even if the majority of the noise is effectively removed in the first few iterations, the algorithm must continue through subsequent iterations until the generated sequence is classified as negative, which serves as the stopping criterion. This results in unnecessary computations, highlighting a key efficiency limitation of the iterative filtering approach.

These limitations motivate the need for a more principled approach that directly captures the distribution of similarity values. This observation leads to the development of the Beta-SOD framework, which models similarity distributions as Beta mixtures and provides a more robust, theoretically grounded solution for noise identification.

4.9.1 Result Analysis

Table 4.9 presents a comparison between the proposed Beta-SOD method and the iterative filtering approach across multiple datasets and varying noise levels. A consistent trend emerges across all settings: Beta-SOD achieves higher precision, whereas iterative filtering achieves higher recall.

Across all datasets, Beta-SOD maintains strong precision (e.g., 85.79% on Market1501 and 87.46% on VeRi-776 at 10% noise), indicating that most of the removed samples are indeed noisy. In contrast, iterative filtering achieves higher recall, of-

Algorithm 2: Iterative Outlier Detection (OD) for Noisy Pair Filtering

Input: Dataset D with image pairs labeled as class 0 or class 1
Output: Filtered dataset $\widehat{D}$
Initialize: MobileNetV2 with ImageNet-pretrained weights;
$R_n \leftarrow 0.02$;
$\mathbf{V} \leftarrow \emptyset$;
filtering_active $\leftarrow$ `False`;
for *epoch* $= 1$ **to** N **do**
  Train MobileNetV2 on D;
  Evaluate validation accuracy Acc_{val};
  if *filtering_active* $=$ `False` **then**
    if Acc_{val} *has stabilized* **then**
      filtering_active $\leftarrow$ `True`;
      $\mathbf{V} \leftarrow \emptyset$;
    continue to the next epoch;
  Compute cosine similarity for all image pairs in D;
  Sort class 1 samples in ascending order of similarity;
  $C_1^{\text{filtered}} \leftarrow$ lowest R_n fraction of class 1 samples;
  $C_1^{\text{remaining}} \leftarrow$ remaining class 1 samples;
  $C_0^{\text{remaining}} \leftarrow$ randomly sampled class 0 samples of equal size to C_1^{filtered};
  Update D by removing C_1^{filtered} and the corresponding sampled class 0 pairs;
  Compute $\mathbf{v} \leftarrow \{m_1, m_2, m_3, v_1, v_2, v_3\}$;
  Append $\mathbf{v}$ to $\mathbf{V}$;
  if $|\mathbf{V}| = 5$ **then**
    $\mathbf{v}^{30} \leftarrow \text{Flatten}(\mathbf{V})$;
    $(y, p) \leftarrow \text{RF}(\mathbf{v}^{30})$;
    if $y = 1$ **then**
      $R_n \leftarrow 3p - 1$;
    else
      break;
$\widehat{D} \leftarrow D$;
return $\widehat{D}$;

Table 4.9: Comparison of iterative filtering and Beta-SOD in terms of outlier removal, precision, and recall across different datasets and various noise levels.

Dataset	Noise	Iterative Filtering			Beta-SOD		
		Outliers	Precision	Recall	Outliers	Precision	Recall
Market1501	10%	10.31	85.79	76.87	12.82	75.66	90.00
	20%	21.26	82.93	80.56	22.61	72.62	90.81
	30%	32.97	75.73	81.67	34.00	68.20	92.01
CUHK03	10%	11.32	82.09	73.63	12.91	75.72	90.22
	20%	21.56	80.47	76.08	23.01	71.83	91.31
	30%	32.28	75.75	81.18	34.11	66.22	91.93
VeRi-776	10%	11.34	87.46	84.8	13.90	72.92	87.21
	20%	22.76	85.66	88.39	23.82	70.11	87.82
	30%	33.05	84.47	89.86	34.97	65.20	89.10

ten exceeding 90%, suggesting that it removes a larger proportion of noisy samples. However, this comes at the cost of significantly lower precision, leading to many clean samples being incorrectly discarded.

This trade-off becomes more pronounced as the noise level increases. Iterative filtering continues to remove a large fraction of noisy samples, but its precision steadily drops, reflecting increasingly aggressive filtering and a higher number of false positives. On the other hand, Beta-SOD remains more selective, preserving a higher proportion of clean samples even at higher noise levels. The difference in behavior can be attributed to the underlying design of the two methods. Iterative filtering relies on similarity-based ranking and repeated removal steps, which makes it inherently aggressive. While this allows it to capture more noisy samples, it lacks a principled mechanism for distinguishing hard clean samples from true outliers. As a result, it often removes informative but challenging examples, which can negatively impact the learning process.

In contrast, Beta-SOD adopts a probabilistic modeling approach based on Beta mixture distributions, thereby explicitly capturing the underlying similarity distributions of clean and noisy pairs. This allows for a more controlled and reliable separation between the two, reducing the likelihood of removing clean samples. Consequently, Beta-SOD achieves higher precision by avoiding unnecessary filtering.

Importantly, although Beta-SOD may leave a small fraction of hard noisy pairs in

the dataset, this does not significantly degrade model performance. Deep CNNs are generally robust to a moderate level of residual noise and can tolerate such samples during training. In fact, preserving hard examples, even if some are noisy, can be beneficial, as it prevents the model from overfitting to overly clean or simplified data. In contrast, the aggressive iterative filtering strategy may discard valuable training samples, leading to reduced generalization.

4.10 Appendix 4C: Extension to Road-side Object Geolocation Pipeline

Learned similarity, obtained via a Siamese-style framework under noisy paired-view conditions, can be effectively incorporated into roadside object geolocation pipelines. This is particularly important in weakly supervised settings, where object detectors are prone to producing many false positives, thereby degrading the reliability of multi-view associations.

A key challenge in this domain is the limited availability of large-scale, annotated datasets for multi-view association of roadside objects. This constraint makes it difficult to learn robust cross-view correspondences directly. To mitigate this, person and vehicle Re-ID datasets are employed as proxy sources, where similarity learning is well-established. By introducing controlled noise that approximates the behavior of false detections in weakly supervised roadside scenarios, these datasets provide a practical means to train models under realistic conditions.

The incorporation of Beta-SOD further strengthens this approach by explicitly modeling the distribution of similarity scores. This enables the identification and suppression of unreliable matches, effectively filtering out spurious associations before they propagate through the pipeline. As a result, the learned similarity becomes more robust to noise and better suited for downstream tasks.

Integrating this similarity-learning framework with Beta-SOD into the roadside object geolocation pipeline provides a principled solution for handling noisy detec-

tions and ambiguous associations. It improves the quality of multi-view matching, which is critical for accurate triangulation and spatial reasoning.

The next chapter builds on this foundation by demonstrating how the proposed similarity modeling and noise-filtering strategy can be incorporated into the full geolocation pipeline, resulting in more reliable and accurate roadside object localization under weak supervision.

Chapter 5

Similarity-Augmented Roadside Object Triangulation from street level images

Object geolocation from street-level imagery under weak supervision is challenging due to noisy detections and the absence of bounding-box annotations. This chapter addresses this task via a similarity-augmented MRF framework that integrates visual consistency into geometry-driven multi-view triangulation. Starting from weakly supervised object detection, this formulation enforces both visual and geometric consistency across paired views to estimate object locations. To discourage visually inconsistent configurations, a Siamese network is trained to estimate pairwise visual similarity. Label noise introduced by weak supervision is addressed using a Beta-distribution–based filtering strategy during Siamese training. We also propose a novel deterministic optimizer for the MRF model that significantly reduces computational requirements and improves reproducibility. Experiments on real-world street-level datasets from Amsterdam and Dublin demonstrate consistent improvements in precision, recall, and localization error over a geometry-only baseline.

5.1 Introduction

Accurate geolocation of roadside objects from street-level imagery is essential for applications such as urban mapping, infrastructure monitoring, and autonomous navigation. Unlike image-level geolocation, object-level geolocation requires estimating precise real-world object coordinates from multiple views under significant uncertainty in depth, bearing, and appearance. Weakly supervised detectors (Li et al. 2025) further amplify this challenge by producing uncertain object locations that generate excessive amounts of low-confidence triangulation candidates. Recent work has therefore shifted toward geometry-driven multi-view inference under weak supervision, where object locations are estimated from noisy geometric cues (Ahmad and Krylov 2024). While video-based and track-driven approaches improve stability by exploiting temporal continuity (Wilson et al. 2022), this inherently assumes ordered observations and reliable tracking assumptions that do not hold for large-scale, crowdsourced image datasets.

Classical learning-based cross-view geolocation estimates camera locations by matching ground-level images to georeferenced aerial or satellite imagery using learned metric representations (Hu et al. 2018; Shi et al. 2019; Zhu et al. 2020). While effective, these methods rely on strong appearance cues and dense georeferenced databases, making them impractical for object-level localization of custom roadside assets. Beyond purely geometric formulations, recent works (Weng et al. 2020; Pang et al. 2023) explore joint object association and localization in 3D using tracking and graph-based reasoning. Such approaches rely on dense image streams and strong calibration, limiting their applicability to sparser street-level imagery.

More recently, graph neural networks (GNNs) have been adopted to explicitly model relational structure among views, objects, or observations, enabling learned aggregation of geometric and contextual cues (Ye et al. 2022c). Nassar *et al.* (Nassar et al. 2020) formulate localization as a graph inference problem, where spatial relationships between observations are encoded as edges and refined through message passing to improve global consistency. While effective in structured settings, most

graph-based approaches require dense supervision or explicit object grouping.

In contrast, purely geometric triangulation-based approaches are prone to failure under dense view coverage, where accumulated measurement uncertainty and ambiguous geometric configurations degrade localization accuracy. To address this, object geolocation is formulated as an energy minimization problem within a MRF, enforcing multi-view geometric consistency through unary and pairwise terms (Krylov et al. 2018). While effective, this formulation relies predominantly on geometric cues, limiting its ability to suppress visually inconsistent yet geometrically plausible configurations in weakly supervised settings. Under weak supervision, MRF-based triangulation groups detections into view pairs that may include both true multi-view observations and spurious pairs arising from false positives (Ahmad and Krylov 2024). Since both satisfy geometric constraints, geometry alone is insufficient.

In our work we integrate visual consistency into the MRF framework. Specifically, we train a Siamese network to estimate pairwise visual similarity and propose a novel Beta-distribution–based filtering strategy to remove visually inconsistent pairs during training. The resulting similarity scores are incorporated as an additive energy term in the MRF, penalizing unreliable configurations. We introduce a novel deterministic optimizer for the MRF model that significantly reduces computational requirements and improves reproducibility. Experiments with real-world street-level imagery demonstrate consistent improvements in object geolocation accuracy.

5.2 Methodology

This work proposes a noise-aware geolocation framework that improves multi-view triangulation performance under weak supervision. The proposed method retains a weakly supervised object detection and an MRF-based triangulation framework, while explicitly addressing the false positives introduced by weak supervision through similarity-driven noise filtering, as shown in Fig. 5.1. A learned image similarity term is incorporated into the MRF inference at the inference stage to enhance robustness and geolocation accuracy, see Fig. 5.2.

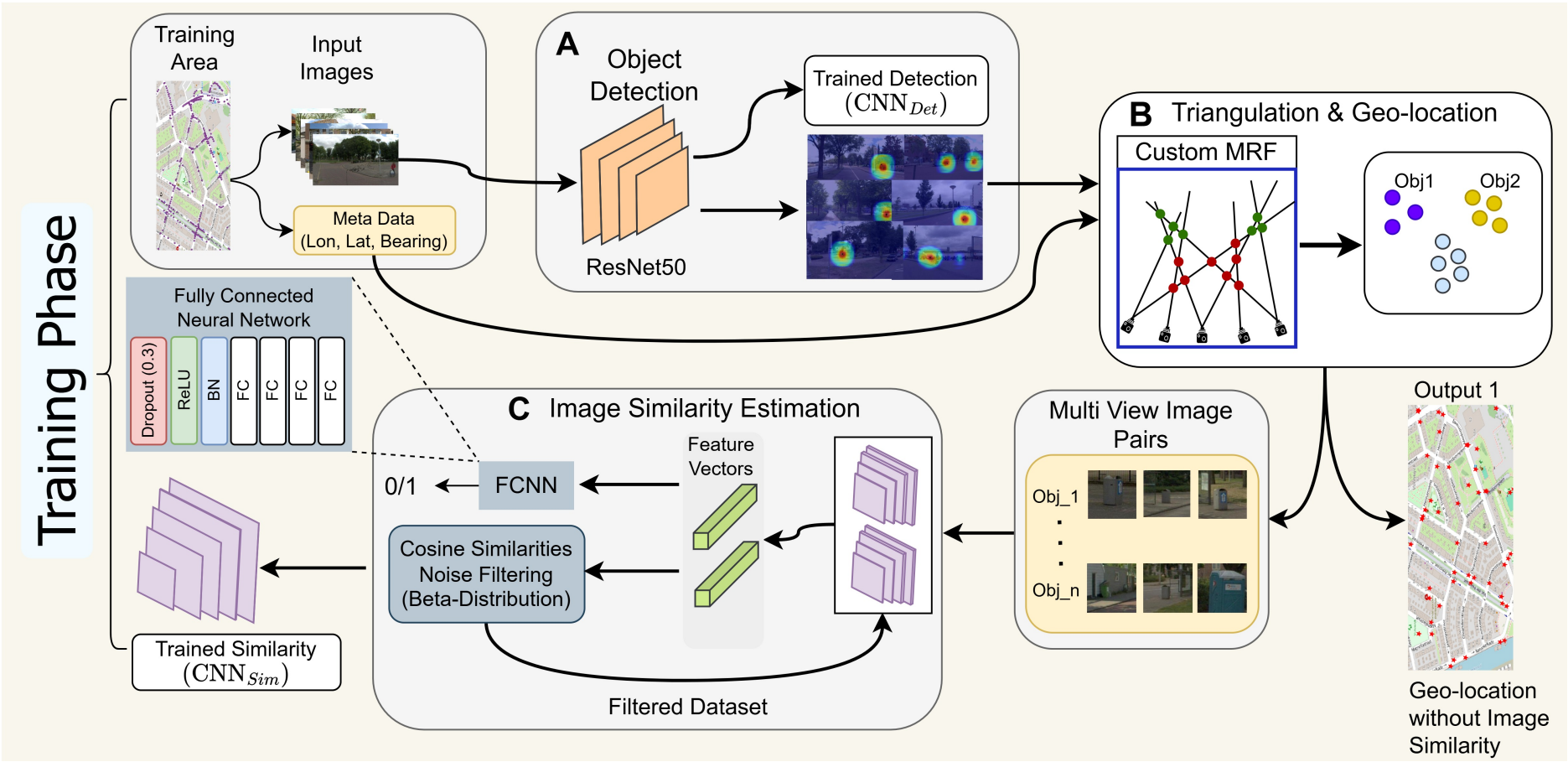


Figure 5.1: Proposed training pipeline: (A) weakly supervised object classification and Grad-CAM–based localization, (B)MRF-based triangulation to generate multi-view image pairs, and (C) Siamese similarity learning with Beta-mixture filtering.

Object detection and geolocation. In the absence of bounding-box annotations, we employ a weakly supervised object detection strategy similar to (Ahmad and Krylov 2024), where a CNN trained with image-level supervision generates class activation maps (e.g., Grad-CAM) to localize roadside objects in street-level imagery. While this enables scalable detection, false positives are unavoidable and propagate to downstream stages if not explicitly addressed. Object geolocation is performed using an MRF-based probabilistic triangulation framework following (Krylov et al. 2018). The method operates on paired camera views, enforcing geometric consistency through ray intersection and the removal of invalid configurations. However, as the inference relies solely on geometric cues, false-positive detections can still form geometrically valid view pairs, resulting in incorrect triangulations and degraded geolocation accuracy.

Noise filtering. The proposed geolocation framework employs the same noise filtering method introduced in Chapter 4, Sec. 4.3 to improve the reliability of multi-view object associations. Candidate image pairs generated by the MRF-based pipeline are used to train a Siamese network, where positive pairs are expected to represent the same roadside object and negative pairs correspond to different objects. However, because the detections are obtained under weak supervision, the positive set may contain incorrect associations caused by false or imprecise detections.

To distinguish between genuine and noisy image pairs within class 1, we model the distribution of similarity scores using a two-component Beta mixture model, defined as:

$$p(s) = \omega \ \mathrm{Beta}(s \mid a_1, b_1) + (1 - \omega) \ \mathrm{Beta}(s \mid a_0, b_0), \tag{5.1}$$

where $\mathrm{Beta}(\cdot)$ denotes the Beta probability density function. The component $\omega \ \mathrm{Beta}(s \mid a_1, b_1)$ captures high-similarity, visually consistent image pairs, while $(1 - \omega) \ \mathrm{Beta}(s \mid a_0, b_0)$ models low-similarity pairs corresponding to noisy or mismatched observations.

A shared MobileNetV2 backbone is used to extract feature embeddings from each image pair, and cosine similarity is used to measure their visual consistency. The

resulting scores are processed using exactly the same filtering procedure described in Chapter 4, Sec. 4.3 to separate reliable pairs from noisy correspondences. The Siamese network is first trained on the full dataset and then fine-tuned on the filtered data, with this process repeated until the proportion of newly removed samples falls below the predefined stopping threshold.

The same multi-loss objective used in Chapter 4, sec. 4.5 is adopted here:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{\text{COS}} + (1 - \lambda)\mathcal{L}_{C}. \tag{5.2}$$

where L_{CE}, L_{COS} and L_C denote the binary cross-entropy, cosine embedding, and contrastive losses, respectively. The final similarity model estimates reliable visual similarity between multi-view detections, and these scores are incorporated into the MRF as an additional energy term to reduce false associations and improve geolocation accuracy.

5.2.1 Information fusion with MRF

Object detections in images combined with the image metadata are used to construct rays (in 2D) that originate at the camera GPS coordinates (assumed identical to camera centre). The bearing (direction of ray) is selected towards the middle of the segmented object. The rays create a 2D graph with nodes formed by intersections of these rays. Edges are established between nodes on the intersected rays. An MRF model is employed to estimate binary decisions ($z_i \in \{0, 1\}$) indicating presence of the object of interest at a particular node i (Krylov et al. 2018). The MRF energy of a node z_i is governed by two geometric unary terms and one pairwise term. We extend this formulation by introducing an additive unary term u_2, which incorporates visual information by rewarding or penalizing a node based on the Siamese similarity of the two views that generate it. Accordingly, the total energy

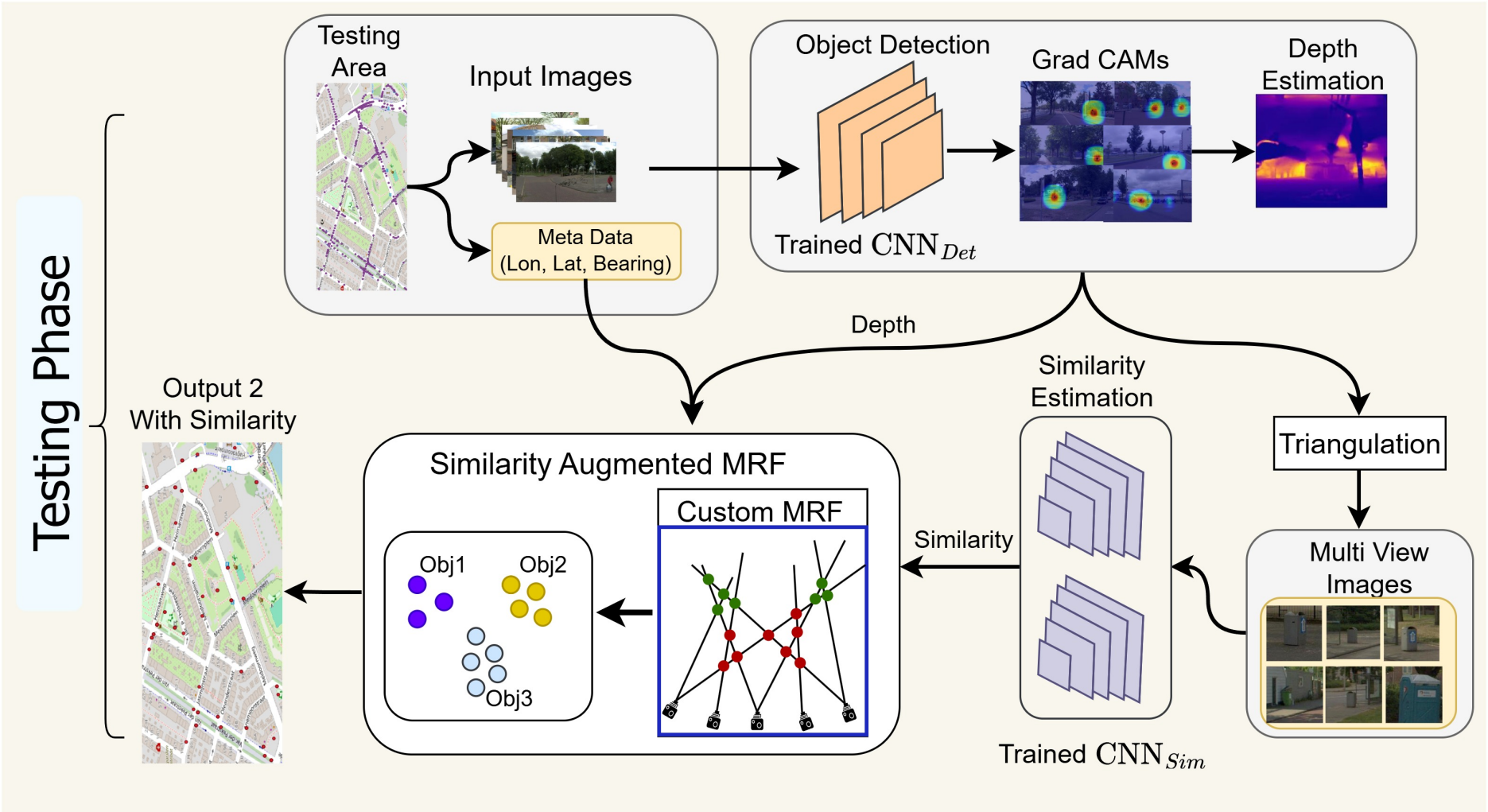


Figure 5.2: Inference pipeline: object detection and localization, triangulation-based view pairing, Siamese similarity estimation, and final geolocation using a similarity-augmented MRF.

is defined over all $N_{\mathcal{Z}}$ nodes as:

$$\mathcal{U}(\mathbf{z}) = \sum_{i=1}^{N_{\mathcal{Z}}} \sum_{j=0}^{2} \alpha_j u_j(z_i) + (1 - \sum_{j=0}^{2} \alpha_j) \sum_{i=1}^{N_{\mathcal{Z}}} \sum_{j=1}^{N_{\mathcal{Z}}} p(z_i, z_j) \tag{5.3}$$

The first unary term penalizes discrepancy between distance from the camera to the node estimated with monocular depth CNN (Δ_{ij}) and with triangulation (d_{ij}): $u_1(z_i) = z_i \sum_{j=1,2} \|\Delta_{ij} - d_{ij}\|$. The u_0 unary term promotes inclusion of terms on any ray R and is inversely proportional to the number of intersections on the ray: $u_0(z_i) = -z_i/(\sum_{x_n \in R} 1)$. The presence of this term ensures the minimum of energy is not reached on empty configuration $z_i = 0$ for all i. The pairwise term penalizes distances between any two nodes simultaneously active on any view-ray R: $p(z_m, z_n) = \sum_{x_m, x_n \in R} z_m z_n \|x_m - x_n\|$, where x_i is a GPS position of node controlled by state z_i. The similarity unary term u_2 penalizes nodes generated from visually inconsistent view pairs: $u_2(z_i) = z_i(1 - s_i)$, where $s_i \in [0, 1]$ denotes the similarity score predicted by the Siamese network. Nodes supported by visually consistent views ($s_i \approx 1$) incur minimal penalty, while intersections formed by visually dissimilar views are penalized, thereby suppressing spurious nodes that are geometrically admissible but visually inconsistent. The contributions of the terms are governed by $\alpha_i \in [0, 1)$, $\sum \alpha_i < 1$.

The state $\mathbf{z} = \{z_i\}_{i=1,\cdots,N_{\mathcal{Z}}}$ that minimizes the energy $\mathcal{U}$ (Eq. 5.3) is computed using the Quadratic pseudo-Boolean optimization (QPBO) algorithm (Rother et al. 2007), which is based on graph-cut optimizations. QPBO has two substantial advantages over the original Simulated Annealing based optimization proposed in (Krylov et al. 2018): firstly, the optimization result is deterministic, which avoids the need to rerun the procedure multiple times to arrive at an averaged (stable) field configuration. Secondly, the procedure has predictable complexity and is fast. The use of QPBO is possible by eliminating higher-order penalty terms originally used in (Krylov et al. 2018) that did not allow for energy function composition compatible with this type of optimization. Specifically, the stand-alone penalty term u_3 used in (Krylov et al. 2018) has been replaced by the unary term u_0 (Eq. 5.3). The

pairwise term is supermodular (Rother et al. 2007), but the number of such terms is relatively small, and empirically allows for optimal state assignments to all graph nodes using QPBO/QPBOP algorithm.

5.3 Datasets and Experimentation

5.3.1 Dataset Description

We construct two street-level datasets centered around Dublin and Amsterdam, with public wastebins as the target object.

Dublin. street-level images are obtained via the Mapillary API (Mapillary 2024), selecting only images flagged as containing wastebins. Public street furniture data from Smart Dublin bins dataset (Smart Dublin n.d.) is used for spatial supervision: images within a 15 m radius of DCC bin locations are labeled as class 1 (9,900 images), while class 0 samples are selected from images located 30-100 m away from any bin. All images are resized to 1500×900. We evaluate on two test sets from Mapillary and DCC dataset, covering 2.7 km of roads with images sampled every 3 m, six 68.77° views per location.

Amsterdam. Panoramic images are obtained via the Mapillary API (Mapillary 2024), restricted to city-provided panoramas flagged as containing wastebins and captured with a fixed north-facing orientation (0°). The ground-truth locations of wastebins are obtained from the OSM API (OpenStreetMap Contributors 2024) and used to associate panoramas with object positions. A total of 4,100 panoramas (8000×4000) are split into 2000×2000 sub-images and labeled using Mapillary masks, yielding 20,000 samples, of which 85% are used for training and 15% for testing.

5.3.2 Training Setup

The training follows a two-stage setup comprising weakly supervised object detection and Siamese-based similarity learning, with MRF-based geolocation performed at

inference time.

Backbone architectures. ResNet-50 is used as the primary image classification backbone, with Grad-CAM employed to localize the object within each image. MobileNetV2 is adopted as the backbone feature extractor in the Siamese architecture due to its efficiency and proven effectiveness in metric learning-based similarity estimation (Yang et al. 2023).

Implementation details. All models are implemented in PyTorch and trained on an NVIDIA Titan X GPU (12 GB). The detection network is trained for 20 epochs using Adam with a batch size of 8 and a step learning-rate scheduler (step size 5, $\gamma = 0.1$). For similarity learning, the Siamese network employs a MobileNetV2 backbone with an FCNN head $(512, 512, 256, 128)$, batch normalization and ReLU activations, and dropout (0.3) applied to the final two layers. Training uses Adam optimizer with an initial learning rate of 10^{-3}, batch size 32, and learning-rate decay every 5 epochs. Optimization alternates between Siamese training and Beta-SOD–based pair filtering, with $K \in [3, 7]$ epochs per cycle until validation accuracy stabilizes, and terminates when the estimated noisy-pair ratio ω_k falls below 2%. The combined loss is optimized with $\lambda = 0.45$, selected via grid search.

5.3.3 Results Analysis

Although several multi-view object geolocation datasets have been proposed, no widely accepted benchmark or evaluation protocol exists (Wilson et al. 2023). Existing results remain dataset- and implementation-specific, while relevant datasets such as Pasadena Multi-View (Nassar et al. 2020) are not publicly accessible. Therefore, we introduce two publicly available street-level datasets from Dublin and Amsterdam, using public waste bins as the target objects, to support future comparisons.

We report results for weakly supervised object detection and geolocation on the Dublin and Amsterdam datasets, including impact analysis of Beta-distribution-based noise filtering.

Object detection derformance. ResNet-34, ResNet-50, and ResNet-101 are eval-

Table 5.1: Classification results on the Amsterdam dataset

Model	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
ResNet-34	94.7	88.2	91.3	92
ResNet-50	**95.81**	**89.10**	**92.33**	**93.1**
Resnet-101	93.1	88	90.47	92.4

Table 5.2: Classification results with Resnet-50 architecture

Test Set	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Amsterdam	95.81	89.10	92.33	93.1
Dublin (DCC)	93.1	87.2	90.05	92.8
Dublin (Mapillary)	94.21	89	91.53	93

uated for image classification on the Amsterdam dataset. As shown in Table 5.1, ResNet-50 achieves the best overall performance, with 95.81% precision, 89.10% recall, and 93.1% accuracy, and is therefore selected as the primary classification backbone. Using the selected ResNet-50 backbone, classification performance across all test sets is reported in Table 5.2. On the Amsterdam dataset, the model achieves 95.81% precision, 89.10% recall, 92.33% F1-score, and 93.1% accuracy. Comparable performance is observed on the Dublin datasets, with 93.1% precision, 87.2% recall, and 92.8% accuracy on Dublin (DCC), and 94.21% precision, 89.0% recall, and 93.0% accuracy on Dublin (Mapillary). This indicates that the detection model generalizes well across cities and data sources, providing reliable inputs for the geolocation.

Effect of noise filtering. We next analyze the effect of the proposed Beta-distribution–based noise filtering on image pairs generated after detection and MRF-based triangulation. Approximately **8.5%** of pairs are filtered on the Dublin dataset and **12.3%** on the Amsterdam dataset, reflecting higher visual ambiguity in panoramic imagery (Fig. 5.3). In the absence of ground-truth pair labels, manual inspection indicates that about **85%** of the removed pairs correspond to genuine mismatches. This is further corroborated at inference time, where the pretrained Siamese network consistently assigns low similarity scores to the filtered pairs, confirming denoising effectiveness.

Geolocation performance. We now evaluate the geolocation performance of the proposed framework on the Amsterdam and Dublin datasets using the similarity-



Figure 5.3: Filtered image pairs after Beta-distribution-based denoising for Dublin (top) and Amsterdam (bottom) datasets

Table 5.3: Geolocation performance on the Amsterdam and Dublin datasets against MRF (Krylov et al. 2018)

Dataset	Method	Precision (%)	Recall (%)	Error (m)
Amsterdam	MRF	64.25	79.9	2.54
	Similarity-Aug. MRF	68.2	86.1	1.87
Dublin (Mapillary)	MRF	66.21	80.3	2.51
	Similarity-Aug. MRF	71.4	87.1	1.52
Dublin (DCC)	MRF	65.1	80.8	2.38
	Similarity-Aug. MRF	72.8	87.9	1.49

augmented MRF. At inference, test images are processed by the pretrained detection model, triangulated to generate admissible intersections, and refined using similarity scores from the Siamese network.

Table 5.3 reports the geolocation results in terms of precision, recall, and absolute localization error. Compared to the baseline MRF, incorporating similarity cues consistently improves performance: on the Amsterdam dataset, precision increases by 4% and recall by 6.2%, while the mean localization error is reduced by 0.7 m; similar trends are observed on the Dublin datasets. These results show that integrating similarity-aware filtering into the MRF effectively suppresses spurious intersections arising from noisy detections, leading to higher precision and recall and substantially

lower localization error. Overall, the proposed similarity-augmented MRF yields more robust and accurate geolocation across three diverse datasets. The QPBO solver for MRF reduces the computational complexity by an order of magnitude and leads to deterministic and hence better reproducible geolocation results.

Ablation Analysis. Table 5.4 presents an ablation study evaluating the contribution of the proposed similarity augmentation and statistical noise filtering to the MRF-based geolocation framework. Incorporating visual similarity directly into the MRF without noise filtering consistently degrades the geolocation performance across all three datasets, resulting in lower precision, reduced recall, and higher localization error compared to the original MRF. This is because weakly supervised object detection generates numerous false positives, and assigning similarity scores to all multi-view detections allows visually similar but incorrect object pairs to influence the MRF energy optimization. In urban environments, roadside bins are often surrounded by visually similar structures, such as bollards, poles, and other street furniture, which can receive non-negligible similarity scores and therefore introduce incorrect cross-view associations. Consequently, these false matches bias the energy minimization process and lead to less accurate object localization. By contrast, applying the proposed noise filtering framework filters unreliable similarity estimates before they are incorporated into the MRF, effectively removing noisy correspondences while preserving reliable object associations. As a result, the similarity-augmented MRF with noise filtering achieves the best performance on all datasets, improving both precision and recall while reducing the average localization error. These results demonstrate that the benefit of visual similarity in multi-view geolocation is highly dependent on robust noise filtering, highlighting the importance of the proposed noise filtering framework.

5.4 Conclusion

This chapter presents a similarity-augmented MRF framework for object geolocation under weak supervision. By integrating Siamese-based visual similar-

Table 5.4: Ablation study of the proposed similarity-augmented MRF, evaluating the impact of visual similarity and the proposed statistical noise filtering on geolocation performance.

Dataset	Method	Precision (%)	Recall (%)	Error (m)
Amsterdam	MRF	64.25	79.9	2.54
	Similarity-Aug. MRF (no Noise filtering)	63.11	79.1	2.81
	Similarity-Aug. MRF (with Noise filtering)	**68.2**	**86.1**	**1.87**
Dublin (Mapillary)	MRF	66.21	80.3	2.51
	Similarity-Aug. MRF (no Noise filtering)	64.90	80.10	2.77
	Similarity-Aug. MRF (with Noise filtering)	**71.4**	**87.1**	**1.52**
Dublin (DCC)	MRF	65.1	80.8	2.38
	Similarity-Aug. MRF (no Noise filtering)	63.52	80.7	2.61
	Similarity-Aug. MRF (with Noise filtering)	**72.8**	**87.9**	**1.49**

ity into a geometry-driven triangulation model and mitigating label noise via Beta-distribution–based filtering, the proposed method effectively suppresses false-positive detections. We deploy a new deterministic optimizer for MRF, resulting in more stable and computationally feasible results. Experiments on real-world street-level datasets from Amsterdam and Dublin show consistent gains in precision, recall, and localization accuracy over a geometry-only MRF baseline. Future work will focus on joint optimization of detection and similarity learning.

Chapter 6

Edge-Aware Graph Neural Network for Roadside Object Geolocation

Accurate geolocation of roadside objects from street level imagery remains a challenging problem due to noisy detections, which are typical for real image datasets with limited annotations availability. This leads to inaccurate camera-to-object distance and bearing estimates. Traditional triangulation methods struggle under these conditions, as they rely heavily on clean geometric measurements and consistent multi-view observations. To address this we propose a scalable GNN model, coined EdgeAwareVH, that learns to reason directly over view-hypothesis relationships. The model integrates edge-aware attention with geometric message passing, enabling it to selectively emphasize reliable observations and suppress spurious intersections. Our pipeline combines geometric cues such as depths, bearing angles, camera pose information and OpenStreetMap-derived spatial context providing robustness in complex urban environments. We introduce a realistic noise simulation pipeline that mimics the behavior of weakly supervised detectors and evaluate our method under multiple noise configurations. Our experiments with wastebins geolocation on noisy simulated and real-world data demonstrate that EdgeAwareVH

achieves high localization accuracy and is resilient to depth and bearing noise. The implementation of our model and the data sets used are publicly available[1]

6.1 Introduction

The relevant literature on roadside object detection, visual geolocation, multi-view association, and triangulation-based localization has already been reviewed in Chapter 2. As discussed in Chapter 3, this study focuses on custom roadside objects for which detailed object-level and multi-view annotations are limited. A weakly supervised detection pipeline is therefore employed, followed by triangulation and a custom MRF model to filter geometrically inconsistent location hypotheses. In Chapter 5, the custom MRF is further extended by incorporating an additional similarity-based energy term, enabling visual appearance information to contribute to hypothesis selection. Although the custom MRF and its similarity-augmented formulation provide reasonable geolocation performance, both remain dependent on predefined energy terms and continue to face limitations under severe and realistic noise conditions, including inaccurate depth and bearing estimates, false detections, viewpoint variation, and unreliable triangulation intersections. These limitations motivate the development of a more adaptive, data-driven graph-based approach capable of jointly learning from heterogeneous geometric, visual, and contextual information.

Learning-based approaches to triangulation have been explored. In particular, Zhang et al. (Zhang et al. 2018) integrating CNN-based object detections and line of bearing estimation to localize roadside utility poles from street level imagery. Branson et al. (Branson et al. 2018) propose a CNN-based multi-view framework for large scale urban tree detection and species recognition using publicly available street level and aerial imagery, combined through probabilistic fusion. Ahmad et al. (Ahmad et al. 2022) propose a framework for street level 3D object localization that combines multi-view sparse point cloud reconstruction with 2D instance segmentation.

[1]Implementation: `https://github.com/waqar3411/EdgeAwareVH`

However, these methods are difficult to adopt in our setting due to limited access to densely annotated multi-view datasets and noisy annotations coming from public street level imagery datasets like Mapillary (Mapillary 2024), which often leads to unreliable detections and degraded localization accuracy.

Two stage geolocalization pipelines have been developed to detect and track dense, multi-class roadside objects from street level video, using visual appearance cues and positional offsets to estimate object locations (Wilson et al. 2022). Roadside object geolocation has also been approached through multi model sensor fusion, where airborne LiDAR is combines with large scale video-log imagery to recover missing camera parameters and improve bearing estimates (Yi et al. 2025). Cross-view geolocalization networks such as GLQINet (Jin et al. 2025) use enhanced global and local attention and spatial partitioning to match drone and satellite imagery across viewpoints. These methods require densely collected high quality imagery rely and pose metadata, making them unsuitable for a setting with significant noise and sparsely collected datasets, typically one panoramic image every 5-10 meters.

The aim of this study is to identify and classify the triangulated candidate intersection positions, referred to as hypotheses, into spurious or true, if closest to the ground truth object locations. This objective naturally motivates the use of the GNNs when the data is structured as a bipartite graph of view and hypothesis nodes. Kajla et al. (Kajla et al. 2024) show that GNNs are effective for classification on structured graph data, as well as feasible compared to alternative graph similarity methods. GNNs have been successfully applied beyond generic classification task to structured document analysis. In particular, Renton et al. (Renton et al. 2021) model floor-plan images as region adjacency graphs and exploit edge information within GNNs to jointly perform node classification and symbol extraction. Eliasof et al. (Eliasof and Treister 2024) proposed GLGNN that uses GAT (Veličković et al. 2018) and GCN (Kipf and Welling 2017) as backbones, that jointly learn node and label representation to leverage both local and global information, leading to improved node classification.

The surveys relating to multi-view clustering (Fang et al. 2023) and multi-view learning (Xiao et al. 2024) highlight the use of deep GNN-based frameworks to fuse complementary information across multiple data sources for unified representation learning. Nassar et al. (Nassar et al. 2020) adopt a GNN to re-identify objects across multiple views by leveraging both visual appearance features and camera pose information. This strategy enable accurate geolocation of objects from images in real-world environments. This GNN is used for cross-view object re-identification exploiting fully annotated multi-view data.

Several datasets have been previously proposed for multi-view object geolocation, primarily targeting structured roadside objects. ARTSv2 (Wilson et al. 2022) focuses on the classification and geolocation of traffic signs from street level imagery, while (Nassar et al. 2020) introduces the Pasadena Multi-View dataset for geolocating and re-identifying street-side trees. GeoSign leverages Mapillary imagery for traffic sign geolocation (Nassar and Lefèvre 2019). More recently, the Auto Arborist dataset proposes a large-scale benchmark for multi-view urban forest monitoring using paired aerial and street level imagery under significant domain shift (Beery et al. 2022). However, Auto Arborist is designed for species classification and domain generalization rather than precise object-level geolocation, and it relies on city-scale tree census annotations rather than accurate per-object triangulation ground truth. Collectively, these datasets illustrate the growing interest in multi-view object geolocation under real-world conditions, but they differ substantially in object categories, data acquisition pipelines, and problem formulations (Wilson et al. 2023).

We focus our investigation on a practical scenario where the detection is sought for a custom roadside object class. Hence the detections are weakly supervised, and yield uncertain depth and bearing estimates, making appearance cues unreliable. With no annotated multi-view pairings, we rely purely on geometric reasoning over triangulated hypotheses. To score these effectively, we develop a graph-based framework that integrates the complementary strengths of graph attention (GAT) (Veličković et al. 2018) and edge-conditioned message passing (ECC) (Si-

monovsky and Komodakis 2017). More specifically, graph attention enables adaptive weighting of reliable view–hypothesis (VH) relationships, while edge-aware convolutions incorporate geometric cues such as intersection distance and bearing error. Unified in the edge-aware view-hypothesis (EdgeAwareVH) layer, these mechanisms allow the model to suppress spurious hypotheses and localize objects robustly despite noisy detections and weak supervision. Furthermore, we incorporate OSM flags to provide additional regularizing GIS layer to penalize intersections that fall inside buildings, waterways, or rail tracks, improving interpretability in the real-world urban environments.

The primary contribution of this study is a GNN framework, coined EdgeAwareVH, that combines graph attention with edge-conditioned message passing to jointly exploit node features and geometric edge attributes for robust roadside object geolocation. We further incorporate OSM-derived spatial context to suppress invalid intersections and enhance localization stability. Extensive experiments show that EdgeAwareVH outperforms both the MRF-based triangulation baseline and attention-based models such as Set Transformer (Lee et al. 2019), while providing better scalability and stronger resilience to distance and bearing noise.

6.2 Proposed Method

We first formulate the triangulation problem and then construct an appropriate GNN model to address it.

6.2.1 Triangulation problem

The raw input consists of multi-view observations of roadside objects, where each view is annotated with its camera coordinates (lat, lon), along with the estimated object bearing θ_b and distance d. To enable stable metric-scale geometric reasoning, all geographic coordinates are transformed into a local ENU frame, and angular quantities are expressed in vector form. The transformation from geographic coordinates

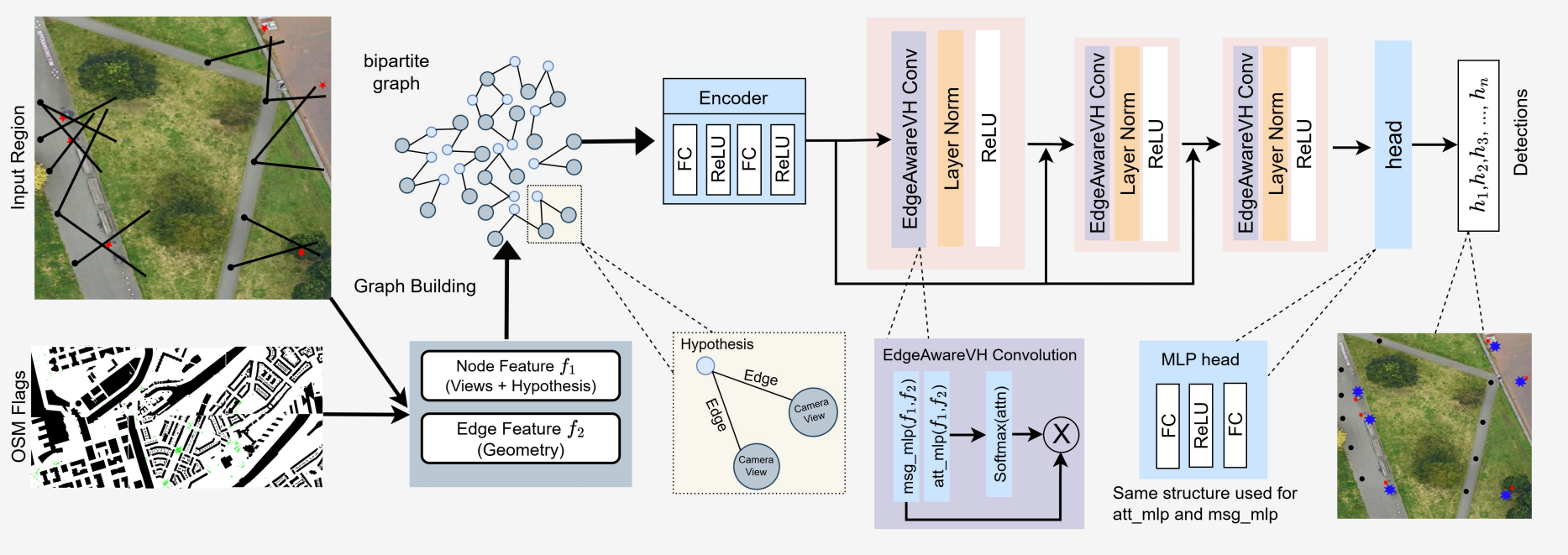


Figure 6.1: Overview of the EdgeAwareVH model. The input includes view, hypothesis and edge features. These inputs form a bipartite graph connecting views to their corresponding triangulated hypotheses. An encoder processes the node features, followed by stacked EdgeAwareVH convolution blocks (EdgeAwareConv + LayerNorm + ReLU). A final MLP head outputs a score for each hypothesis, allowing the model to select the most viable object locations.

(lat, lon) to the local ENU frame is performed by first converting the coordinates into the Earth-Centered, Earth-Fixed (ECEF) reference frame, followed by a rotation from ECEF to ENU about a fixed reference origin located near the center of the scene:

$$\begin{bmatrix} E \\ N \end{bmatrix} = \mathbf{R}_{\text{ECEF}\to\text{ENU}} \left(\begin{bmatrix} X \\ Y \end{bmatrix} - \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \right),$$

where (X, Y) and (X_0, Y_0) denote the ECEF coordinates of the point of interest and the reference origin, respectively, and $\mathbf{R}_{\text{ECEF}\to\text{ENU}}$ is the rotation matrix that aligns the global ECEF axes with the local tangent plane. The resulting EN coordinates enable direct Euclidean operations such as triangulation and distance computations in meters. Camera orientation in the horizontal plane is given by the bearing angle θ measured clockwise from north. To avoid the discontinuity at 360° and to make angular information differentiable, we employ sine and cosine thereof instead.

The candidate object locations are estimated by intersecting the viewing rays from pairs of observations $\mathcal{V}_i$ and $\mathcal{V}_j$, producing intersection points, referred to as hypotheses $\mathcal{H}_k$. The candidate hypotheses are then filtered by removing geometrically unstable triangulated locations that would typically lead to excessive error in position estimation. Specifically, we discard view pairs that are too close by enforcing:

$$\|\mathcal{V}_i - \mathcal{V}_j\|_2 > d_{\min},$$

where each view is represented by the EN coordinates of the camera position where the image has been recorded $\mathcal{V}_i = (E_i, N_i)$ and $\mathcal{V}_j = (E_j, N_j)$, and $d_{\min} = 3\,\text{m}$. We also discard hypotheses that are located unrealistically far from the cameras by requiring:

$$\|\mathcal{H}_k - \mathcal{V}_i\|_2 < d_{\max},$$

where $\mathcal{H}_k$ denotes the EN coordinates of the hypothesis (E_k, N_k) associated with view (camera location) $\mathcal{V}_i$, and $d_{\max} = 30\,\text{m}$. Finally, intersections arising from near-parallel or opposite viewing directions are removed by constraining the angular

difference between the view bearing Θ_{ij} to satisfy $10° < \Theta_{ij} < 170°$, ensuring that only geometrically reliable view pairs are retained.

6.2.2 Graph Construction and GNN

Each object in the dataset is represented by a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{H}, e)$, where $\mathcal{V}$ denotes the set of camera views, $\mathcal{H}$ is the set of triangulated hypotheses and e denotes the edges connecting each view to its associated hypotheses. This representation enable the model to learn the spatial geometry and view consistency.

Node Types and Features

The graph nodes are described by the following features $\mathcal{V} = \Big[E,\ N,\ \sin\theta,\ \cos\theta,\ d\Big]$, and $\mathcal{H} = \Big[E,\ N,\ f^{(\mathrm{OSM})}\Big]$. Here, E and N are the local ENU coordinates of the node and $f^{(\mathrm{OSM})} \in \{0, 1\}$ is an auxiliary OSM-derived flag.

Edges Connectivity and Features

Each edge connects a view node $\mathcal{V}_i$ to its associated hypothesis node $\mathcal{H}_k$ and contains geometric features: $e_{ik} = \Big[d_i, \Delta_{ik}, \varepsilon_{ik}\Big]$, where d_i is the estimated monocular camera-to-object depth, $\Delta_{ik} = \|\mathcal{H}_k - \mathcal{V}_i\|$ is the geometric distance between the camera and the associated hypothesis and ε_{ik} is the bearing misalignment that measures the angle between a camera's viewing direction, which is estimated based on image segmentation, and the geometric direction from that camera to a hypothesis. For a camera i at (E_i, N_i) with bearing θ_i, and a hypothesis k at (E_k, N_k), the bearing misalignment is defined as

$$\varepsilon_{ik} = \arccos\left(\frac{(E_k - E_i)\,\sin\theta_i + (N_k - N_i)\,\cos\theta_i}{\sqrt{(E_k - E_i)^{\,2} + (N_k - N_i)^{\,2}}}\right).$$

Graph Labels

Each constructed graph is unique and corresponds to a specific configuration of real-world objects, where multiple hypothesis nodes represent candidate triangulated

positions. Among these, for each true object only one hypothesis that is geometrically closest to its location is labeled 1 while the all the remaining are assigned label 0.

The proposed model draws from the strengths of GAT (Veličković et al. 2018) and ECC (Simonovsky and Komodakis 2017) models. In GAT, nodes are updated based on the weighted sum of the neighbor features and the attention weights depend on the similarity between the connected nodes. Generally node updates are written as:

$$h_k = \sigma\left(\sum_{i \in \mathcal{N}(\mathcal{H}_k)} \alpha_{ik} \mathbf{W}(\mathcal{V}_i)\right),$$

where $\mathbf{W}$ is a learnable weight matrix corresponding to the view node $\mathcal{V}_i$, $\sigma(\cdot)$ is a non-linear activation (e.g., ReLU), and α_{ik} is the normalized attention coefficient between $\mathcal{V}_i$ and $\mathcal{H}_k$. $\mathcal{N}(\mathcal{H}_k)$ is the set of all other hypotheses that share one view with $\mathcal{H}_k$. Note that in case of a simple two views intersection hypothesis $\mathcal{H}_s$ without any other views nearby, $\mathcal{N}(\mathcal{H}_s) = \{\mathcal{H}_s\}$.

GAT learns which neighbor is more informative by using the nodes features. However, it does not consider the edge attributes explicitly. In our case, geometric relations, like distance between the view node and the hypothesis, are crucial but cannot be learned through the node features alone. To overcome this limitation, the ECC methodology can be adopted to capture the edge attributes as well. In ECC the information exchange between the nodes are modulated by the connecting edge features e_{ik}. The node update in ECC is given by:

$$h_k = \sigma\left(\sum_{i \in \mathcal{N}(\mathcal{H}_k)} \mathbf{W}(e_{ik})\, \mathcal{V}_i\right).$$

ECC allows the model to incorporate geometric relationships into message propagation, but it lacks the dynamic weighting mechanism of GAT. In ECC, all neighbors contribute through a deterministic modulation of e_{ij}, without learning their relative importance among competing hypotheses.

In EdgeAwareVH we combine both mechanisms:

$$h_k = \sigma\left(\sum_{i \in \mathcal{N}(\mathcal{H}_k)} \alpha_{ik}\left[\mathbf{W}_1\big(\psi(\mathcal{V}_i)\big) + \mathbf{W}_2\big(\psi(e_{ik})\big)\right]\right),$$

with $\psi(\cdot)$ being a Multilayer Perceptron (MLP) encoding the node and edge attributes, and α_{ik} is the attention coefficient of both $\mathcal{V}_i$ and e_{ik}. The resulting model learns to assign higher weights to view nodes that strongly support a given hypothesis while down-weighting or ignoring views affected by depth errors or near-parallel directions. EdgeAwareVH preserves spatial interpretability, as the attention weights and the edge embeddings retain direct geometric meaning.

EdgeAwareVH Model Architecture

The input to the model is the node feature matrix $X \in \mathbb{R}^{N \times d}$, where each of the N rows represent either a view or hypothesis node, $d = 6$. An encoder module transforms these raw features into a latent representation that feeds into the subsequent message-passing layers. This encoder consist of two fully connected layers, each followed by ReLU activation, see Figure 6.1.

The encoded node embeddings are then passed to the EdgeAwareVH convolution layers. Each layer propagates information from a view node i to a hypothesis node k using two lightweight MLPs: a message MLP that computes edge-conditioned messages

$$m_{ik} = \mathrm{MLP}_{\mathrm{msg}}([\mathcal{V}_i, e_{ik}]),$$

and an attention MLP that predicts the corresponding attention coefficient:

$$\alpha_{ik} = \phi\big(\mathrm{MLP}_{\mathrm{att}}([\mathcal{V}_i, e_{ik}])\big),$$

where ϕ denotes a softmax function. The attention weights are normalized across all incoming edges to ensure that messages are combined adaptively based on geometric consistency.

Each EdgeAwareVH convolution layer is then followed by Layer normalization and *ReLU* activation and a residual connection ensuring stable gradient propagation through deeper networks. After several propagation layers, the model produces refined embeddings given by:

$$z_k = \mathrm{MLP}_{head}(h_k + \sum_{i \in N(i)} \alpha_{ik} m_{ik}),$$

for all hypothesis nodes. The MLP_{head} consists of two fully connected layers with an intermediate ReLU activation and is optimized by cross-entropy loss function given by:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{k \in \mathcal{N}(\mathcal{H}_k)} \mathbf{1}(k = y^*) \log \frac{\exp(z_k)}{\sum_k \exp(z_k)},$$

where z_k is the logit corresponding to hypothesis k, y^* denotes the index of the hypothesis closest to the ground truth location, and $\mathbf{1}(\cdot)$ is the indicator function that equals 1 when its argument is true and 0 otherwise.

The overall layer configuration of the proposed GNN model is summarized in Table 6.1.

Table 6.1: Layer configuration of the EdgeAwareVH model, showing node encoding ($N = N_v + N_h$), geometric message passing, and final scoring stage.

Operation	**Layer structure**	**Output**
Encode node features	$2 \times$ (FC $\rightarrow$ ReLU)	[N,128]
GNN module	$3 \times$ EdgeAwareVH layer ($\mathrm{MLP}_{msg} \rightarrow \mathrm{MLP}_{att} \rightarrow$ Softmax)	[N,128]
Normalize + residual	LayerNorm $\rightarrow$ ReLU	[N,128]
Scoring head	FC $\rightarrow$ ReLU $\rightarrow$ FC	[N_h,1]
Loss	CrossEntropy	Scalar

6.3 Experimental Setup

6.3.1 Datasets

Despite several proposed datasets for multi-view object geolocation, no widely accepted benchmark exists for quantitative comparisons (Wilson et al. 2023). Evalua-

tion protocols in this field remain fragmented: definitions of true and false positives, object association rules, and distance-error metrics differ substantially across studies, leading to results that are inherently dataset- and implementation-specific. Notably, some relevant datasets like Pasadena Multi-View (Nassar et al. 2020) and Auto Arborist (Beery et al. 2022) are not publicly accessible. As a consequence, direct comparison with previously reported numbers is not practical. In the absence of a common benchmark datasets, we employ publicly available Mapillary and OSM data, which are openly accessible for future comparisons.

To train EdgeAwareVH, a synthetic dataset was generated over a $20\,\text{km} \times 20\,\text{km}$ region centered around Amsterdam at 52.38488 lat, 4.89346 lon.

We use public wastebins as target objects, with ground-truth locations sourced from OSM (OpenStreetMap Contributors 2024). We employ Mapillary API (Mapillary 2024) to collect all the panoramic images camera locations in the area. For each bin, all Mapillary viewpoints within a 10 m radius are collected, providing multiple observations. For each view, the object bearing is computed from the relative angle between the camera and the bin, and distance is obtained as their Euclidean distance. Each bin is viewed from multiple viewpoints, and all pairwise intersections are computed using the triangulation, typically producing multiple location hypotheses per object. To add spatial context, OSM is rasterized, and each hypothesis receives a binary flag indicating if it is plausible (1), i.e. outside building, waterway and railway footprints, or unrealistic (0).

For testing, we use a separate $500\,\text{m} \times 500\,\text{m}$ region of Amsterdam centered at 52.38923 lat, 4.90903 lon. The same data generation and annotation pipeline was applied to both training and test sets.

6.3.2 Experimental Setup

The real-world conditions involve detections that are affected by false positives, as well as inaccurate bearing and distance estimates. In this work, we train on an easily scalable simulated dataset, that does not require preliminary manual annotation.

We thus need to inject comparable noise amount into the dataset, to mimic realistic data collection. Firstly, only 80% of the views per object are retained by randomly dropping 20% to simulate occlusions and detector false negatives. Then, an additional 10% of spurious views are synthesized by assigning random depths (5–25 m) and random object bearings (0°–360°), representing non-systematic false positives. Finally, Gaussian noise is added to the remaining bearings and depths to simulate typical prediction inaccuracies. We thus put together a collection of progressively noisier training configurations:

- **Setting $\mathbf{N_0}$**: Retain 80% of views and add 10% noisy views with clean depth and bearing.

- **Setting $\mathbf{N_1}$**: Same as N_0, with added perturbations: depth + $\mathcal{N}(0, 0.5\sqrt{d})$ and bearing + $\mathcal{N}(0, 1.5\, O_b)$.

- **Setting $\mathbf{N_2}$**: Same as N_0 with more noise: depth + $\mathcal{N}(0, \sqrt{d})$ and bearing + $\mathcal{N}(0, 3\, O_b)$.

- **Setting $\mathbf{N_3}$**: Same as N_0 with even more noise: Higher noise level: depth + $\mathcal{N}(0, 2\sqrt{d})$ and bearing +
  $\mathcal{N}(0, 6\, O_b)$.

- **Setting $\mathbf{N_4}$**: Same as N_3, but retain only 60% of the true views and add 20% random noisy views.

Here d and O_b are the camera-to-object distance and bearing, respectively. The noise configurations are designed to assess the model's ability to generalize under realistic conditions. We evaluate performance on three test sets: a real-world dataset with detections obtained using the weakly supervised detector (Ahmad and Krylov 2024) on Mapillary panoramic images (D_{Real}), a simulated set with injected noise ($D_{Sim+Noise}$), and a clean simulated set without noise (D_{Sim}).

6.3.3 Model Training

The proposed model is trained for 50 epochs using a batch size of 4, where each batch consist of four object-level graphs, each representing a single object with its associated view nodes, hypothesis nodes and edges. The model was optimized using the Adam optimizer with a learning rate of 1×10^{-3}. Training was carried out on NVIDIA Titan X GPU. During validation, the model assigns a score to each hypothesis, and the top-K hypotheses are averaged to estimate the object location. The mean localization error between this estimate and the ground truth serves as the validation metric, and the model with the lowest error is selected for final testing.

6.3.4 Model Testing

The model is evaluated on a held-out test set. Since ground-truth locations are unavailable, all views in the region are triangulated pairwise to generate candidate hypothesis. Admissible intersections formed from the same view ray of the camera are grouped as a single potential object and a graph is built. The model assigns a confidence score for every hypothesis node. Rather than relying solely on the top prediction, the top-20% hypotheses are retained and their ENU coordinates are averaged to obtain the final estimate. Multiple graphs may correspond to the same object, several nearby hypotheses can appear. To merge them, a DBSCAN hierarchical clustering is applied as a post processing step. The centroid of each cluster using a fixed neighborhood radius parameter K, is reported as the final predicted object location:

$$\hat{p} = \frac{1}{|N(\mathcal{H}_k)|} \sum_{k \in N(\mathcal{H}_k)} h_k,$$

where $N(\mathcal{H}_k)$ denotes the set of hypotheses within the same cluster. The value of K is set to be under 5 meters, based on the assumption that objects in the scene are spatially sparse and typically separated by at least this distance.

Importantly, we also evaluate our method on a non-simulated test set generated

using a weakly supervised CNN-based detector (Ahmad and Krylov 2024). This is key to assess the real world geolocation performance for roadside objects.

6.4 Results Analysis

The simulated training set contains approximately 33,000 objects and 250,000 view images. Such amount of data is challenging to obtain with real street level panoramas-based detections, but easily accessible given a realistic GIS set of object locations. Our experiments showed that reducing this set by 50% resulted in approx. 20% drop in precision, highlighting the importance of sufficient training data for a GNN; increasing the set by 10% yielded negligible improvement and introduced overfitting. The simulated test set (D_{Sim}) contains 160 objects and about 3,100 view images with no overlap with the training, while the real-world test set (D_{Real}) includes the same number of objects and about 1600 views.

6.4.1 Noise setting for training

We first analyze the noise settings $N_0 - N_4$ introduced above to assess which leads to the best generalization for the real world data. To this end we analyze the performance on D_{Real} and ($D_{Sim+Noise}$) geolocation test sets, see Table 6.2. The overall best performance on D_{Real} is obtained with N_2, yielding the highest precision (75.8%), highest recall (85.6%), and lowest error (1.93 m). Performance degrades slightly as noise increases, with precision and recall dropping by several few points. It can also be easily observed that data without noise N_0 leads to very poor generalization. We thus choose N_2 as the representative noise configuration for our experiments.

6.4.2 Benchmark comparisons

We compare the proposed method against (i) the MRF-based triangulation approach (Krylov et al. 2018), which scores intersections using hand-crafted energy

Table 6.2: EdgeAwareVH performance with different noise configurations on real D_{Real} and noisy simulated $D_{Sim+Noise}$ test sets.

Noise	D_{Real}			$D_{Sim+Noise}$		
	Precision	Recall	Error (m)	Precision	Recall	Error (m)
N_0	64.1	81.0	3.864	**87.6**	85.1	0.323
N_1	73.7	84.4	1.951	87.1	**87.9**	0.903
N_2	**75.8**	**85.6**	**1.930**	87.1	86.2	**0.901**
N_3	71.7	81.7	1.952	83.3	83.4	0.931
N_4	71.5	83.1	2.003	79.2	79.1	1.301

Table 6.3: Comparative performance of EdgeAwareVH against MRF (Krylov et al. 2018) and Set Transformer (Lee et al. 2019), evaluated on both real and simulated test sets.

Dataset	Model	Precision	Recall	Error (m)
D_{Real}	MRF	64.25	79.90	2.541
	Set Transformer + OSM	72.10	83.50	1.902
	EdgeAwareVH	74.00	85.00	1.931
	EdgeAwareVH + OSM	**75.80**	**85.60**	**1.930**
$D_{Sim+Noise}$	MRF	73.60	82.32	1.102
	Set Transformer + OSM	85.20	85.30	0.912
	EdgeAwareVH	86.40	86.20	0.903
	EdgeAwareVH + OSM	**87.10**	**87.00**	**0.901**
D_{Sim}	MRF	85.12	93.20	1.910
	Set Transformer + OSM	90.10	96.00	0.852
	EdgeAwareVH	89.00	**96.70**	0.823
	EdgeAwareVH + OSM	**91.10**	**96.70**	**0.820**

terms; and (ii) the Set Transformer (Lee et al. 2019) as a strong attention-based baseline. Although the latter has not previously been applied to this task, it provides a meaningful comparison for assessing how an alternative attention mechanisms performs relative to EdgeAwareVH. As can be seen in Table 6.3, EdgeAwareVH achieves the best performance on D_{Real}, reaching 74% precision, 85% recall, and a 1.93 m localization error. Incorporating lightweight spatial context yields additional gains: the OSM flag increases precision to 75.8% and recall to 86%, showing that simple map-derived cues help suppress invalid hypotheses and improve performance.

The MRF baseline (Krylov et al. 2018) shows the weakest performance across all datasets, largely due to its reliance on hand-crafted energy terms that lack the flexibility of learned models. It may possibly achieve betters results with manual parameter selection, and also custom integration with OSM. The Set Transformer

Table 6.4: Performance of Proposed method against Set Transformer (Lee et al. 2019) on D_{Real} at varying detection distance thresholds (in meters)

Model	Metric (%)	3m	5m	7m
Set Transformer + OSM	Precision	72.1	75.2	76.7
	Recall	83.5	85.0	87.3
EdgeAwareVH + OSM	Precision	75.8	77.8	79.2
	Recall	85.6	86.1	89.6

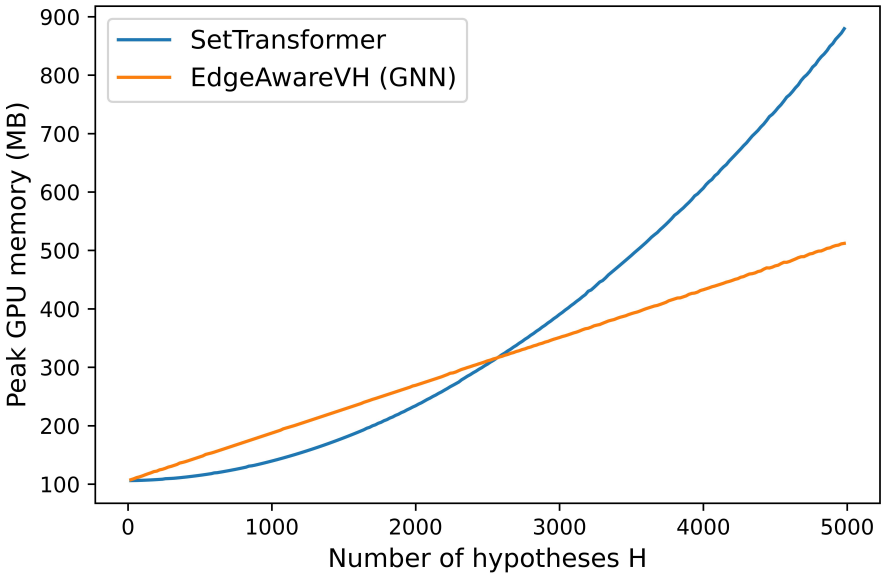


Figure 6.2: GPU memory usage: linear with respect to the number of hypotheses for EdgeAwareVH, and quadratic for Set Transformer (Lee et al. 2019).

performs competitively, only three points behind on D_{Real}, but incurs significantly higher computational cost due to its quadratic attention complexity. The Set Transformer does not incorporate any edge features and therefore cannot exploit geometric relationships between views and hypotheses. In contrast, EdgeAwareVH learns these relations directly from the graph structure, avoiding explicit energy-term design while outperforming both baselines.

Table 6.4 reports precision and recall at different spatial thresholds K, where predictions within K meters of the ground truth are counted as true positive. Increasing K values relaxes the matching criterion and generally boost the performance. Across all thresholds, EdgeAwareVH consistently outperforms the Set Transformer demonstrating higher localization accuracy and robustness to spatial noise.

6.4.3 Computation Complexity

Figure 6.2 compares GPU memory consumption for Set Transformer and EdgeAwareVH as the number of hypotheses H increases. EdgeAwareVH exhibits near-linear scaling consistent with its $\mathcal{O}(V \cdot H)$ message-passing complexity and remains below 500 MB even at $H \geq 5000$. In contrast, Set Transformer shows much steeper growth due to its quadratic $\mathcal{O}(H^2)$ attention matrices, exceeding 800 MB at similar scales. Importantly on D_{Real} we observed the top-5 objects by the numbers of hypotheses have been detected by 100-150 views and had 3000-6000 hypotheses associated with them. These results indicate that while Set Transformer is feasible for small hypothesis sets, the memory consumption becomes a problem at larger scales, whereas EdgeAwareVH remains more efficient and scalable for large triangulation scenarios.

6.4.4 Ablation Study

Table 6.5 presents an ablation study on the D_{Real} test set, illustrating the impact of the individual message-passing mechanisms. The results highlight the importance of incorporating geometric and spatial features on both nodes and edges, compared with the methods that ignore these cues. The comparison include GAT (Veličković et al. 2018), ECC (Simonovsky and Komodakis 2017) and the proposed EdgeAwareVH, which integrates both attention and edge-conditioned aggregation. All three methods integrate OSM flags for better spatial awareness.

Table 6.5: Ablation comparisons for EdgeAwareVH on D_{real}. The evaluated variants include Graph Attention Networks (GAT) (Veličković et al. 2018), Edge-Conditioned Convolutions (ECC) (Simonovsky and Komodakis 2017), and the proposed EdgeAwareVH, each integrated with OSM-based priors

Method	**Precision (%)**	**Recall (%)**	**Error (m)**
GAT + OSM	65.4	82.0	2.374
ECC + OSM	67.7	83.7	2.273
EdgeAwareVH + OSM	**75.8**	**85.6**	**1.930**

6.5 Conclusion

This chapter introduces EdgeAwareVH, a scalable and geometry-driven GNN approach for robust roadside object geolocation from weekly supervised street level imagery. Practically, any existing GIS record of the asset of interest can be used to train the model without the need for costly image matching annotations. By formulating the multi-view object triangulation as view-hypothesis graph and integrating the edge-aware attention with geometric message passing, the proposed method effectively distinguishes viable from spurious intersections. The incorporation of the OSM-derived information further enhances the model performance by penalizing hypotheses inconsistent with the spatial context. Our experiments with wastebins geolocation show that EdgeAwareVH outperforms both traditional geometric baseline as well as attention-based alternatives. The future work may focus on integrating the visual features and self-attention blocks to strengthen view-to-view relationships and further enhance robustness.

Chapter 7

Discussion

This chapter concludes the research by summarizing key contributions, weakly supervised localization, similarity-based noise filtering, and graph-based spatial reasoning, and how these methods address roadside object geolocation under weak supervision. It identifies main limitations: coarse localization, noisy cross-view associations, and scalability in large urban environments. Finally, it suggests future research: richer training data, improved visual representations, multi-modal integration, and scalable graph-based methods.

7.1 Conclusion

This thesis addressed the problem of automated geolocation of roadside objects from street-level imagery under weak supervision. The central challenge is estimating accurate object locations from imperfect inputs. These include noisy detections, uncertain geometric measurements, and unreliable cross-view associations. In large urban settings, sparse annotations, inconsistent observations, and multi-view geometric ambiguity further complicate the task. As a result, existing approaches that rely solely on geometric or learning-based methods often struggle to achieve robust, scalable results.

A key research gap identified is the limited availability and reliability of object-level annotations in street-level imagery datasets. Traditional supervised methods

require detailed annotations, which are expensive and difficult to obtain at scale. To address this, the thesis proposed a weakly supervised object localization framework that uses image-level labels and deep classification models. Grad-CAM, a technique that highlights image regions important to a model's predictions, is combined with unsupervised segmentation, which separates objects in images without labeled data. This approach extracts object locations without requiring bounding-box annotations. These detections are then integrated with camera metadata like geolocation (the image's location), bearing (the direction the camera is facing), and depth estimates (distance information). This serves as the basis for multi-view geometric reasoning, which combines information from multiple views to better understand object positions.

However, weak supervision introduces noise, such as false positives and inaccurate detections. This leads to a critical gap: the lack of reliable ways to associate object observations across views. Incorrect associations directly affect triangulation, resulting in spurious intersections and degraded localization accuracy. To address this, the thesis redefined cross-view association as a similarity learning problem. It introduced a Siamese network to learn visual similarity between image pairs, making it easier to identify consistent observations across views. To further improve robustness, a probabilistic outlier detection framework based on Beta mixture modeling was developed. This explicitly models the distribution of similarity scores and filters out unreliable pairs, thereby reducing the impact of noisy supervision on later tasks.

The integration of visual similarity with geometric reasoning was explored to improve triangulation-based localization. Traditional triangulation relies heavily on geometric consistency and is sensitive to noise in detections and measurements. The proposed approach uses learned similarity scores as an extra cue during triangulation. By combining visual and geometric information, it suppresses incorrect associations and reduces spurious intersection points. This leads to more reliable object locations and better geolocation accuracy in the presence of noisy observations.

Despite improvements, statistical and geometric methods still cannot capture

complex dependencies between observations. This highlights another gap: handcrafted models cannot fully use the relationships among views, detections, and candidates. To address this, the thesis introduced a graph-based learning framework using Graph Neural Networks. Here, multi-view observations and triangulated hypotheses are nodes, and their relationships are edges that encode geometric and visual consistency. The edge-aware graph neural network reasons over these relationships, combining multiple information sources to find the most plausible object locations. This approach improves robustness to noise and scales better to real-world scenarios.

This thesis establishes a robust framework for geolocating roadside objects. It integrates weakly supervised detection, similarity learning, probabilistic noise modeling, geometric triangulation, and graph-based reasoning. The main contribution is showing that reliable geolocation is possible even under weak supervision. This is achieved by combining visual and geometric cues and accounting for noise and uncertainty. The proposed methods demonstrate that similarity learning improves cross-view association, while graph-based reasoning boosts robustness and scalability of localization.

7.2 Limitations

Building on these contributions, it is important to consider the limitations of the proposed approaches. The thesis shows robust roadside object geolocation is possible with weak supervision, but challenges remain for complex scenarios. The main limitations are as follows:

- First, the weakly supervised localization framework uses image-level labels and Grad-CAM activation maps to estimate object locations. This reduces the need for detailed annotations but can result in coarse localization. The output may include false positives or imprecise boundaries, which can affect later stages. This is more problematic in cluttered urban scenes with similar objects

close together. Expanding the training set with more diverse samples could address this, covering viewpoint, scale, lighting, occlusion, and background variations.

- Second, the effectiveness of the similarity learning framework depends on the quality of the learned visual representations. While the Siamese network and Beta mixture modeling improve robustness, visually similar but physically different objects may still be incorrectly matched, especially under viewpoint or lighting changes. These incorrect matches can affect the triangulation stage and lower geolocation accuracy. Training the model with more diverse samples, hard negatives, and various conditions could help address this.

- Another limitation is the reliance on geometric cues such as camera position, bearing, and estimated depth. These are often noisy, especially when derived from monocular imagery or imperfect metadata. Since triangulation is sensitive to such errors, even small inaccuracies can cause large deviations in object location estimates. Although the proposed methods include probabilistic reasoning and similarity-based filtering, performance still depends on the quality of geometric information.

- The graph-based learning framework improves the modeling of relationships between observations, but increases computational complexity. As the number of views and candidates increases, the constructed graph grows, impacting scalability and memory usage. While the model is more robust, handling very large urban datasets with dense observations may need further optimization. Solutions include graph sparsification, filtering candidates, mini-batch processing, and memory-aware design.

- A further limitation is the lack of standardized evaluation metrics and benchmarking protocols for object-level geo-localization. Unlike related fields, there is no widely accepted metric to capture both geometric accuracy and association correctness. This thesis used task-specific criteria, but direct comparison

with other approaches remains hard due to differences in datasets, evaluation setups, and assumptions. This makes it difficult to establish a unified comparison framework across methods. This issue is also highlighted in the survey by D. Wilson (Wilson et al. 2023), where the object-level geo-localization evaluation is identified as an open challenge. In particular, the survey emphasizes that evaluation protocols often vary significantly across studies, with different works using inconsistent distance thresholds, matching criteria, and dataset-specific definitions of correctness. As a result, reported performance metrics are not always directly comparable, and improvements in one setting may not generalize to others. This lack of standardization limits the ability to objectively assess progress and identify the true strengths and weaknesses of different approaches.

7.3 Future Work

The inherent complexity of the problem motivates several directions for future research and further improvement. Despite the associated challenges, the proposed framework provides a strong foundation for robust roadside object geolocation under weak supervision and highlights key areas for future research. The following directions outline possible extensions of this work:

- **Incorporation of additional data modalities:** Combining street-level imagery with aerial or satellite data could provide complementary spatial context, helping to resolve ambiguities that are difficult to address using a single modality. In addition, incorporating LiDAR data could provide accurate depth and structural information about the surrounding scene, thereby strengthening geometric reasoning and improving object localization. Similarly, integrating multimodal cues, such as image features, depth information, LiDAR point clouds, and aerial context, within the Siamese similarity learning framework could support richer feature learning and more effectively distinguish true object

matches from noisy pairwise associations. Leveraging temporal information from video sequences could further improve robustness in dynamic environments.

- **Scalable graph construction with integrated similarity reasoning:** As the number of views and candidate hypotheses increases, the graph size can grow rapidly, affecting computational efficiency and memory requirements. Future work could explore graph pruning, sparsification, or hierarchical graph representations to control complexity while preserving the most informative geometric relationships. At the same time, visual similarity could be incorporated directly into the graph neural network, rather than being treated as a separate filtering step. By jointly reasoning over geometry and appearance within a unified graph-based framework, the model could improve cross-view consistency, reduce noisy associations, and support more scalable deployment on large-scale urban datasets.

- **Learning-based reasoning for improved object geolocation:** Future work could further explore learning-based graph formulations, where object hypotheses are modeled jointly and refined through contextual reasoning rather than being selected through fixed or sequential post-processing steps. By using learnable queries and attention-based matching between observations and candidate hypotheses, the model could jointly reason about multiple possible object locations in a more flexible manner. This would allow the network to suppress redundant hypotheses, handle ambiguous multi-view associations, and directly predict a compact set of reliable object locations. Such an approach could improve the existing graph-based framework by combining geometric reasoning, attention-based hypothesis selection, and end-to-end optimization within a unified model.

- **Improving Beta-SOD through adaptive loss weighting:** A further extension of the Beta-SOD framework could involve designing adaptive loss func-

tions that explicitly account for hard and noisy examples during similarity learning. Instead of relying solely on post-training filtering, the training objective could be modified to penalize false-positive associations more strongly while upweighting reliable true-positive pairs that are difficult to distinguish. This would encourage the Siamese network to learn more discriminative embeddings for challenging cross-view cases, such as visually similar but different objects or true matches observed under strong viewpoint and illumination changes. As a result, hard true positives would be less likely to be incorrectly removed during Beta-SOD filtering, while noisy or inconsistent pairs could be identified more effectively.

- **Improving localization and feature representation through richer visual learning:** In addition to the future directions discussed above, this extension could further address the challenges of weak supervision by improving both Grad-CAM-based localization and similarity estimation. Increasing the scale and diversity of the training data, including a wider range of object instances under different viewpoints, illumination conditions, occlusions, scales, and background settings, could help the model learn more representative features and produce more stable activation maps for object localization. In parallel, enhancing the Siamese network with attention-based or transformer-driven architectures could improve the discriminative power of the learned embeddings. Together, these improvements could support more accurate weakly supervised localization, better handling of visually similar objects, and more reliable cross-view similarity estimation.

Bibliography

Ahmad, Javed, Matteo Toso, Matteo Taiana, Stuart James, and Alessio Del Bue (2022). "Multi-view 3D Objects Localization from Street-Level Scenes". In: *Image Analysis and Processing – ICIAP 2022*. Cham: Springer International Publishing, pp. 89–101. ISBN: 978-3-031-06430-2.

Ahmad, Waqar and Vladimir A. Krylov (2024). "Roadside Object Geolocation from Street-Level Images with Reduced Supervision". In: *2024 32nd European Signal Processing Conference (EUSIPCO)*, pp. 641–645. DOI: 10.23919/EUSIPCO63174.2024.10715092.

Ahmed, Ejaz, Michael Jones, and Tim K. Marks (2015). "An improved deep learning architecture for person re-identification". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3908–3916. DOI: 10.1109/CVPR.2015.7299016.

Aldubaikhi, Ali and Sarosh Patel (2025). "Advancements in Small-Object Detection (2023–2025): Approaches, Datasets, Benchmarks, Applications, and Practical Guidance". In: *Applied Sciences* 15.22. ISSN: 2076-3417. DOI: 10.3390/app152211882. URL: https://www.mdpi.com/2076-3417/15/22/11882.

Arandjelović, Relja, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic (2018). "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6, pp. 1437–1451. DOI: 10.1109/TPAMI.2017.2711011.

Atienza, N, J Garcia-Heras, and JM Munoz-Pichardo (2006). "A new condition for identifiability of finite mixture distributions". In: *Metrika* 63.2, pp. 215–221.

Bai, Yan, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan (2018). "Group-Sensitive Triplet Embedding for Vehicle Reidentification". In: *IEEE Transactions on Multimedia* 20.9, pp. 2385–2399.

Beery, Sara, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang (June 2022). "The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21294–21307.

Benedek, Csaba, Xavier Descombes, and Josiane Zerubia (2012). "Building Development Monitoring in Multitemporal Remotely Sensed Image Pairs with Stochastic Birth-Death Dynamics". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.1, pp. 33–50. DOI: 10.1109/TPAMI.2011.94.

Biljecki, Filip and Kohei Ito (2021). "Street View Imagery in Urban Analytics and GIS: A Review". In: *Landscape and Urban Planning* 215, p. 104217. DOI: 10.1016/j.landurbplan.2021.104217.

Branson, Steve, Jan Dirk Wegner, David Hall, Nico Lang, Konrad Schindler, and Pietro Perona (2018). "From Google Maps to a fine-grained catalog of street trees". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 135, pp. 13–30. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2017.11.008.

Braso, Guillem and Laura Leal-Taixe (June 2020). " Learning a Neural Solver for Multiple Object Tracking ". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 6246–6256. DOI: 10.1109/CVPR42600.2020.00628.

Cao, Renyu, Jianfeng Zhu, Wei Tu, Qingquan Li, Jie Cao, and Bing Liu (2018). "Integrating Aerial and Street View Images for Urban Land Use Classification". In: *Remote Sensing* 10.10, p. 1553. DOI: 10.3390/rs10101553.

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko (2020). "End-to-End Object Detection with Transformers". In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*. Glasgow, United Kingdom: Springer-Verlag, pp. 213–229. ISBN: 978-3-030-58451-1. DOI: 10.1007/978-3-030-58452-8_13.

Chaabane, Mohamed, Lionel Gueguen, Ameni Trabelsi, Ross Beveridge, and Stephen O'Hara (2021). "End-to-end Learning Improves Static Object Geo-localization from Video". In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2062–2071. DOI: 10.1109/WACV48630.2021.00211.

Chen, Qi, Wei Wang, Fangyu Wu, Suparna De, Ruili Wang, Bailing Zhang, and Xin Huang (2019). "A Survey on an Emerging Area: Deep Learning for Smart City Data". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 3.5, pp. 392–410. DOI: 10.1109/TETCI.2019.2907718.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). "A simple framework for contrastive learning of visual representations". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 149. ICML'20, p. 11.

Chen, Xiaobo, Haoze Yu, Chuan Hu, and Hai Wang (2024). "Multi-branch feature learning network via global-local self-distillation for vehicle re-identification". In: *IEEE Transactions on Vehicular Technology* 73.9, pp. 12415–12425.

Chen, Yanbei, Xiatian Zhu, and Shaogang Gong (2017). "Person Re-identification by Deep Learning Multi-scale Representations". In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 2590–2600. DOI: 10.1109/ICCVW.2017.304.

Chen, Yukang, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia (June 2023). "VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21674–21683.

Cordts, Marius, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele (2015). "The cityscapes dataset". In: *CVPR Workshop on the Future of Datasets in Vision*. Vol. 2. 1. IEEE, pp. 1–8.

Dalal, N. and B. Triggs (2005). "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

Daryani, Amir Etefaghi, M. Usman Maqbool Bhutta, Byron Hernandez, and Henry Medeiros (June 2025). "CaMuViD: Calibration-Free Multi-View Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1220–1229.

Defense Mapping Agency (1987). *Department of Defense World Geodetic System 1984: Its Definition and Relationships*. Technical Report.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1, pp. 1–22.

Diebolt, Jean and Christian P Robert (1994). "Estimation of finite mixture distributions through Bayesian sampling". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.2, pp. 363–375.

Du, Yuwen, Anning Hu, Zichen Chao, Yifan Lu, Junhao Ge, Genjia Liu, Weitao Wu, Lanjun Wang, and Siheng Chen (Oct. 2025). "RoCo-Sim: Enhancing Roadside Collaborative Perception through Foreground Simulation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26977–26986.

Eliasof, Moshe and Eran Treister (2024). "Global-local graph neural networks for node-classification". In: *Pattern Recognition Letters* 184, pp. 103–110. ISSN: 0167-8655.

Fan, Hehe, Liang Zheng, Chenggang Yan, and Yi Yang (Oct. 2018). "Unsupervised Person Re-identification: Clustering and Fine-tuning". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 14.4. ISSN: 1551-6857. DOI: 10.1145/3243316.

Fang, Uno, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang (2023). "A Comprehensive Survey on Multi-View Clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 35.12, pp. 12350–12368. DOI: 10.1109/TKDE.2023.3270311.

Farooq Bhat, Shariq, Ibraheem Alhashim, and Peter Wonka (2021). "AdaBins: Depth Estimation Using Adaptive Bins". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4008–4017. DOI: 10.1109/CVPR46437.2021.00400.

Felzenszwalb, Pedro F., Ross B. Girshick, David McAllester, and Deva Ramanan (2010). "Object Detection with Discriminatively Trained Part-Based Models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9, pp. 1627–1645. DOI: 10.1109/TPAMI.2009.167.

Feng, Lei, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An (2021). "Can cross entropy loss be robust to label noise?" In: *Proceedings of the international joint conference on artificial intelligence*, pp. 2206–2212.

Galanakis, George, Xenophon Zabulis, and Antonis A Argyros (2024). "Noise-robust person re-identification through nearest-neighbor sample filtering". In: *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, pp. 1–8.

Ge, Yixiao, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li (2020). "Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.

Ghosh, Aritra, Himanshu Kumar, and P Shanti Sastry (2017). "Robust loss functions under label noise for deep neural networks". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1.

Girshick, Ross (2015). "Fast R-CNN". In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.

Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik (2014). "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587. DOI: 10.1109/CVPR.2014.81.

Gupta, Arjun K and Saralees Nadarajah (2004). *Handbook of beta distribution and its applications*. CRC press.

Han, Bo, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama (2018). "Co-teaching: Robust training of deep neural networks with extremely noisy labels". In: *Advances in neural information processing systems* 31.

Hartley, Richard and Andrew Zisserman (2004). *Multiple View Geometry in Computer Vision*. 2nd ed. Cambridge University Press.

Hartley, Richard I. and Peter Sturm (1997). "Triangulation". In: *Computer Vision and Image Understanding* 68.2, pp. 146–157. ISSN: 1077-3142. DOI: https://doi.org/10.1006/cviu.1997.0547.

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick (2017). "Mask R-CNN". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hoffmann, Eric J., Yu Wang, and Matthias Werner (2019). "Model Fusion for Building Type Classification from Aerial and Street View Images". In: *Remote Sensing* 11.11, p. 1259. DOI: 10.3390/rs11111259.

Hofmann-Wellenhof, Bernhard, Herbert Lichtenegger, and Elmar Wasle (2008). *GNSS – Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and More*. 1st ed. Springer Vienna, p. 518. ISBN: 978-3-211-73017-1. DOI: 10.1007/978-3-211-73017-1.

Hu, Sixing, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee (2018). "CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7258–7267. DOI: 10.1109/CVPR.2018.00758.

Jin, Ming, Yizhen Zheng, Yuan-Fang Li, Chen Gong, Chuan Zhou, and Shirui Pan (2021). "Multi-scale contrastive siamese networks for self-supervised graph representation learning". In: *International Joint Conference on Artificial Intelligence 2021*. Association for the Advancement of Artificial Intelligence (AAAI), pp. 1477–1483.

Jin, Xu, Yin Junping, Zhang Juan, and Gao Tianyan (Sept. 2025). "Enhancing cross view geo localization through global local quadrant interaction network". In: *Scientific Reports* 15. DOI: 10.1038/s41598-025-18935-6.

Kajla, Nadeem Iqbal, Malik Muhammad Saad Missen, Mickael Coustaty, Hafiz Muhammad Sanaullah Badar, Maruf Pasha, and Faiza Belbachir (2024). "A histogram-based approach to calculate graph similarity using graph neural networks". In: *Pattern Recognition Letters* 186, pp. 286–291. ISSN: 0167-8655.

Karim, Nazmul, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah (2022). "Unicon: Combating label noise through uniform selection and contrastive learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9676–9686.

Kipf, Thomas N. and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Krylov, Vladimir and Rozenn Dahyot (2018). "Object Geolocation Using MRF-Based Multi-Sensor Fusion". In: *Proceedings of the IEEE International Conference on Image Processing*, pp. 2745–2749.

Krylov, Vladimir A., Eamonn Kenny, and Rozenn Dahyot (2018). "Automatic Discovery and Geotagging of Objects from Street View Imagery". In: *Remote Sensing* 10.5. ISSN: 2072-4292. DOI: 10.3390/rs10050661.

Kuma, Ratnesh, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram (2019). "Vehicle Re-identification: an Efficient Baseline Using Triplet Embedding". In: *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. DOI: 10.1109/IJCNN.2019.8852059.

Kumar, Ratnesh, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram (2019). "A Strong and Efficient Baseline for Vehicle Re-Identification Using Deep Triplet Embedding". In: *Journal of Artificial Intelligence and Soft Computing Research* 10.1, pp. 27–45. DOI: 10.2478/jaiscr-2020-0003. URL: https://doi.org/10.2478/jaiscr-2020-0003.

Lang, Alex H., Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom (June 2019). "PointPillars: Fast Encoders for Object Detection From Point Clouds". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Laumer, David, Lukas Lang, Niels van Doorn, Oisin Mac Aodha, Pietro Perona, and Jan Dirk Wegner (2020). "Geocoding of Trees from Street Addresses and Street-Level Images". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 162, pp. 125–136.

Lee, Juho, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh (2019). "Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 3744–3753.

Li, Heng-Chao, Vladimir A Krylov, Ping-Zhi Fan, Josiane Zerubia, and William J Emery (2015). "Unsupervised learning of generalized Gamma mixture model

with application in statistical modeling of high-resolution SAR images". In: *IEEE Transactions on Geoscience and Remote Sensing* 54.4, pp. 2153–2170.

Li, Jia, Yafei Song, Jianfeng Zhu, Lele Cheng, Ying Su, Lin Ye, Pengcheng Yuan, and Shumin Han (2019). "Learning from large-scale noisy web data with ubiquitous reweighting for image classification". In: *IEEE transactions on pattern analysis and machine intelligence* 43.5, pp. 1808–1814.

Li, Ming, Xinming Huang, and Ziming Zhang (Oct. 2021). " Self-supervised Geometric Features Discovery via Interpretable Attention for Vehicle Re-Identification and Beyond ". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA, pp. 194–204.

Li, Wei, Rui Zhao, Tong Xiao, and Xiaogang Wang (2014). "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159. DOI: 10.1109/CVPR.2014.27.

Li, Wei, Xiatian Zhu, and Shaogang Gong (2018). "Harmonious Attention Network for Person Re-identification". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294. DOI: 10.1109/CVPR.2018.00243.

Li, Xiaotian, Baojie Fan, Jiandong Tian, and Huijie Fan (June 2024). "GAFusion: Adaptive Fusing LiDAR and Camera with Multiple Guidance for 3D Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21209–21218.

Li, Zheng, Yongcheng Wang, Dongdong Xu, Yunxiao Gao, and Tianqi Zhao (2025). "TBNet: A texture and boundary-aware network for small weak object detection in remote-sensing imagery". In: *Pattern Recognition* 158, p. 110976. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2024.110976.

Liang, Tingting, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang (2022). "BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Bel-

grave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 10421–10434. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/43d2b7fbee8431f7cef0d0afed51c691-Paper-Conference.pdf.

Lin, Tsung-Yi, Serge Belongie, and James Hays (2013). "Cross-View Image Geolocalization". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898. DOI: 10.1109/CVPR.2013.120.

Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie (2017). "Feature Pyramid Networks for Object Detection". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944. DOI: 10.1109/CVPR.2017.106.

Liu, Min, Fei Wang, Xueping Wang, Yaonan Wang, and Amit K Roy-Chowdhury (2024). "A two-stage noise-tolerant paradigm for label corrupted person re-identification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.7, pp. 4944–4956.

Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg (2016a). "SSD: Single Shot MultiBox Detector". In: *Computer Vision – ECCV 2016*, pp. 21–37. ISBN: 978-3-319-46448-0.

Liu, Xiaobin, Shiliang Zhang, Qingming Huang, and Wen Gao (2018a). "RAM: A Region-Aware Deep Model for Vehicle Re-Identification". In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. DOI: 10.1109/ICME.2018.8486589.

Liu, Xinchen, Wu Liu, Huadong Ma, and Huiyuan Fu (2016b). "Large-scale vehicle re-identification in urban surveillance videos". In: *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. DOI: 10.1109/ICME.2016.7553002.

Liu, Xinchen, Wu Liu, Tao Mei, and Huadong Ma (2018b). "PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance". In:

IEEE Transactions on Multimedia 20.3, pp. 645–658. DOI: 10.1109/TMM.2017.2751966.

Loce, Robert, Edgar Bernal, Wencheng Wu, and Raja Bala (Dec. 2013). "Computer vision in roadway transportation systems: A survey". In: *Journal of Electronic Imaging* 22. DOI: 10.1117/1.JEI.22.

Lou, Yihang, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan (2019a). "Embedding Adversarial Learning for Vehicle Re-Identification". In: *IEEE Transactions on Image Processing* 28.8, pp. 3794–3807. DOI: 10.1109/TIP.2019.2902112.

Lou, Yihang, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan (June 2019b). "VERI-Wild: A Large Dataset and a New Method for Vehicle Re-Identification in the Wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lowe, David G. (Nov. 2004). "Distinctive Image Features from Scale-Invariant Keypoints". In: *Int. J. Comput. Vision* 60.2, pp. 91–110. ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000029664.99615.94. URL: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Lumnitz, Sebastian, Tim Devisscher, Jérémy Mayaud, Vlado Radic, Nicholas Coops, and Volker Griess (2021). "Mapping Trees Along Urban Street Networks with Deep Learning and Street-Level Imagery". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 175, pp. 144–157.

Ma, Jianbo, Fei Wu, Chunjiang Li, Chuanming Tang, Jianlin Zhang, Zhiyong Xu, Meihui Li, and Dongxu Liu (2024). "G2EMOT: Guided Embedding Enhancement for Multiple Object Tracking in Complex Scenes". In: *IEEE Transactions on Instrumentation and Measurement* 73, pp. 1–14. DOI: 10.1109/TIM.2024.3449952.

Mapillary (2024). *Mapillary: Street-level Imagery Platform*. https://www.mapillary.com.

Melekhov, Iaroslav, Juho Kannala, and Esa Rahtu (2016). "Siamese network features for image matching". In: *2016 23rd international conference on pattern recognition (ICPR)*. IEEE, pp. 378–383.

Miao, Yunqi, Jiankang Deng, Guiguang Ding, and Jungong Han (2024). "Confidence-guided centroids for unsupervised person re-identification". In: *IEEE Transactions on Information Forensics and Security* 19, pp. 6471–6483.

Milan, Anton, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler (2016). "MOT16: A Benchmark for Multi-Object Tracking". In: *ArXiv* abs/1603.00831. URL: https://api.semanticscholar.org/CorpusID:15430338.

Murphy, Evan, Marco Viola, and Vladimir Krylov (Sept. 2025). *A Stochastic Birth-and-Death Approach for Street Furniture Geolocation in Urban Environments*. DOI: 10.48550/arXiv.2509.10310.

Nassar, Ahmed Samy, Stefano D'Aronco, Sébastien Lefèvre, and Jan D. Wegner (2020). "GeoGraph: Graph-Based Multi-View Object Detection with Geometric Cues End-to-End". In: *European Conference on Computer Vision (ECCV)*.

Nassar, Ahmed Samy and Sébastien Lefèvre (2019). "Automated Mapping Of Accessibility Signs With Deep Learning From Ground-level Imagery and Open Data". In: *2019 Joint Urban Remote Sensing Event (JURSE)*, pp. 1–4. DOI: 10.1109/JURSE.2019.8808961.

Neuhold, Gerhard, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder (2017). "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes". In: *ICCV*.

OpenStreetMap Contributors (2024). *OpenStreetMap*. https://www.openstreetmap.org.

Pang, Ziqi, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, and Yu-Xiong Wang (2023). "Standing Between Past and Future: Spatio-Temporal Modeling for Multi-Camera 3D Multi-Object Tracking". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17928–17938. DOI: 10.1109/CVPR52729.2023.01719.

Pedersen, Kasper F. and Kristian Torp (2020). "Geolocating Traffic Signs using Crowd-Sourced Imagery". In: *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. SIGSPATIAL '20. Seattle, WA,

USA: Association for Computing Machinery, pp. 199–202. ISBN: 9781450380195. DOI: 10.1145/3397536.3422340.

Pham, Hieu, Zihang Dai, Qizhe Xie, and Quoc V. Le (2021). "Meta Pseudo Labels". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11552–11563. DOI: 10.1109/CVPR46437.2021.01139.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi (2016). "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

Reed, Scott, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich (2014). "Training deep neural networks on noisy labels with bootstrapping". In: *arXiv preprint arXiv:1412.6596*.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun (2017). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6, pp. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

Renton, Guillaume, Muhammet Balcilar, Pierre Héroux, Benoît Gaüzère, Paul Honeine, and Sébastien Adam (2021). "Symbols Detection and Classification using Graph Neural Networks". In: *Pattern Recognition Letters* 152, pp. 391–397. ISSN: 0167-8655.

Rodríguez-Puerta, F., C. Barrera, B. García, F. Pérez-Rodríguez, and A. M. García-Pedrero (2022). "Mapping Tree Canopy in Urban Environments Using Point Clouds from Airborne Laser Scanning and Street Level Imagery". In: *Sensors* 22.9, p. 3269.

Roth, Karsten, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen (2020). "Revisiting training strategies and generalization performance in deep metric learning". In: *International Conference on Machine Learning*. PMLR, pp. 8242–8252.

Rother, C., V. Kolmogorov, V. Lempitsky, and M. Szummer (2007). "Optimizing binary MRFs via extended roof duality". In: *2007 IEEE conference on computer vision and pattern recognition*. IEEE, pp. 1–8.

Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520.

Sattler, Torsten, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, and Tomas Pajdla (2018). "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schönberger, Johannes L. and Jan-Michael Frahm (2016). "Structure-from-Motion Revisited". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113. DOI: 10.1109/CVPR.2016.445.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

Shehzadi, Tahira, Khurram Azeem Hashmi, Shalini Sarode, Didier Stricker, and Muhammad Zeshan Afzal (Oct. 2025). "STEP-DETR: Advancing DETR-based Semi-Supervised Object Detection with Super Teacher and Pseudo-Label Guided Text Queries". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3069–3079.

Shehzadi, Tahira, Khurram Azeem Hashmi, Didier Stricker, and Muhammad Zeshan Afzal (June 2024). "Sparse Semi-DETR: Sparse Learnable Queries for Semi-Supervised Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5840–5850.

Shen, Fei, Yi Xie, Jianqing Zhu, Xiaobin Zhu, and Huanqiang Zeng (2023). "Git: Graph interactive transformer for vehicle re-identification". In: *IEEE Transactions on Image Processing* 32, pp. 1039–1051.

Shen, Yantao, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang (2017). "Learning Deep Neural Networks for Vehicle Re-ID with Visual-spatio-Temporal Path Proposals". In: *ICCV*, pp. 1918–1927. DOI: 10.1109/ICCV.2017.210.

Shi, Yujiao, Liu Liu, Xin Yu, and Hongdong Li (2019). "Spatial-Aware Feature Aggregation for Image based Cross-View Geo-Localization". In: *Advances in Neural Information Processing Systems*. Vol. 32.

Shorfuzzaman, Mohammad and M. Shamim Hossain (2021). "MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients". In: *Pattern Recognition* 113, p. 107700.

Shu, Jun, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng (2019). "Meta-weight-net: Learning an explicit mapping for sample weighting". In: *Advances in neural information processing systems* 32.

Simonovsky, Martin and Nikos Komodakis (2017). "Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29–38.

Smart Dublin (n.d.). https://data.smartdublin.ie/dataset/dcc-public-bin-locations. Accessed: Jan. 2026.

Snavely, Noah, Steven M. Seitz, and Richard Szeliski (July 2006). "Photo tourism: exploring photo collections in 3D". In: *ACM Trans. Graph.* 25.3, pp. 835–846. ISSN: 0730-0301. DOI: 10.1145/1141911.1141964. URL: https://doi.org/10.1145/1141911.1141964.

Snyder, John P. (1987). *Map Projections: A Working Manual*. US Geological Survey.

Song, Chunli, Yucan Zhou, Wenqiu Zhu, Jiang Zhou, Xiaoyan Gu, and Bo Li (2024). "Noise-Aware Person Re-identification via Local Uncertainty Estimation". In: *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Song, Hwanjun, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee (2022). "Learning from noisy labels with deep neural networks: A survey". In: *IEEE transactions on neural networks and learning systems* 34.11, pp. 8135–8153.

Su, Chi, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian (2017). "Pose-Driven Deep Convolutional Model for Person Re-identification". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3980–3989. DOI: 10.1109/ICCV.2017.427.

Sun, Yifan, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang (2018). "Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)". In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*. Berlin, Heidelberg: Springer-Verlag, pp. 501–518. ISBN: 978-3-030-01224-3. DOI: 10.1007/978-3-030-01225-0_30.

Szeliski, Richard (2010). *Computer Vision: Algorithms and Applications*. Springer.

Tang, Yi, Di Wu, Zhi Jin, Wenbin Zou, and Xia Li (2017). "Multi-modal metric learning for vehicle re-identification in traffic surveillance environment". In: *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2254–2258. DOI: 10.1109/ICIP.2017.8296683.

Tian, Yicong, Chen Chen, and Mubarak Shah (2017). "Cross-View Image Matching for Geo-Localization in Urban Environments". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1998–2006. DOI: 10.1109/CVPR.2017.216.

Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio (2018). "Graph Attention Networks". In: *International Conference on Learning Representations (ICLR)*.

Vo, Nam, Nathan Jacobs, and James Hays (2017). "Revisiting IM2GPS in the Deep Learning Era". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2640–2649. DOI: 10.1109/ICCV.2017.286.

Wang, Ao, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding (2024). "YOLOv10: real-time end-to-end object detection". In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. NIPS '24. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9798331314385.

Wang, Bowen, Yafei Wang, Wei Gong, Siheng Chen, Genjia Liu, Minhao Xiong, and Chin Long Ng (Oct. 2025). "V2XScenes: A Multiple Challenging Traffic Conditions Dataset for Large-Range Vehicle-Infrastructure Collaborative Perception". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28385–28395.

Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao (2023). "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7464–7475. DOI: 10.1109/CVPR52729.2023.00721.

Wang, Hongbo, Jiaying Hou, and Na Chen (2019a). "A Survey of Vehicle Re-Identification Based on Deep Learning". In: *IEEE Access* 7, pp. 172443–172469. DOI: 10.1109/ACCESS.2019.2956172.

Wang, Xun, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott (2019b). "Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5017–5025. DOI: 10.1109/CVPR.2019.00516.

Wang, Yisen, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey (2019c). "Symmetric cross entropy for robust learning with noisy labels". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 322–330.

Wang, Zhixue, Chaoyong Peng, Yu Zhang, Nan Wang, and Lin Luo (2021). "Fully convolutional siamese networks based change detection for optical aerial images with focal contrastive loss". In: *Neurocomputing* 457, pp. 155–167.

Wang, Zhongdao, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang (2017). "Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-identification". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 379–387. DOI: 10.1109/ICCV.2017.49.

Wegner, Jan Dirk, Steve Branson, David Hall, Nico Lang, Konrad Schindler, and Pietro Perona (2016). "Cataloging Public Objects Using Aerial and Street-Level Images – Urban Trees". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weng, Xinshuo, Jianren Wang, David Held, and Kris Kitani (2020). "3D Multi-Object Tracking: A Baseline and New Evaluation Metrics". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10359–10366. DOI: 10.1109/IROS45743.2020.9341164.

Weyand, Tobias, Ilya Kostrikov, and James Philbin (2016). "Planet-photo geolocation with convolutional neural networks". In: *European conference on computer vision*. Springer, pp. 37–55.

Wilson, Daniel, Thayer Alshaabi, Colin Van Oort, Xiaohan Zhang, Jonathan Nelson, and Safwan Wshah (2022). "Object Tracking and Geo-Localization from Street Images". In: *Remote Sensing* 14.11. ISSN: 2072-4292. DOI: 10.3390/rs14112575.

Wilson, Daniel, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah (Nov. 2023). "Image and Object Geo-Localization". In: *Int. J. Comput. Vision* 132.4, pp. 1350–1392. ISSN: 0920-5691.

Wojke, Nicolai, Alex Bewley, and Dietrich Paulus (2017). "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649. DOI: 10.1109/ICIP.2017.8296962.

Wolf, Paul R. and Charles D. Ghilani (2012). *Elementary Surveying: An Introduction to Geomatics*. Pearson.

Workman, Scott, Richard Souvenir, and Nathan Jacobs (2015). "Wide-Area Image Geolocalization with Aerial Reference Imagery". In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3961–3969. DOI: 10.1109/ICCV.2015.451.

Wu, Chih-Wei, Chih-Ting Liu, Cheng-En Chiang, Wei-Chih Tu, and Shao-Yi Chien (2018a). "Vehicle Re-identification with the Space-Time Prior". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 121–1217. DOI: 10.1109/CVPRW.2018.00024.

Wu, Fangyu, Shiyang Yan, Jeremy S. Smith, and Bailing Zhang (2018b). "Joint Semi-supervised Learning and Re-ranking for Vehicle Re-identification". In: *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 278–283. DOI: 10.1109/ICPR.2018.8545584.

Wu, Wenhao, Hau-San Wong, Si Wu, and Tianyou Zhang (June 2024). "Relational Matching for Weakly Semi-Supervised Oriented Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27800–27810.

Wu, Yichen, Jun Shu, Qi Xie, Qian Zhao, and Deyu Meng (2021). "Learning to purify noisy labels via meta soft label corrector". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12, pp. 10388–10396.

Xiang, Xuezhi, Zhushan Ma, Xiaoheng Li, Lei Zhang, and Xiantong Zhen (2025). "Vehicle re-identification with large separable kernel attention and hybrid channel attention". In: *Image and Vision Computing* 155, p. 105442. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2025.105442.

Xiao, Shunxin, Jiacheng Li, Jielong Lu, Sujia Huang, Bao Zeng, and Shiping Wang (Oct. 2024). "Graph neural networks for multi-view learning: a taxonomic review". In: *Artificial Intelligence Review* 57. DOI: 10.1007/s10462-024-10990-1.

Yang, Haiping, Yuanyuan Chen, Wei Wu, Shiliang Pu, Xiaoyang Wu, Qiming Wan, and Wen Dong (2023). "A Lightweight Siamese Neural Network for Building Change Detection Using Remote Sensing Images". In: *Remote Sensing* 15.4. ISSN: 2072-4292.

Ye, Mang, He Li, Bo Du, Jianbing Shen, Ling Shao, and Steven C. H. Hoi (2022a). "Collaborative Refining for Person Re-Identification With Label Noise". In: *IEEE Transactions on Image Processing* 31, pp. 379–391.

Ye, Mang, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi (June 2022b). " Deep Learning for Person Re-Identification: A Survey and Outlook". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 44.06, pp. 2872–2893. ISSN: 1939-3539.

Ye, Mang and Pong C. Yuen (2020). "PurifyNet: A Robust Person Re-Identification Model With Noisy Labels". In: *IEEE Transactions on Information Forensics and Security* 15, pp. 2655–2666.

Ye, Zi, Yogan Jaya Kumar, Goh Ong Sing, Fengyan Song, and Junsong Wang (2022c). "A Comprehensive Survey of Graph Neural Networks for Knowledge Graphs". In: *IEEE Access* 10, pp. 75729–75741. DOI: 10.1109/ACCESS.2022.3191784.

Yi, Hong, Matthew Satusky, David Borland, Meghna Chakraborty, Mike Vann, Raghavan Srinivasan, and Chris Bizon (Oct. 2025). "Automated Roadside Object Detection and Geolocation by Leveraging Data Fusion of Airborne LiDAR and Videolog Images". In: *Data Science for Transportation* 7. DOI: 10.1007/s42421-025-00132-6.

Yi, Li, Sheng Liu, Qi She, A Ian McLeod, and Boyu Wang (2022). "On learning contrastive representations for learning with noisy labels". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16682–16691.

Yin, Junbo, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang (June 2024). "IS-Fusion: Instance-Scene Collaborative Fu-

sion for Multimodal 3D Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14905–14915.

Yu, Tianyuan, Da Li, Yongxin Yang, Timothy Hospedales, and Tao Xiang (2019). "Robust Person Re-Identification by Modelling Feature Uncertainty". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 552–561.

Yu, Zhipeng, Qianqian Xu, Yangbangyan Jiang, Yingfei Sun, and Qingming Huang (2024). "Enhancing Samp le Utilization in Noise-Robust Deep Metric Learning With Subgroup-Based Positive-Pair Selection". In: *IEEE Transactions on Image Processing* 33, pp. 6083–6097.

Yuan, Ye, Wuyang Chen, Yang Yang, and Zhangyang Wang (2020). "In Defense of the Triplet Loss Again: Learning Robust Person Re-Identification with Fast Approximated Triplet Loss and Label Distillation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1454–1463. DOI: 10.1109/CVPRW50498.2020.00185.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (Feb. 2021). "Understanding deep learning (still) requires rethinking generalization". In: *Commun. ACM* 64.3, pp. 107–115.

Zhang, Shiwei, Zhengzheng Wang, and Wei Ke (2025). "One point is all you need for weakly supervised object detection". In: *Pattern Recognition* 159, p. 111087. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2024.111087.

Zhang, Weixing, Chandi Witharana, Weidong Li, Chuanrong Zhang, Xiaojiang Li, and Jason Parent (2018). "Using Deep Learning to Identify Utility Poles with Crossarms and Estimate Their Locations from Google Street View Images". In: *Sensors* 18.8. ISSN: 1424-8220. DOI: 10.3390/s18082484.

Zhang, Yiheng, Dong Liu, and Zheng-Jun Zha (2017). "Improving triplet-wise training of convolutional neural network for vehicle re-identification". In: *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1386–1391. DOI: 10.1109/ICME.2017.8019491.

Zhang, Zhilu and Mert R. Sabuncu (2018). "Generalized cross entropy loss for training deep neural networks with noisy labels". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 8792–8802.

Zhang, Zizhao and Tomas Pfister (2021). "Learning Fast Sample Re-weighting Without Reward Data". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 705–714. DOI: 10.1109/ICCV48922.2021.00076.

Zheng, Liang, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian (2015). "Scalable Person Re-identification: A Benchmark". In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124.

Zhong, Zhun, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang (June 2019). " Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification ". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 598–607. DOI: 10.1109/CVPR.2019.00069.

Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba (2016). "Learning Deep Features for Discriminative Localization". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929. DOI: 10.1109/CVPR.2016.319.

Zhou, Yi, Li Liu, and Ling Shao (2018). "Vehicle Re-Identification by Deep Hidden Multi-View Inference". In: *IEEE Transactions on Image Processing* 27.7, pp. 3275–3287. DOI: 10.1109/TIP.2018.2819820.

Zhouy, Yi and Ling Shao (2018). "Viewpoint-Aware Attentive Multi-view Inference for Vehicle Re-identification". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6489–6498. DOI: 10.1109/CVPR.2018.00679.

Zhu, Jianqing, Huanqiang Zeng, Yongzhao Du, Zhen Lei, Lixin Zheng, and Canhui Cai (2018). "Joint Feature and Similarity Deep Learning for Vehicle Re-identification". In: *IEEE Access* 6, pp. 43724–43731. DOI: 10.1109/ACCESS.2018.2862382.

Zhu, Sijie, Taojiannan Yang, and Chen Chen (2020). "VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5316–5325.

Zimmer, Walter, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C. Knoll (June 2024). "TUMTraf V2X Cooperative Perception Dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22668–22677.