Advanced Image Data Augmentation Strategies to enhance Robustness, Generalization and Bias Mitigation

Teerath Kumar, MS

Supervised by Prof. Alessandra Mileo and

Dr. Malika Bendechache, School of Computing, University of Galway, Galway, Ireland

A thesis presented for the degree of Doctor of Philosophy

(Ph.D.)

School of Computing,

Dublin City University

December 2025

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 is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, 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.

Teerath Kumar

Teerath Kumar

ID no. 21267958

31 December 2025

GenAI Declaration

I declare that no Generative AI tools were used in the preparation of this thesis. However, grammarly was utilized for grammatical corrections. Additionally, Google Gemini was employed to assist with LaTeX table and equation formatting. The final content and intellectual conclusions remain entirely my own.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Alessandra Mileo and Dr. Malika Bendechache, for their invaluable guidance, expertise, and continuous support throughout the course of these years.

I would also like to thank my wife, Amrata Meghwar, for her constant love and support during the challenging moments. Your unwavering presence has been a source of strength.

My heartfelt thanks go to my father, Hanso Mal, and my mother, Baya, for their unconditional love and encouragement, which have been a pillar of support throughout my journey.

I am also grateful to my brother, Rawal Rai, and my sisters, Gudi Bai and Shevta Kumari, for their constant motivation and belief in me. Your encouragement has been essential to my progress.

Finally, I would like to express my appreciation to the entire Majnani family for their support and understanding during this time. Thank you all for being part of this important journey.

Contents

List of Figures

List of Tables

Advanced Image Data Augmentation Strategies to enhance Robustness, Generalization and Bias Mitigation

Teerath Kumar

Abstract

Data augmentation plays a crucial role in improving deep learning models, yet conventional approaches often result in feature loss or biased learning. This thesis introduces novel enhancement techniques that address these challenges in multiple domains. First, Random Slices Mixing Data Augmentation (RSMDA) enhances feature diversity by strategically combining image slices while leveraging label smoothing, improving model generalisation. Second, RandSaliencyAug (RSA) balances feature loss and contextual information retention by selectively occluding salient regions using six new strategies, outperforming existing occlusion-based methods. Third, KeepOriginalAugment integrates salient regions into non-salient areas, striking a balance between data diversity and information preservation through optimised placement strategies. We evaluated these methods in FashionMNIST, STL10, CIFAR10, CIFAR100, TinyImageNet, ImageNet, and VOC 2007. In addition, we provide depth analysis and comparison including class activation maps (explainability), robustness, and time complexity.

Furthermore, FaceSaliencyAug and FaceKeepOriginalAugment mitigate geographical, gender, and stereotypical biases in computer vision models, improving fairness and diversity. We also explore and propose two novel data augmentation techniques from an image-mixing perspective: Noise Addition (NA) and Partial Mix (PM). These techniques were evaluated on FFHQ, WIKI, IMDB, LFW, and UTK Faces datasets, as well as diverse datasets, demonstrating significant improvements in di-

versity. We evaluate their impact on mitigating gender bias across CEO, Engineer, Nurse, and School Teacher datasets, using the Image-Image Association Score (IIAS) in convolutional neural networks (CNNs) and vision transformers (ViTs). In addition, we introduce a Saliency-Based Diversity and Fairness Metric to quantify bias reduction across datasets. Our approaches are rigorously tested on CNNs and ViTs with extensive evaluations on facial recognition datasets, proving their effectiveness in enhancing both model performance and fairness.

Chapter 1

Introduction

1.1 Background

Deep learning models have gained popularity and achieved tremendous progress in computer vision (CV) tasks such as image classification [1, 2, 3], object detection [4, 5] and image segmentation [6, 7, 8, 9]. These advances have been fostered by effective deep neural network architectures, powerful computational resources, and extensive availability of data [10]. Deep learning has significantly transformed computer vision, with Convolutional Neural Networks (CNNs) excelling in feature extraction through hierarchical convolution operations. Early layers detect basic patterns like edges, while deeper layers capture complex structures. Their success has fueled research in vision tasks. More recently, Vision Transformers (ViTs) [11] have gained attention for leveraging self-attention in image analysis.

Deep neural networks require large amounts of data and are prone to overfitting [12], where they perform well on training data but struggle with unseen samples. This problem worsens with limited data, often due to privacy concerns or the high cost of manual labeling [10]. Even with large open datasets like ImageNet [13], overfitting persists, as models tend to focus on key regions while neglecting less prominent but generalisable features [14]. To mitigate overfitting, two broad strategies are commonly used: model regularisation and data augmentation. Model regularisation [15] controls the complexity of neural networks to improve generalisation,

with techniques like batch normalisation [16] and dropout [17]. On the other hand, data augmentation [18, 3, 19, 10] enhances model robustness by generating diverse training samples from existing data. Traditional augmentation techniques, such as flipping [18, 10], cropping [18, 10], and resizing [18], have been widely used but often lack sufficient diversity for deep CNN architectures [19, 20, 21, 22, 23]. This limitation has driven interest in advanced augmentation methods, including random erasing (RE) [18], hide-and-seek (HS) [24], GridMask [25], CutOut [26], MixUp [27], CutMix [14], and RICAP [19], among others [10].

Data augmentation can be broadly classified into five main categories: spatial augmentations [2], color distortion [2], image mixing [28], information-erasing [28], and saliency-based augmentation [29].

Spatial augmentations involve geometric transformations such as rotation, flipping, and cropping, which help improve model invariance. Color distortion techniques manipulate image color properties, including random brightness adjustments and jittering, to enhance robustness. Image mixing methods, such as MixUp [27] and CutMix [14], create augmented samples by blending multiple images.

Information-erasing augmentation removes certain regions of an image, compelling the model to learn alternative features. For instance, erasing a crucial part (e.g., a cat's face) forces the model to focus on secondary features like legs or the tail, while removing less critical features helps reinforce the importance of key structures. Techniques in this category include CutOut [26], Random Erasing (RE) [18], Hide-and-Seek (HaS) [24], and GridMask (GM) [25]. CutOut randomly replaces image patches with a fixed value, whereas RE masks out regions with randomly determined aspect ratios and sizes. HaS divides an image into a grid and selectively removes squares, while GM applies a uniform mask to disrupt patterns within the image.

Saliency-based augmentation methods leverage saliency detection to transform images while preserving or modifying their most important regions. These techniques enhance dataset diversity while maintaining critical image structures. Exam-

ples include SaliencyMix [29], SalfMix [30], and KeepAugment [31]. SaliencyMix [29] blends salient regions from two images to generate new samples, reinforcing semantically meaningful features. SalfMix [30] duplicates and integrates salient features with non-salient regions, potentially enhancing model focus but risking overfitting. KeepAugment [31] selectively applies transformations such as cropping and resizing to non-salient areas while preserving critical features, though it may introduce domain shifts that affect contextual learning.

Although these methods enhance data diversity, they may also introduce noise and impact feature fidelity. For instance, SalfMix's emphasis on salient feature duplication could lead to overfitting, reducing generalisation capability. KeepAugment's selective augmentation, while maintaining context, may introduce domain shifts that hinder the effective exchange of crucial contextual information, ultimately affecting model comprehension.

These issues become more important when assessing the relationship between data augmentation and bias, we define bias as a systematic and unfair difference in model performance or predictions across groups, caused by imbalanced or non-representative training data. We quantify such bias as disparities in accuracy and associations across groups (e.g. using the image-to-image association score (IIAS) or image similarity score (ISS)). Image datasets are known to exhibit biases such as such as geography [32, 33], race [34, 35], and gender [34, 36]. Recent work has shown that certain data-augmentation policies, while reducing overall error, can actually worsen model bias by decreasing error for majority subgroups and increasing error for minorities [37]. Therefore, if data augmentation is not designed carefully, it may amplify existing biases in the effective training distribution rather than mitigate them.

Computer vision models often exhibit social biases, including gender [34, 36], geographical [32, 33], and racial biases [34, 35]. For instance, facial recognition systems perform less accurately for women and individuals with darker skin tones [34]. Similarly, pulse oximeters, which measure blood oxygen levels, tend to overestimate

oxygen saturation in non-white patients, leading to under-detection of hypoxia, particularly in Black patients who are three times more likely to be affected [38].

A primary source of bias is the training data, often compiled from online sources like Google and Flickr (e.g., FFHQ [39]), which may reinforce existing biases. Audits of facial datasets primarily focus on race (via skin tone) and gender [32, 34, 35], highlighting the importance of facial region adjustments for debiasing. To address these biases, various mitigation strategies have been proposed [40, 41, 42].

Another important issue is the adversarial attack: a small, often unrecognisable perturbation to the input. Such perturbations mislead the network and degrade its performance [43]. A simple way to prevent an attack is to generate an unseen input sample [44]. For adversarial attacks, we assume that the attacker has complete information about the model, i.e., a white box attack. Adversarial attacks are very important because very small, barely recognisable perturbations can reliably deceive deep models, directly undermining the reliability of computer-vision systems. This vulnerability means robustness must be explicitly evaluated alongside accuracy, not assumed from clean-data performance. In this thesis, robustness is treated as a first-order concern: models are stress-tested with standard attacks (e.g., FGSM and FGM) across datasets and architectures, using controlled perturbation parameter (ϵ), to quantify how quickly performance degrades under attack.

To summarise, modern vision models have advanced rapidly but remain prone to overfitting and shortcut learning, which motivates stronger data augmentation beyond basic flips and crops. This work reviews five augmentation families—spatial, colour distortion, image mixing, information-erasing, and saliency-based—and notes their trade-offs: while they boost diversity, some erode feature fidelity or induce domain shift. Therefore we propose different data augmentation techniques discussed in next chapters. We validate our appraoched not only from performance metrics (accuracy, error rate etc) but also time comparison, class activation maps. Furthermore, we connect these limitations to fairness concerns: poorly designed augmentation can entrench demographic biases, whereas saliency-guided approaches can help

measure and reduce disparities (e.g., with IIAS). Finally, because small, imperceptible perturbations can reliably fool deep models, robustness must be evaluated alongside accuracy; this thesis therefore treats adversarial resilience as a first-order objective and will test standard attacks (FGSM/FGM) under controlled budgets.

1.2 Problem statement

Data augmentation plays a crucial role in improving the robustness and generalisation of deep learning models. However, advanced data augmentation, especially information-erasing augmentations such as Random Erasing (RE), CutOut, and GridMask, can remove key features or contextual information from images. Complete object removal creates noisy data - which deteriorate the model performance, and contextual information removal generates object-focused images - which results in an overfitting problem. Both cases are important; either of them compromises the model's ability to make reliable inferences, especially in complex and sensitive applications. This limitation is particularly problematic for tasks that rely on subtle details in images to distinguish between classes.

Simultaneously, computer vision models face a challenge due to the presence of the different biases in training datasets. These biases, like gender, geographical, and racial biases, are based on unbalanced or non-representative data. For example, facial recognition systems show more accuracy for women with white skin than for non-white women.

Given these challenges, there is a clear need for advanced data augmentation methods that can enhance data diversity while preserving essential features and improving model generalisation. Furthermore, there is a critical need to address and mitigate bias in computer vision models, ensuring that they perform equitably across diverse demographic groups.

This thesis proposes innovative data augmentation techniques such as Random Slices Mixing Data Augmentation (RSMDA, where a slice is a contiguous strip of pixels aligned with the image axes) , RandSaliencyAug, and KeepOriginalAugment,

designed to tackle these issues. The proposed augmentations aim to increase data diversity without losing important image features, thus improving model performance and robustness. We also explore these techniques to mitigate biases in computer vision models, guaranteeing more equitable performance. This thesis aims to improve the fairness and accuracy of computer vision models by using these new augmentation strategies. This will make them more useful for real-world tasks like facial recognition by making them more reliable and unbiased.

1.3 Hypotheses and Research Questions

1.3.1 Hypothesis 1 (H1):

We can combine slices from different images to design an image data augmentation approach that reduces feature loss and increases feature diversity, leading to improved model generalisation.

Research Question 1 (RQ 1): What is the most effective way to combine slices from different images for data augmentation so that we can have the best results in reducing feature loss and enhancing feature diversity for better model generalisation?

1.3.2 Hypothesis 2 (H2):

A data augmentation strategy that incorporates both saliency detection in image-erasing technique in a clever way, can help balance feature loss and contextual information loss, improving the quality of the augmented data.

Research Question 2 (RQ 2): How can we design a strategy that combines saliency and image-erasing in a way that maximises the trade off between feature loss and contextual information loss in data augmentation? Specifically, when tackling this question, we break down the problem into two as follows:

- **RQ 2.1:** How can we handle occlusion while preserving important features?

- **RQ 2.2:** How can we mitigate overfitting while retaining contextual information?

1.3.3 Hypothesis 3 (H3):

Instead of focusing on salient regions only, a data augmentation strategy that considers both salient and non-salient features can better increase data diversity at the same time preserving contextual information

Research Question 3 (RQ 3): How can we balance the use of salient and non-salient regions to improve model generalisation? In order to investigate this research question, we break it down in two parts: the first is related to where the augmented region needs to be placed, while the second is concerned with what type of region is augmented. As a result, this question is split in two parts as follows:

- **RQ 3.1:** What role do different placement strategies (e.g., minimum, maximum, random) play in optimising the balance between data diversity and information preservation?

- **RQ 3.2:** How does the choice of augmented region (salient, non-salient, or both) affect model performance and generalisation?

1.3.4 Hypothesis 4 (H4):

Data augmentation approaches that carefully consider salient and nonsalient regions and where they are placed can not only improve model fairness, but also reduce geographical, gender, and stereotypical biases in computer vision models.

Research Question 4 (RQ 4): How can we mitigate geographic, gender, and stereotypical biases in computer vision models with data augmentation? In tackling this question, we separate gender from other demographics, and also focus on assessment metrics, thus generating three subquestions as follows:

- **RQ 4.1:** How can data augmentation impact model fairness and reduce gender bias?

- **RQ 4.2:** How can data augmentation improve diversity in data sets while maintaining fairness among different demographic groups?
- **RQ 4.3:** How can we design a new metric that is able to capture and measure fairness and diversity regardless of whether the data is balanced or imbalanced?

1.4 Thesis Structure

1. **Introduction (Chapter 1):** The current chapter provides an overview of the context, motivation, and background knowledge for the work. It introduces the core concepts related to image data augmentation, debiasing techniques, and the role of saliency in improving model robustness and fairness. The chapter also outlines the hypothesis and research questions tackled in the thesis and explains how Chapters 3–6 answer them in sequence, with each chapter building on the findings of the previous one; this roadmap is made explicit in the handoff lines below.
2. **Literature Review (Chapter 2):** This chapter reviews state-of-the-art research in the areas of data augmentation, saliency-based augmentation, and debiasing techniques in computer vision. The literature is categorised into the following topics:
    - (i) Data augmentation techniques in computer vision.
    - (ii) Saliency-based augmentation methods.
    - (iii) Data augmentation for fairness.

    The review identifies concrete gaps (feature loss in erasing, lack of structure in mixing, and fairness risks) that motivate H1–H4 and directly lead to the designs in Chapters 3–6.
3. **Random Slice Mixing Data Augmentation (Chapter 3):** This chapter addresses Hypothesis 1 (H1). We propose Random Slice Mixing Data Augmentation (RSMDA) and investigate sliced image mixing as a strategy to minimise

feature loss and enhance diversity while improving model generalisation. The observation that preserving discriminative detail is critical motivates Chapter 4, which uses saliency to decide *what* to alter and *where*.

4. **Saliency-Based Data Augmentation Techniques (Chapter 4):** This chapter addresses Hypothesis 2 (H2) and delves into a saliency-based data augmentation method, which utilises saliency detection to guide augmentation. It builds on Chapter 3 by targeting modifications to salient regions to reduce shortcut learning while retaining key information. These results expose an object–context trade-off, setting up Chapter 5 to combine salient and non-salient regions in a balanced way.

5. **Combining salient and non-salient regions in Data Augmentation (Chapter 5):** This chapter tackles Hypothesis 3 (H3) and explores the intelligent integration of salient and non-salient regions in data augmentation, guided by saliency detection. It extends Chapter 4 to preserve context while scaling diversity and mitigating domain shift. The resulting toolkit and hyperparameter choices are transferred in Chapter 6 to fairness-sensitive face datasets and evaluation.

6. **Improving Fairness with Data Augmentation (Chapter 6):** This chapter addresses Hypothesis 4 (H4). We explore the challenges of bias or fairness in computer vision and propose methods for mitigating such biases through augmentation strategies, reusing and adapting the approaches from Chapter 4 and 5. The chapter also presents fairness-with-diversity metrics and links fairness outcomes back to robustness and generalisation. These findings complete the RQ chain and feed directly into the overall conclusions.

7. **Conclusions and Future Work (Chapter 7):** This final chapter summarises the research conducted in this thesis by revisiting the hypotheses and research questions. It discusses the contributions of the work, highlights limitations in the proposed methods, and presents future directions for further

research in image augmentation and fairness, reflecting how the stepwise progression (Chapters 3–6) supports the overall aims.

Chapter 2

Literature review

In this chapter, we will explore state-of-the-art image data augmentation techniques, with a focus on image mixing and saliency-based methods. The discussion will encompass their impact not only on model performance but also on mitigating gender, racial, and geographical biases in computer vision.

As previously discussed in the introductory chapter, regularisation is the common approach to prevent the model from overfitting. We have mentioned that there are two main categories of regularisation: (i) dropout as a regularisation technique and (ii) data augmentation an explicit form of regularisation [2, 1, 3, 18]. Despite this PhD is focused on data augmentation, we believe it is essential to briefly discuss dropout state of the art (Section 2.1), before we move onto Data Augmentation (Section 2.2), Bias in Computer Vision (Section 2.3) and our contribution (Section 2.5).

2.1 Dropout

Extensive research has been conducted on dropout [45, 17, 46, 47, 48], a regularisation technique introduced by Hinton et al. [45, 17]. It randomly sets neurons to zero during training, effectively averaging multiple sub-networks to enhance generalisation and robustness against adversarial attacks. However, dropout is more effective in fully connected layers than in convolutional layers (CLs) [49], primarily because

(i) CLs have fewer parameters and require less regularisation, and (ii) dropping pixels does not significantly affect image-level features due to redundant neighboring information.

Several dropout variants have been introduced to improve its effectiveness. Adaptive dropout [46] adjusts neuron drop probabilities using a binary belief network. DropConnect [47] randomly zeros subsets of weights instead of activations. SpatialDropout [49] discards entire feature maps to mitigate redundant pixel information. Stochastic pooling [48] selects activations from a multinomial distribution and can be combined with data augmentation or dropout for enhanced regularisation.

Since dropout applies at the neuron level and does not modify the data but modifies the network architecture, which may not be sufficient when the model is provided with limited or non-variant data. On the other hand, data augmentation 2.2, which applies on data level directly to improve model generalisation by diversifying input data, helps the model to regularise the better when data is limited or non-variant. This difference shows that dropout has its limitations as regularisation, while data augmentation offers more advantages.

2.2 Image Data Augmentation

Data augmentation is the regularisation technique we are focusing on, in this thesis. Technically, data augmentation enlarges the number of training samples using existing training data samples; the goal is that having more (good quality) training sample will increase the performance of deep learning models. Most data augmentation techniques create different flavors of existing samples during training, thereby providing many different views of such samples to increase the diversity of data. Recently, there has been a lot of research work on data augmentation [18, 24, 25, 26, 19, 50, 10]. In order to better help the reader navigate through the important approaches in this area, we classify data augmentation approaches into different subcategories, as shown in Figure 2.1 and discussed in depth in our survey [28].

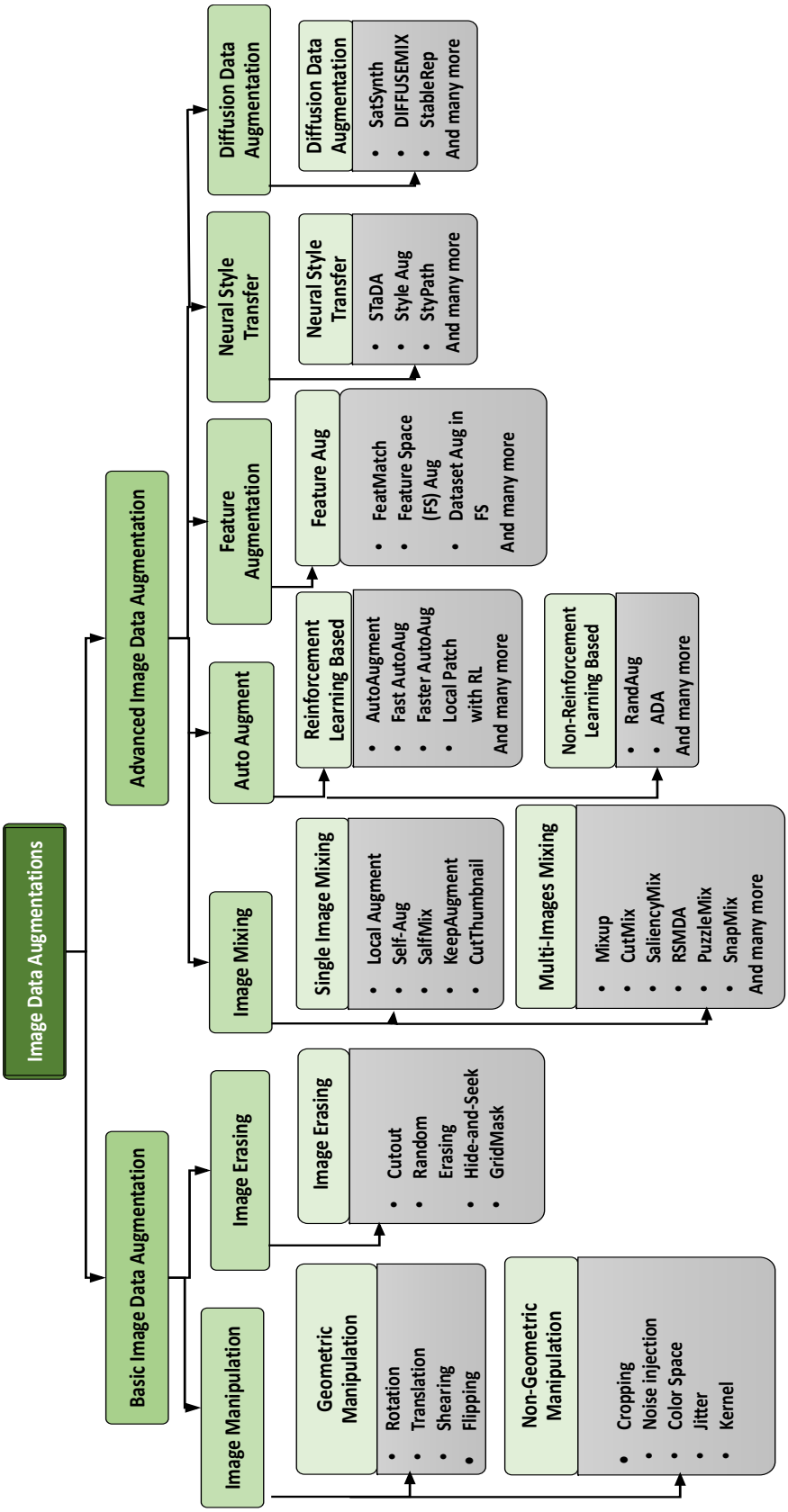


Figure 2.1: Image data augmentation taxonomy. Note: Due to space constraints, not all image data augmentation techniques are included in this taxonomy. The details of each technique are provided in the survey [28]

2.2.1 Basic Augmentation Techniques

Basic image augmentation can be categorised into image manipulation and information erasing. Image manipulation includes geometric and non-geometric transformations, while information erasing selectively removes image parts.

Geometric Data Augmentation. Geometric augmentation alters an image's position, orientation, or aspect ratio using techniques such as rotation, translation, shearing, and flipping. Rotation modifies orientation but requires careful selection to prevent unintended distortions. Translation shifts an image, introducing diversity, though excessive shifts can degrade its structure. Shearing alters shape asymmetrically, and flipping enhances diversity but must align with dataset characteristics to avoid introducing noise.

Non-geometric Data Augmentation. Non-geometric augmentation modifies pixel values without altering the structure. Common techniques include cropping, resizing, noise injection, color space manipulation, jitter, and kernel filtering. Cropping extracts a portion of an image and resizes it, while noise injection improves robustness against adversarial attacks. Color space manipulation adjusts image channels to prevent biases, though extreme alterations may mislead models. Jitter applies random brightness, contrast, saturation, and hue changes, requiring careful tuning to preserve key features. Kernel filtering, such as Gaussian blur or edge detection, balances detail enhancement and noise reduction.

2.2.2 Information Erasing Augmentation

Information erasing augmentation occludes parts of an image, compelling models to learn features from unmasked regions, is focus of this thesis as we address RQ1 and RQ2. Techniques in this category include random erasing [18], cutout [26], grid mask [25], and hide-and-seek [24]. Cutout randomly masks image regions, while random erasing dynamically determines occlusion areas based on aspect ratio and

size. Grid mask applies a uniform masking pattern, whereas hide-and-seek divides an image into squares and randomly removes portions.

While these methods encourage robust feature learning, they may introduce challenges. Excessive occlusion can produce noisy examples, negatively impacting performance: when large regions containing both object and background are removed, the remaining evidence may be insufficient for reliable classification, so the model is pushed to memorise idiosyncratic residual patterns or artefacts, a classical form of overfitting [25]. Moreover, if contextual cues are systematically erased during training but present at test time, this induces a train–test distribution shift; conversely, when context is kept but spuriously correlated with the label, models can overfit to it, as shown on the Waterbirds dataset [51]. This trade-off highlights the need for augmentation techniques that break spurious shortcuts while preserving sufficient object and contextual information to maintain feature fidelity.

2.2.3 Advanced Image Data Augmentation

Innovative methods for image data augmentation, including reinforcement learning, feature-based, and style-based techniques, have emerged. These advancements are classified into key categories, offering a framework to explore the field and identify areas for further research.

Image Mixing: This technique involves blending one or more images, including the same image, resulting in improved deep neural network model accuracy. We categorise image mixing data augmentation into two sub-categories: single image mixing - requires single to mix with itself and multi-images mixing - requires more than one image. Within single-image mixing, this thesis also focuses on information-preserving techniques that enhance diversity while retaining essential image features, as the thesis addresses RQ3. Notable approaches include RandAugment, which simplifies the process by randomly selecting transformations, reducing computational cost while maintaining effectiveness [52].

Other related methods, such as SalfMix [30] and KeepAugment [31], leverage

saliency detection to guide augmentation. SalfMix relocates salient regions to non-salient areas, forcing models to prioritise key features, though it risks overfitting due to redundant salient features (Figure 5.1b). KeepAugment preserves salient regions while applying RandAugment only to non-salient areas, promoting feature fidelity. However, this selective augmentation may introduce a domain shift, disrupting contextual relationships between salient and non-salient regions (Figure 5.1c).

AutoAugment: AutoAugment aims to identify optimal data augmentation policies through a discrete search process. It includes a search algorithm and space, with techniques categorised into reinforcement learning and non-reinforcement learning sub-categories, more detailed in survey [28].

Feature augmentation: Feature augmentation is another category of data augmentation, where images are transformed into embedding or representation then data augmentation is performed on the embedding of the image. This technique enhances the feature space used for model training by modifying these embeddings, rather than directly altering the raw image data. The details about sub-categories is provided in the survey [28].

Neural Style Transfer: It is another category of data augmentation, which can transfer the artist style of one image to another without changing semantics at a high level. It brings more variety to the training set. The main objective of this neural style transfer is to generate a third image from two images, where one image provides texture content and another provides high-level semantic content. Some of the SOTA augmentation methods for the sub-category are discussed in the survey [28].
Diffusion Data Augmentation This technique generates new data samples by simulating the diffusion process, adding random perturbations followed by smoothing. It creates realistic data variations, enhancing the training set's diversity and improving model robustness and generalisation. The subcategories are discussed in the survey [28].

2.3 Understanding Bias in Computer Vision

Social biases related to ethnicity [34], gender [53], geographical region and culture [54, 32, 33] is now a well-documented problem in computer vision. These biases mainly originate in training data primarily sourced from the Internet and is propagated and amplified throughout the machine learning pipeline [53, 54]. Such issues can cause a multitude of problems when models are deployed in real-world applications, including variances in accuracy in facial recognition systems depending on gender and race [34] and the generation of stereotypical images related to gender [55]. Such biases can cause harm, foster discrimination, and stymie progress towards a more equitable and just society [54, 53].

2.3.1 Bias and data augmentation for bias mitigation

Bias in computer vision models can be understood as a systematic tendency to perform differently across subgroups or to rely on spurious correlations that do not reflect the true underlying task. In the context of face analysis and image classification, this often manifests as higher error rates for certain demographic groups (e.g., darker-skinned or female faces) or strong dependence on backgrounds, co-occurring objects, or geographic cues rather than on the target object itself [53, 32]. Such behaviour is not only a robustness problem under distribution shift but also a fairness issue, as it can propagate and amplify existing social and cultural inequalities when deployed at scale.

Bias mitigation strategies in machine learning are commonly grouped into three broad families: pre-processing, in-processing, and post-processing. Pre-processing methods modify the data before training, for example by rebalancing class or group frequencies, filtering or relabelling problematic samples, or learning fair representations [56, 57]. In-processing methods incorporate debiasing directly into the learning algorithm, such as through adversarial debiasing, group-aware regularisation, or distributionally robust optimisation that upweights worst-case groups [40, 58]. Post-processing methods adjust model outputs after training, for example via group-

specific thresholds or calibration to equalise certain error rates across groups [59]. Each category has its own trade-offs: pre-processing is model-agnostic but relies heavily on data quality and group labels, in-processing can be effective but often increases training complexity, and post-processing is limited because it cannot change the underlying representation learned by the model.

Within this landscape, data augmentation sits naturally as a pre-processing strategy. Traditional augmentation operations (random flips, crops, colour jittering, noise) are primarily designed as generic regularisers to improve robustness and reduce overfitting, but they do not explicitly target bias. Nonetheless, they already hint at a connection: by exposing the model to more diverse views of the same underlying content, they can reduce sensitivity to incidental factors such as viewpoint or illumination. To address bias more directly, augmentation must be used in a targeted way: either to increase variability for minority or under-represented subgroups, or to disrupt specific spurious correlations that the model might otherwise exploit.

Numerous strategies have been proposed to mitigate bias in computer vision models. These include the expansion of dataset diversity, as outlined in Kärkkäinen et al.'s work [53], as well as the deployment of adversarial debiasing techniques [40]. In the context of image data augmentation for debiasing, previous research is relatively scarce [37, 60]. The aforementioned studies have primarily employed data augmentation to address different facets of bias. Li et al. [60] have focused on leveraging data augmentation for enhancing cross-bias generalisation. Smith et al. [37] have also explored data augmentation within an evolutionary framework to combat gender and age bias. Zhang et al. [40] have explored data augmentation in two key ways: first, to balance class representation, and second, to generate adversarial examples aimed at reducing bias in image classification, thus enhancing model fairness. BiasSwap [41] debiases neural networks by using unsupervised sorting and style transfer, effectively altering the biases present in the training data. While this method has shown promise, it lacks exploration of how partial erasure of salient

regions can further enhance debiasing. Lee et al. [42] have proposed a technique that improves debiasing through the synthesis of diverse bias-conflicting samples. This approach is valuable as it introduces variability in training data, but it may not adequately address the underlying biases in the salient features of images.

Moreover, a critical examination of these methods reveals a gap in addressing geographical and stereotypical biases. Understanding these biases is essential in the context of localised and personalised search results. Research indicates that implicit AI biases contribute to geographical bias, where the localisation and personalisation of search engines can perpetuate stereotypes, such as the prevalent stereotype of the "white, middle-aged female nurse" in Western societies [32]. Existing augmentation-based methods rarely consider such geographical and cultural dimensions explicitly, and most either treat augmentation as a generic robustness tool or rely on single-image perturbations that remove potentially biased cues without replacing them with unbiased evidence.

This motivates the direction taken in this thesis: to treat data augmentation as a bias-aware pre-processing mechanism that is tightly integrated with saliency information and designed not only to increase diversity, but also to replace or redistribute biased visual evidence while preserving label-relevant content. The subsequent sections detail how the proposed methods build on and extend these ideas.

2.4 Motivation for Our Contributions

The existing literature reveals three persistent limitations across the state of the art. First, most single-image erasing and mixing methods (e.g. Cutout, Random Erasing, GridMask, Hide-and-Seek, SalfMix) operate on each image in isolation: once regions are removed or heavily occluded, the lost information is not replaced, yielding low-information samples that increase estimator variance and encourage memorisation of whatever fragments remain. This motivates the proposed Random Slice Mixing Data Augmentation (RSMDA), which reconstructs richer training examples by mixing randomly sliced regions from multiple images while preserving

global semantics. Second, more advanced saliency-aware augmentations (e.g. KeepAugment, SalfMix, RandAugment) partially address feature loss but still struggle to jointly maximise data diversity and information preservation: they tend to freeze or repeatedly duplicate salient regions, induce domain shift between salient and non-salient areas, and can distort contextual cues. This motivates our RandSaliencyAug and KeepOriginalAugment strategies, which combine saliency detection with targeted erasing and controlled mixing to retain both object and contextual evidence. Third, work on bias-aware augmentation is comparatively sparse and fragmented, with limited exploration of saliency-guided erasing for debiasing, little attention to geographical and stereotypical biases, and a lack of saliency-based quantitative metrics for dataset diversity and fairness. The thesis contributions directly target these gaps by (i) recovering information lost in single-image erasing through RSMDA, (ii) refining saliency-driven augmentation via RandSaliencyAug and KeepOriginalAugment to balance diversity and feature fidelity, and (iii) extending these mechanisms to explicitly assess and mitigate bias in computer vision models through bias-aware facial variants and saliency-based fairness metrics.

2.5 Our Contribution Beyond the State-of-the-art

Our contributions aim to address specific challenges and relate to our Research Questions as discussed in the remainder of this chapter. Each contribution corresponds to one or more hypotheses and chapters, and together they form a progression from mixed slicing, to saliency-guided erasing, to balancing diversity and information preservation, and finally to fairness.

2.5.1 Contribution 1 - Mixed slicing

Single-image deletion techniques, as discussed in Section 2.2.2, may lose features, consequently deteriorating the performance of DL models, and multi-image mixing techniques have explored augmentations from different perspectives, but none have

considered data augmentation based on slice mixing. To the best of our knowledge, we are the first to explore random slice mixing to check the effect of the proposed data augmentation technique using different strategies, namely the horizontal (row-wise) slices mixing strategy, the vertical (column-wise) slices mixing strategy, and a mixture of both, as shown in Figure 3.2. The research question we address is: *can the proposed data augmentation technique preserve the feature information lost in single-image data augmentations?* This challenge has been suggested as a future direction in the survey [28], and more detail is provided in Chapter 3, where mixed slicing is introduced as the foundation for the later saliency-based contributions.

2.5.2 Contribution 2 - Combining saliency and erasing strategies

Building on Contribution 1, which highlights the importance of preserving discriminative slices, we next combine slicing with saliency. Information-erasing data augmentation techniques, as discussed in Section 2.2.2, promote diversity by providing occlusion perspectives in different ways, but there are high chances of either completely erasing targeted objects (as shown in Figure 4.1a), leading to noisy image data, or erasing contextual information (as shown in Figure 4.1b) that may force the model to learn only the most important information and result in overfitting. To trade off between these erasing issues, while still providing an occlusion perspective, we propose a simple yet effective data augmentation, RandSaliencyAug, which detects salient regions in the image and applies any of the proposed strategies (Row Slice Erasing, Column Slice Erasing, Row-Column Saliency Erasing, Partial Saliency Erasing, Horizontal Half Saliency Erasing and Vertical Half Saliency Erasing) from the search space either randomly or based on the performance of the model. The proposed approach neither removes the complete object in images like RE and Cutout nor masks a large number of squares like HaS and GM. The overall process of the RandSaliencyAug approach is shown in Figure 4.3. More detail is provided in Chapter 4, where RandSaliencyAug is analysed as a saliency-aware

extension of the mixed slicing framework.

2.5.3 Contribution 3 - Balancing data diversity and information preservation

Extending the saliency-based ideas from Contribution 2, we next focus on explicitly balancing diversity and information preservation. Methods such as SalfMix [30], KeepAugment [31], and RandAugment [52], as discussed in Section 2.2.3, have addressed computational and feature fidelity challenges. However, it is important to note that these methods, while promoting data diversity, can introduce noise and reduce feature fidelity, thus impacting overall model performance. For instance, SalfMix can lead to overfitting by repeating the salient part in the image [30], while KeepAugment can introduce a domain shift between salient and non-salient regions, hindering contextual information [31]. To address these challenges and simultaneously increase feature fidelity and diversity, we propose a simple yet effective data augmentation technique named KeepOriginalAugment. Our approach differs from KeepAugment in two key aspects. Firstly, in KeepAugment, non-salient regions remain unchanged during data augmentation, whereas our approach allows augmentation of these regions. Secondly, KeepAugment fixes the placement of salient regions, whereas our proposed method varies the placement to achieve scaled augmentation. More detail is provided in Chapter 5, where KeepOriginalAugment complements RandSaliencyAug by explicitly balancing salient and non-salient content.

2.5.4 Contribution 4 - Assessing and mitigating bias

Finally, building directly on Contributions 2 and 3, we transfer the proposed augmentations to fairness-sensitive settings. We extend RandSaliencyAug and KeepOriginalAugment (renamed as FaceRandSaliencyAug and FaceKeepOriginalAugment for facial data) to assess dataset diversity and reduce bias (a detailed literature review is provided in Section 2.3.1) across different datasets and professions categories. Additionally, we evaluate gender bias in CNNs and ViTs using these augmentations.

Lastly, we introduce a saliency-based metric to measure the bias and fairness of these datasets. Further details are provided in Chapter 6, where the robustness-oriented contributions are connected to fairness and bias mitigation.

Chapter 3

Random Slice Mixing: Minimising Feature Loss while Enhancing Diversity

3.1 Introduction

This chapter answers Research Question (RQ1), which focuses on finding the most effective way to combine slices from different images for data augmentation so that we can have the best results in reducing feature loss and enhancing feature diversity for better model generalisation. It does so by outlining the methodology used and analysing the results. Prior to that, we discuss the limitations of existing data augmentation techniques. Single-image deletion data augmentation methods as Random Erasing (RE), Hide-and-Seek (HS), GridMask, and CutOut—have been widely explored. Likewise, multi-image mixing approaches like CutMix, MixUp, and RICAP have also been studied [10]. These two categories are illustrated in Figure 3.1. While single-image deletion methods risk removing important features and potentially reducing model performance, multi-image mixing techniques introduce variation from different perspectives. However, none of these methods consider augmentation through slice-based mixing. To the best of our knowledge, we are the first

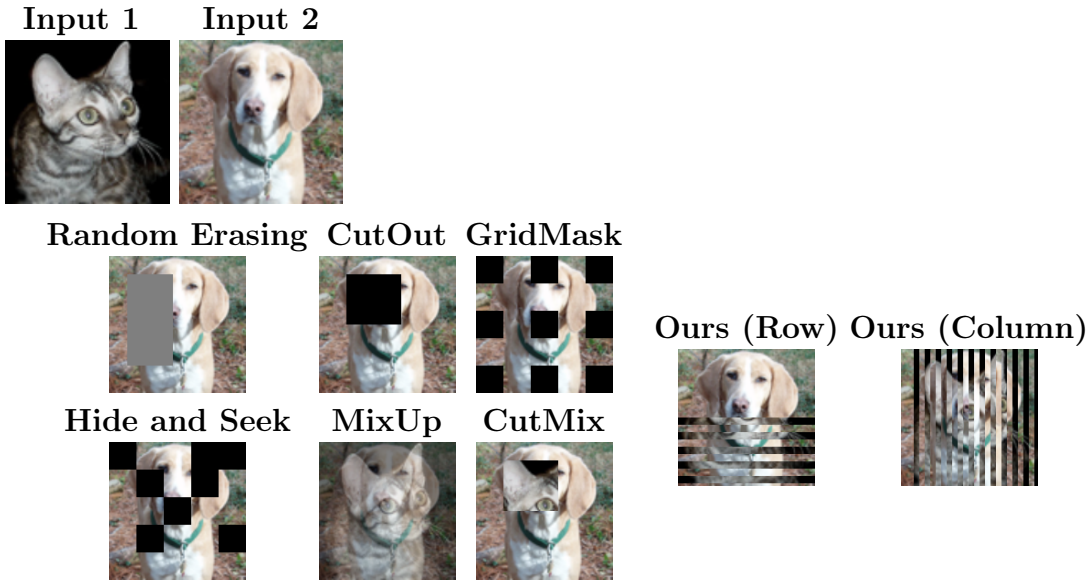


Figure 3.1: Comparison of different data augmentations against the proposed RSMDA.

to explore random slices mixing data augmentation (RSMDA) to check the effect of the proposed data augmentation technique using different strategies, namely, the horizontal (row-wise) RSDMA strategy, the vertical (column-wise) RSDMA strategy, and a mixture of both, as shown in Figure 3.2.

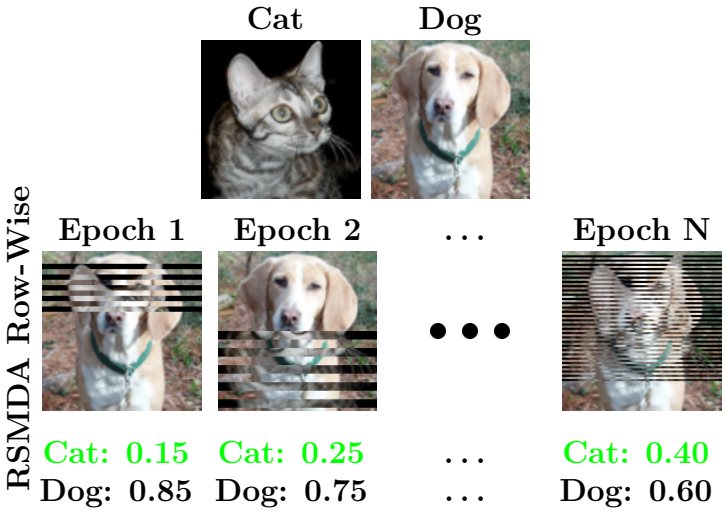
3.2 Methodology

Let $x \in \mathbf{R}^{W \times H \times C}$ and y represent a training image and its label, respectively. The main idea of RSMDA is to create a new training image with its label $(\tilde{x}, \tilde{y})$. For that purpose, we select two training samples with corresponding labels, (x_1, y_1) and (x_2, y_2). A combination of these training samples can be defined as:

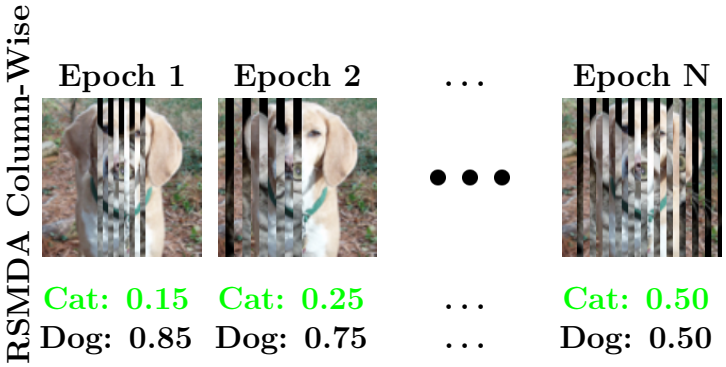
$$\tilde{x} = M \odot x_1 + (1 - M) \odot x_2 \tag{3.1}$$

Additionally, their labels' combination is defined as:

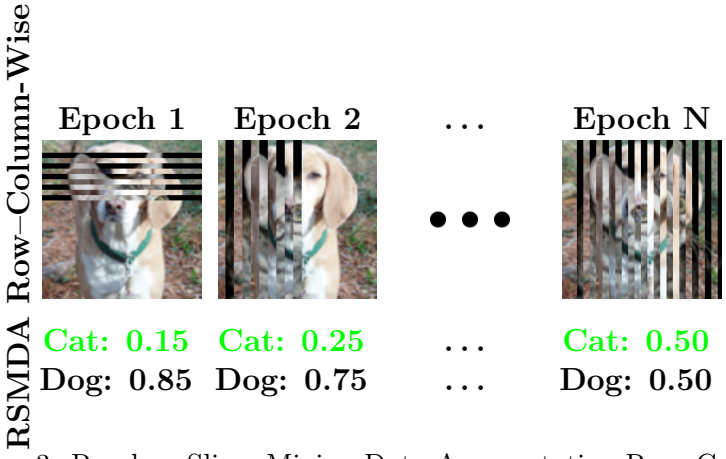
$$\tilde{y} = \lambda \odot y_1 + (1 - \lambda) \odot y_2 \tag{3.2}$$



(a) Strategy 1: Random Slices Mixing Data Augmentation Row-Wise (RSMDA-R)



(b) Strategy 2: Random Slices Mixing Data Augmentation Column-Wise (RSMDA-C)



(c) Strategy 3: Random Slices Mixing Data Augmentation Row–Column-Wise (RSMDA-RC)

Figure 3.2: Three strategies for Random Slices Mixing Data Augmentation (RSMDA).

where $M \in [0,1]^{W \times H}$ is a binary mask that is filled with 0 and 1, where 0 and 1 are for excluding and including the image pixel, respectively. The symbol $\odot$ shows element-wise multiplication, and λ is the combination ratio of the two images and their labels, which is sampled from a beta distribution, such as CutMix [14] or MixUp [27]. In $Beta(\alpha, \alpha)$, we set alpha to 1, following previous data augmentation, and λ is distributed from the normal distribution.

For sampling the binary mask M, we randomly obtain slices of the size S in a certain range from 1 to half of width or height. To obtain the total number of slices, we divide W or H by S.

In the case of the total number of column slices:

$$TotalSlices = \lfloor W/S \rfloor \tag{3.3}$$

In the case of the total number of row slices:

$$TotalSlices = \lfloor H/S \rfloor \tag{3.4}$$

The next step is to ascertain how many slices should be mixed. To obtain this, we multiply the total slices by λ, which gives the number of slices to be mixed.

$$N_{mix} = \lfloor \lambda \times TotalSlices \rfloor \tag{3.5}$$

In Equation (3.5), N_{mix} is the number of slices to be selected from the target sample and pasted to the source image. To do so, we fill N_{mix} number of slices in mask M with 1 to select the slices from the target image. In order to generate an augmented sample pair $(\tilde{x}, \tilde{y})$, the selected slices from the target image are pasted to a source image. Then, the $(\tilde{x}, \tilde{y})$ augmented pair is used for training the model.

Furthermore, we propose and investigate the three different strategies of the proposed technique. Each of them is discussed below:

- **Random slices mixing row-wise (RSMDA-R)**: In this strategy, we obtain

N_{mix} number of slices horizontally from the target image and paste it to the source image. Their corresponding labels are also mixed following the whole process discussed in Section 3.2. RSMDA-R is shown in Figure 3.2a.

- **Random slices mixing column-wise (RSMDA-C)**: This is another strategy that we explore, in which we follow the same method used in RSMDA-R, except that we obtain the slices vertically, as shown in Figure 3.2b.

- **Random slices mixing row–column-wise (RSMDA-RC)**: This is the third strategy. We apply both RSMDA-R and RSMDA-C based on binary randomness to each learning step, as shown in Figure 3.2c.

3.3 Experiments

In this section, we discuss the experimental setup, the datasets used, and the results.

3.3.1 Experimental Setup

In our work, we used many network types, such as Resnet [1], VGG [3], and PyramidNet [61]. For a fair comparison with single image based mixing data augmentation (SIBDA) techniques , we employed the same parameters as in [18]; the later used 300 epochs, an initial learning rate of 0.1, continuous reduction by 10 at certain epochs (100, 150, 175, and 190), and a batch size set to 64. The probability of performing RSMDA was set to 0.5, similar to RE, and we also checked 10 different probabilities, as mentioned in Section 3.3.3. For a fair comparison with Multi-Image based mixing data augmentation (MIBDA) techniques, we used the same setting as was mentioned in CutMix [14], where the epochs were 300, the batch size was 128, the initial learning rate was 0.1, the momentum was 0.9, and the learning rate decayed by 10 after every 30 epochs. We performed all of the experiments using a PyTorch module with 2 NVIDIA GetForce RTX 2080 Ti GPUs. Similar to the previous settings, each experiment was repeated at least three times unless otherwise mentioned.

3.3.2 Datasets

To validate the proposed approach, four different datasets were used. The datasets include color color datasets of different sizes of images, such as CIFAR10 [62], CIFAR100 [62], and STL10 [63]. Then, a grayscale dataset, such as fashionMNIST [64] is used for the experiments.

FashionMNIST

The fashionMNIST dataset consists of 60,000 training and 10,000 test images. Each image is in grayscale and has the dimensions of 28×28, and there are 10 clothing classes in this dataset, namely, t-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot. It is important to note, we re-performed the baseline and RE experiments for the fashionMNIST dataset as the original experiments [18] performed for the baseline and RE were on the old fashionMNIST dataset. The issue with the old fashionMNIST dataset is that a few test and training images overlapped with each other, as discussed in the GitHub repository (`https://github.com/zhunzhong07/Random-Erasing/issues/9`, accessed on 12/01/2022) of RE [18].

CIFAR10 and CIFAR100

Both the CIFAR10 and CIFAR100 datasets have an equal number of training and test images. Each dataset has 50,000 training and 10,000 test images, and each image is an RGB color image and has the dimensions of $32 \times 32 \times 3$. There are 10 and 100 classes in the cifar10 and cifar100 datasets, respectively.

STL10

We shifted the experiments to slightly greater dimensions, so we chose the STL10 dataset. It has only 500 training images and 8000 test images. Each image is in RGB color and of the dimensions $96 \times 96 \times 3$. There are 10 classes in this dataset. Images in this dataset are taken from one of the biggest datasets, ImageNet [2].

3.3.3 Results

Hyperparameter Study

In our approach, we find the best probability of performing RSMDA and the best slice size using the ResNet20 model and the fashionMNIST dataset. In order to find the best probability, we performed experiments using different probabilities starting from 0.1 to 1.0 with 0.1 intervals, and it was found that 0.5 was the best probability. To find the optimal slice size, we investigate a slice size of two as the minimum, one-third of the image height or width as the maximum, and a random slice size between the minimum and maximum with each batch of images. The best parameters were found to be 0.5 for the best probability, and the random slice size was the best slice size, as shown in Figure 3.3, in which the x-axis, the three different color lines, and the y-axis show the probability of performing RSMDA, the slice size, and accuracy, respectively. For all of the remaining experiments, we used the best parameters. Note, here, accuracy is defined as the percentage of correctly predicted samples,

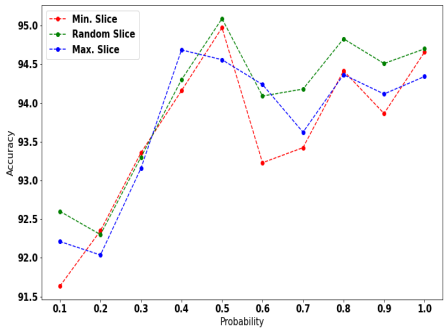


Figure 3.3: Hyperparameters: probability and slice size effect on accuracy.

which is mathematically defined as:

$$A = 100 * C/T \tag{3.6}$$

where A, C, and T are the accuracy percentage, the number of correctly predicted samples, and the total number of samples, respectively. Therefore, higher accuracy is preferred. The error rate is the percentage of incorrectly predicted samples and

can be defined as:

$$E = 100 - A \tag{3.7}$$

where E is the error rate, and A is the accuracy percentage. Therefore, a lower error rate is preferred.

Table 3.1: Performance comparison of the proposed approach with random erasing and baseline using the error rate metric, a lower value is better. First and second best performances are highlighted in blue and red color, respectively.

Models	Baselines	RE	RSMDA (R)	RSMDA(C)	RSMDA(RC)
Fashion-MNIST					
ResNet20	6.21± 0.11	5.04 ± 0.10	4.91 ± 0.12	4.72 ± 0.13	4.76 ± 0.06
Resnet32	6.04 ± 0.13	4.84 ± 0.12	4.81 ± 0.17	4.65 ± 0.15	4.81 ± 0.12
Resnet44	6.08 ± 0.16	4.87 ± 0.1	4.07 ± 0.14	4.784 ± 0.01	4.9 ± 0.25
Resnet56	6.78 ± 0.16	5.02 ± 0.11	5.00 ± 0.19	5.00 ± 0.2	5.09 ± 0.59
CIFAR10					
Resnet20	7.21 ± 0.17	6.73 ± 0.09	7.18 ± 0.13	7.38 ±0.254	7.48 ± 1.08
Resnet32	6.41 ± 0.06	5.66 ± 0.10	6.31 ± 0.14	6.06 ± 0.101	6.21 ± 0.76
Resnet44	5.53 ± 0.0	5.13 ± 0.09	5.09 ± 0.10	5.26 ± 0.262	5.51 ± 0.06
Resnet56	5.31 ± 0.07	4.89 ± 0.0	5.02 ± 0.11	5.28 ± 0.02	5.97 ± 0.47
VGG11	7.88±0.76	7.82±0.65	7.80±0.65	7.82 ± 0.27	7.81 ± 0.57
VGG13	6.33±0.23	6.22±0.63	6.18±0.54	6.31 ± 0.266	6.20 ± 0.38
VGG16	6.42±0.34	6.21±0.76	6.20±0.34	6.26 ± 0.196	6.35 ± 0.76
CIFAR100					
Resnet20	30.84 ± 0.19	29.97 ± 0.11	30.18 ± 0.27	30.28 ± 0.33	30.46 ± 0.79
Resnet32	28.50 ± 0.37	27.18 ± 0.32	27.08 ± 0.34	28.22 ± 0.22	28.42 ± 0.12
Resnet44	25.27 ± 0.21	24.29 ± 0.16	24.49 ± 0.23	25.21 ± 0.57	25.08 ± 0.13
Resnet56	24.82 ± 0.27	23.69 ± 0.33	23.35 ± 0.26	24.33 ± 0.12	24.91 ± 0.57
VGG11	28.97±0.76	28.73±0.67	28.26±0.75	28.92 ± 0.33	28.29 ± 0.43
VGG13	25.73±0.67	25.71±0.54	25.71±0.56	25.72 ± 0.26	25.72 ± 0.42
VGG16	26.64±0.56	26.63±0.75	26.61±0.65	26.63 ± 1.77	26.63 ± 0.66
STL10					
VGG11	22.29±0.13	22.27±0.21	20.68±0.23	21.49 ± 0.02	20.79 ± 0.33
VGG13	20.64±0.26	20.18±0.23	19.91±0.92	19.60 ± 0.12	19.7 ± 0.23
VGG16	20.62±0.34	20.12±0.65	20.09±0.23	20.35 ± 0.03	20.49 ± 0.44

Classification Results

We performed a number of experiments using different networks and datasets with three strategies. First, we compare our three strategies' results with random erasing data augmentation and the baseline of different models, as shown in Table 3.1, where RSMDA(R), RSMDA(C), and RSMDA(RC) show RSMDA row-wise (hor-

izontally), RSMDA column-wise (vertically), and RSMDA row–column-wise, respectively. The error rate is reported, and a lower rate is better. In Table 3.1, RSMDA(R) has better performance than the baseline and random erasing in almost all experiments. In fashionMNIST, RSMDA(C) showed the best performance of all other methods. In the case of the CIFAR10 dataset, RSMDA(R) was more successful, especially using the VGG network type. Using the CIFAR100 dataset, again, RSMDA(R) showed better performance than random erasing, and it showed better performance using the VGG network type. In the case of the STL10 dataset, RSMDA(R) was a winner compared to the baseline and random erasing. Overall, RSMDA(R) has shown a huge performance improvement, and we consider it the best for the rest of our experiments using the optimal hyperparameters discussed in Section 3.3.3. We compare our proposed approach with different dropout

Table 3.2: Comparison of state-of-the-art regularisation methods on CIFAR-100. First and second best performances are highlighted in blue and red color, respectively.

PyramidNet-200 ($\tilde{\alpha} = 240$) (Params: 26.8 M)	**Top-1 Err (%)**	**Top-5 Err (%)**
Baseline	16.45	3.69
+ StochDepth [65]	15.86	3.33
+ Label smoothing ($\epsilon = 0.1$) [66]	16.73	3.37
+ Cutout [26]	16.53	3.65
+ Cutout + Label smoothing ($\epsilon = 0.1$)	15.61	3.88
+ DropBlock [8]	15.73	3.26
+ DropBlock + Label smoothing ($\epsilon = 0.1$)	15.16	3.86
+ Mixup ($\alpha = 0.5$) [27]	15.78	4.04
+ Mixup ($\alpha = 1.0$) [27]	15.63	3.99
+ Manifold Mixup ($\alpha = 1.0$) [67]	16.14	4.07
+ Cutout + Mixup ($\alpha = 1.0$)	15.46	3.42
+ Cutout + Manifold Mixup ($\alpha = 1.0$)	15.09	3.35
+ ShakeDrop [68]	15.08	2.72
+ RSMDA(R)	15.03	3.01
+ CutMix	14.47	2.97

methods and data augmentations using the best hyperparameters and the proposed strategy, as shown in Table 3.2. In this comparison, we employed a large model, PyramidNet-200, with 26.8 million parameters using the CIFAR100 dataset. Our approach, RSMDA(R), outperformed all of the aforementioned dropout methods

and SOTA multi-image methods, except for CutMix, but it showed a competitive performance as compared to CutMix. It placed second showing error rates for top-1 and top-5 of 15.03 and 3.01, respectively. We compare the results based on the top-1 error % by following the trend as mentioned in Table 5 of the work [14].

Table 3.3: Lighter architectures on CIFAR-100. First and second best performances are highlighted in blue and red color, respectively.

Model	**Params**	**Top-1 (%)**	**Top-5 Err (%)**
PyramidNet-110 ($\tilde{\alpha} = 64$) [61]	1.7M	19.85	4.66
PyramidNet-110+ RSMDA	1.7M	19.29	4.42
PyramidNet-110+ CutMix	1.7M	17.97	3.83
ResNet-110	1.1M	23.14	5.95
ResNet-110+ RSMDA	1.1M	22.87	5.93
ResNet-110+ CutMix	1.1M	20.11	4.43

Although CutMix [14] achieves the lowest top-1 error on CIFAR-100 with PyramidNet-200, the proposed RSMDA(R) attains very similar performance (within ∼0.6 percentage points in top-1 error) as shown in Table 3.2 while offering complementary advantages. First, RSMDA introduces a *structured slice-based mixing* scheme: instead of replacing one contiguous rectangular patch as in CutMix, it mixes multiple row- or column-wise slices. This produces a richer set of partial views of both images and distributes mixed content more uniformly across the spatial extent, which is reflected in the CAM analysis (Figure 3.10), where RSMDA encourages the network to focus on fine-grained, discriminative regions such as the dog's tail rather than only large, highly salient parts. Second, the hyperparameter study in Figure 3.3 shows that RSMDA is *stable across a broad range of probabilities and slice sizes*, with a simple default configuration (probability 0.5 and random slice size) that works consistently well across datasets and architectures; in contrast, CutMix requires carefully tuned patch sizes and λ distributions to avoid overly aggressive occlusions. Third, while CutMix is designed purely as a regulariser, RSMDA is explicitly motivated by *feature-loss reduction in single-image erasing methods*: by recombining multiple slices rather than deleting them, it recovers information that Cutout/Random Erasing permanently remove and integrates naturally with the saliency-aware

extensions developed in later chapters. Overall, even when CutMix attains slightly lower error on a specific setting, RSMDA delivers *competitive accuracy with a different inductive bias* prioritising fine-grained feature learning, , controllable mixing patterns, and robust behaviour across architectures and datasets.

Table 3.4: Performance on the Oxford-IIIT Pet Dataset

Model	**Accuracy (%)**
ResNet50 (Baseline)	70.92
MobileNet V2 (Baseline)	77.43
ResNet50 (RSMDA-R)	72.20
MobileNet V2 (RSMDA-R)	80.16

The results in Table 3.4present the classification performance of two backbone architectures, ResNet50 and MobileNetV2, on the Oxford-IIIT Pet dataset. Both baseline models and their enhanced versions using the proposed RSMDA (R) method are evaluated. Incorporating RSMDA (R) leads to consistent performance improvements across architectures. Specifically, ResNet50 improves from 70.92% to 72.20%, while MobileNetV2 increases from 77.43% to 80.16%. These results demonstrate the effectiveness of RSMDA (R) in enhancing model robustness and generalization.

3.3.4 Evaluation using different metrics

In this section, we present class-wise results using additional metrics—precision, recall, F1 score, and ROC AUC—to better characterise performance across classes. Detailed results for the different models and datasets are reported in Table 3.5 for FashionMNIST, Tables 3.6 and 3.7 for CIFAR10, Table 3.8 for CIFAR100, and Table 3.9 for STL10. For CIFAR100, which has 100 classes, it is impractical to present all class-wise results in the main text, so we provide the full results as supplementary material at `https://github.com/kmr2017/ThesisPhDCode/tree/main/RevisionExperiments/RQ1Files/CIFAR100`. Furthermore, we show confusion matrices for the best-performing models in Figures 3.4, 3.5, and 3.6.

Table 3.5: Class-wise metrics on FashionMNIST for different models with RSMDA.

Class	Precision	Recall	F1	ROC AUC
ResNet20 with RSMDA(C)				
T-shirt/top	0.9509	0.9290	0.9398	0.9887
Trouser	0.9697	0.9910	0.9802	0.9972
Pullover	0.9347	0.9310	0.9329	0.9918
Dress	0.9711	0.9410	0.9558	0.9923
Coat	0.9329	0.9310	0.9319	0.9884
Sandal	0.9723	0.9840	0.9781	0.9969
Shirt	0.9395	0.9160	0.9276	0.9876
Sneaker	0.9635	0.9760	0.9697	0.9950
Bag	0.9631	0.9920	0.9773	0.9985
Ankle boot	0.9603	0.9680	0.9641	0.9922
Mean	**0.9558**	**0.9559**	**0.9558**	**0.9929**
ResNet32 with RSMDA(C)				
T-shirt/top	0.9647	0.9280	0.9460	0.9897
Trouser	0.9802	0.9890	0.9846	0.9981
Pullover	0.9451	0.9300	0.9375	0.9895
Dress	0.9745	0.9570	0.9657	0.9931
Coat	0.9465	0.9380	0.9422	0.9918
Sandal	0.9676	0.9870	0.9772	0.9975
Shirt	0.9327	0.9280	0.9303	0.9924
Sneaker	0.9713	0.9810	0.9761	0.9946
Bag	0.9584	0.9900	0.9739	0.9956
Ankle boot	0.9655	0.9790	0.9722	0.9950
Mean	**0.9606**	**0.9607**	**0.9606**	**0.9937**
ResNet44 with RSMDA(R)				
T-shirt/top	0.9511	0.9150	0.9327	0.9858
Trouser	0.9717	0.9940	0.9827	0.9976
Pullover	0.9391	0.9410	0.9401	0.9893
Dress	0.9775	0.9550	0.9661	0.9920
Coat	0.9541	0.9360	0.9450	0.9890
Sandal	0.9707	0.9940	0.9822	0.9983
Shirt	0.9364	0.9130	0.9246	0.9845
Sneaker	0.9724	0.9880	0.9802	0.9968
Bag	0.9640	0.9910	0.9773	0.9966
Ankle boot	0.9694	0.9810	0.9751	0.9943
Mean	**0.9606**	**0.9608**	**0.9606**	**0.9924**
ResNet56 with RSMDA(R)				
T-shirt/top	0.9464	0.9010	0.9232	0.9859
Trouser	0.9734	0.9880	0.9806	0.9972
Pullover	0.9369	0.9200	0.9284	0.9884
Dress	0.9641	0.9400	0.9519	0.9907
Coat	0.9388	0.9350	0.9369	0.9899
Sandal	0.9695	0.9840	0.9767	0.9962
Shirt	0.9091	0.9200	0.9145	0.9897
Sneaker	0.9633	0.9720	0.9676	0.9946
Bag	0.9527	0.9860	0.9690	0.9959
Ankle boot	0.9623	0.9710	0.9667	0.9934
Mean	**0.9516**	**0.9517**	**0.9515**	**0.9922**

Table 3.6: Class-wise metrics on CIFAR10 for different Resenet models with RSMDA.

Class	Precision	Recall	F1	ROC AUC
Resnet20 with RSMDA(R)				
Airplane	0.9194	0.9360	0.9277	0.9815
Automobile	0.9441	0.9630	0.9535	0.9900
Bird	0.9518	0.9080	0.9294	0.9706
Cat	0.9107	0.8870	0.8987	0.9398
Deer	0.9311	0.9330	0.9321	0.9708
Dog	0.9044	0.9180	0.9112	0.9532
Frog	0.9481	0.9310	0.9395	0.9720
Horse	0.9571	0.9360	0.9464	0.9733
Ship	0.9391	0.9720	0.9553	0.9894
Truck	0.9384	0.9590	0.9486	0.9895
Mean	**0.9344**	**0.9343**	**0.9342**	**0.9730**
Resnet32 with RSMDA(C)				
Airplane	0.9483	0.9530	0.9506	0.9869
Automobile	0.9452	0.9660	0.9555	0.9946
Bird	0.9559	0.9320	0.9438	0.9861
Cat	0.9162	0.9070	0.9116	0.9692
Deer	0.9376	0.9160	0.9267	0.9812
Dog	0.9289	0.9270	0.9279	0.9690
Frog	0.9439	0.9420	0.9429	0.9874
Horse	0.9397	0.9350	0.9373	0.9860
Ship	0.9550	0.9760	0.9654	0.9957
Truck	0.9489	0.9660	0.9574	0.9933
Mean	**0.9419**	**0.9420**	**0.9419**	**0.9849**
Resnet44 with RSMDA(R)				
Airplane	0.9557	0.9700	0.9628	0.9942
Automobile	0.9586	0.9720	0.9652	0.9964
Bird	0.9693	0.9470	0.9580	0.9918
Cat	0.9343	0.9100	0.9220	0.9801
Deer	0.9379	0.9370	0.9375	0.9910
Dog	0.9366	0.9160	0.9262	0.9831
Frog	0.9660	0.9670	0.9665	0.9935
Horse	0.9605	0.9490	0.9547	0.9931
Ship	0.9602	0.9900	0.9749	0.9982
Truck	0.9511	0.9730	0.9619	0.9973
Mean	**0.9530**	**0.9531**	**0.9530**	**0.9919**
Resnet56 with RSMDA(R)				
Airplane	0.9547	0.9700	0.9623	0.9963
Automobile	0.9649	0.9910	0.9778	0.9985
Bird	0.9765	0.9150	0.9448	0.9855
Cat	0.9336	0.9140	0.9237	0.9846
Deer	0.9540	0.9340	0.9439	0.9895
Dog	0.9281	0.9300	0.9291	0.9869
Frog	0.9696	0.9570	0.9633	0.9947
Horse	0.9519	0.9700	0.9609	0.9941
Ship	0.9603	0.9910	0.9754	0.9981
Truck	0.9638	0.9850	0.9743	0.9980
Mean	**0.9558**	**0.9557**	**0.9555**	**0.9926**

Table 3.7: Class-wise metrics on CIFAR10 for different VGG models with RSMDA.

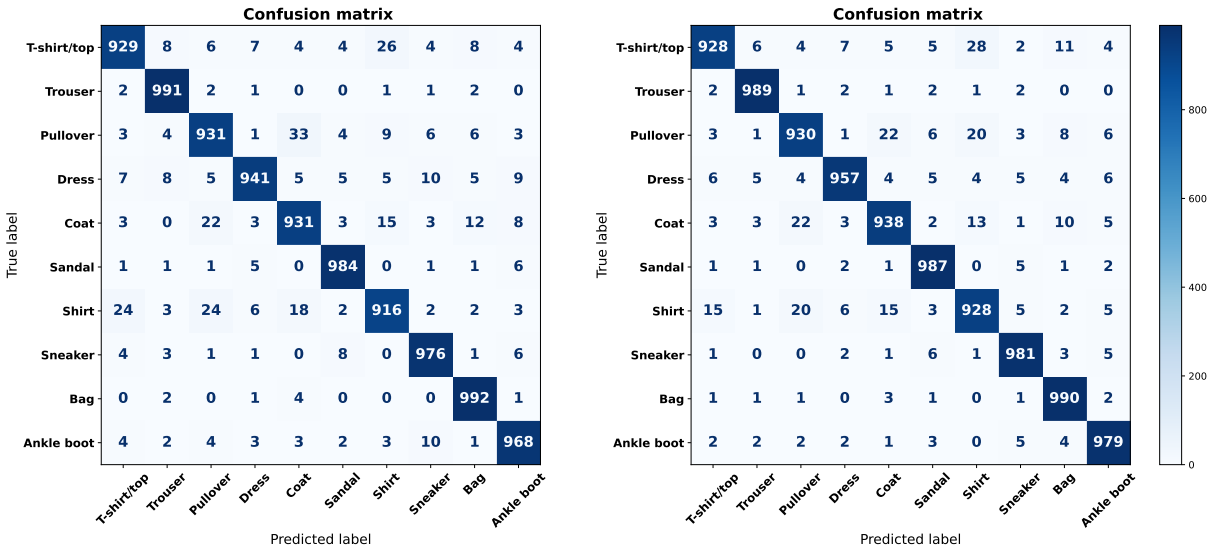
Class	Precision	Recall	F1	ROC AUC
VGG11 with RSMDA(R)				
Airplane	0.9364	0.9420	0.9392	0.9855
Automobile	0.9545	0.9650	0.9597	0.9933
Bird	0.9476	0.9220	0.9346	0.9790
Cat	0.9076	0.9040	0.9058	0.9579
Deer	0.9222	0.9010	0.9115	0.9700
Dog	0.9172	0.9080	0.9126	0.9608
Frog	0.9556	0.9470	0.9513	0.9842
Horse	0.9359	0.9340	0.9349	0.9779
Ship	0.9411	0.9750	0.9578	0.9938
Truck	0.9462	0.9670	0.9565	0.9922
Mean	**0.9364**	**0.9365**	**0.9364**	**0.9795**
VGG13 with RSMDA(R)				
Airplane	0.9500	0.9510	0.9505	0.9925
Automobile	0.9645	0.9770	0.9707	0.9957
Bird	0.9565	0.9240	0.9400	0.9813
Cat	0.9256	0.9200	0.9228	0.9623
Deer	0.9481	0.9310	0.9395	0.9832
Dog	0.9309	0.9290	0.9299	0.9729
Frog	0.9656	0.9530	0.9592	0.9907
Horse	0.9480	0.9480	0.9480	0.9897
Ship	0.9499	0.9860	0.9676	0.9969
Truck	0.9628	0.9830	0.9728	0.9968
Mean	**0.9502**	**0.9502**	**0.9501**	**0.9862**
VGG16 with RSMDA(R)				
Airplane	0.9539	0.9520	0.9530	0.9879
Automobile	0.9596	0.9730	0.9662	0.9921
Bird	0.9577	0.9280	0.9426	0.9827
Cat	0.9250	0.9130	0.9190	0.9574
Deer	0.9454	0.9350	0.9402	0.9789
Dog	0.9329	0.9310	0.9319	0.9649
Frog	0.9636	0.9540	0.9588	0.9862
Horse	0.9462	0.9490	0.9476	0.9873
Ship	0.9516	0.9840	0.9676	0.9937
Truck	0.9617	0.9790	0.9703	0.9931
Mean	**0.9498**	**0.9498**	**0.9497**	**0.9824**

Table 3.8: Mean metrics on CIFAR100 for different models with RSMDA. For full class-wise metrics, refer to the supplementary materialgithub.com.

Class	Precision	Recall	F1	ROC AUC
PyramidNet110 with RSMDA(R)				
Mean	**0.819**	**0.818**	**0.818**	**0.920**
PyramidNet200 with RSMDA(R)				
Mean	**0.861**	**0.860**	**0.860**	**0.936**
ResNet20				
Mean	**0.707**	**0.705**	**0.705**	**0.899**
ResNet32				
Mean	**0.740**	**0.739**	**0.738**	**0.907**
ResNet44				
Mean	**0.765**	**0.763**	**0.763**	**0.914**
ResNet56 with RSMDA(R)				
Mean	**0.780**	**0.779**	**0.779**	**0.937**
Resnet110 with RSMDA(R)				
Mean	**0.791**	**0.790**	**0.789**	**0.921**
VGG11 with RSMDA(R)				
Mean	**0.735**	**0.734**	**0.733**	**0.908**
VGG13 with RSMDA(R)				
Mean	**0.761**	**0.760**	**0.760**	**0.932**
VGG16 with - RSMDA(R)				
Mean	**0.750**	**0.749**	**0.748**	**0.934**

Table 3.9: Class-wise metrics on STL10 for different models with RSMDA.

Class	Precision	Recall	F1	ROC AUC
VGG11 with RSMDA(R)				
Airplane	0.7795	0.8350	0.8063	0.9547
Bird	0.8512	0.7650	0.8058	0.9429
Car	0.8401	0.8275	0.8338	0.9563
Cat	0.7403	0.7662	0.7531	0.9409
Deer	0.8303	0.7887	0.8090	0.9574
Dog	0.7556	0.7538	0.7547	0.9366
Horse	0.8174	0.8225	0.8199	0.9621
Monkey	0.7713	0.7800	0.7756	0.9518
Ship	0.8139	0.8475	0.8304	0.9641
Truck	0.8169	0.8200	0.8185	0.9578
Mean	**0.8017**	**0.8006**	**0.8007**	**0.9525**
VGG13 with RSMDA(RC)				
Airplane	0.8321	0.8425	0.8373	0.9706
Bird	0.8480	0.8087	0.8279	0.9703
Car	0.8239	0.8363	0.8300	0.9704
Cat	0.7515	0.7900	0.7703	0.9716
Deer	0.8374	0.8113	0.8241	0.9757
Dog	0.7982	0.7863	0.7922	0.9700
Horse	0.8254	0.8213	0.8233	0.9764
Monkey	0.7947	0.7837	0.7892	0.9703
Ship	0.8305	0.8450	0.8377	0.9673
Truck	0.8054	0.8175	0.8114	0.9564
Mean	**0.8147**	**0.8143**	**0.8143**	**0.9699**
VGG16 with RSMDA(R)				
Airplane	0.8138	0.8250	0.8194	0.9593
Bird	0.8353	0.8050	0.8199	0.9631
Car	0.8292	0.8313	0.8302	0.9588
Cat	0.7609	0.7638	0.7623	0.9598
Deer	0.8464	0.7850	0.8145	0.9622
Dog	0.7476	0.7887	0.7676	0.9669
Horse	0.8220	0.8137	0.8178	0.9676
Monkey	0.7917	0.7837	0.7877	0.9644
Ship	0.8292	0.8375	0.8333	0.9643
Truck	0.8048	0.8400	0.8220	0.9501
Mean	**0.8081**	**0.8074**	**0.8075**	**0.9617**



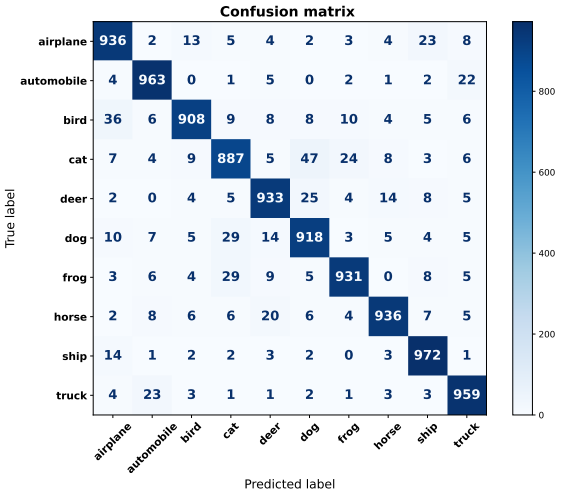
(a) ResNet-20 (RSMDA(C))

(b) ResNet-32 (RSMDA(C))

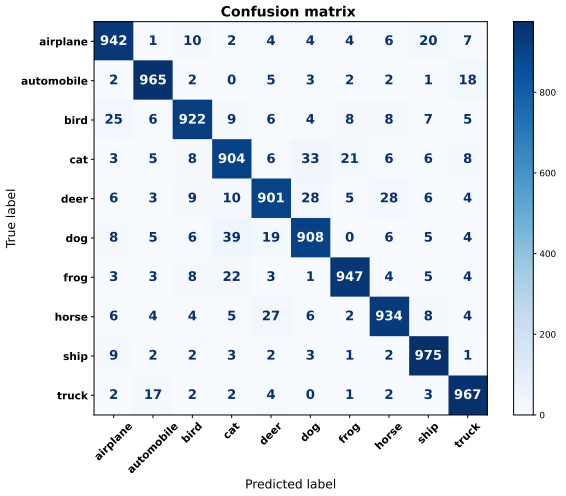
(c) ResNet-44 (RSMDA(R))

(d) ResNet-56 (RSMDA(R))

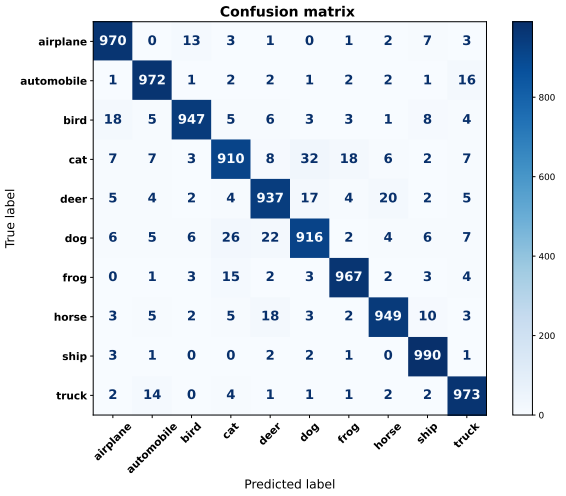
Figure 3.4: Confusion matrices on FashionMNIST for different ResNet models using RSMDA variants.



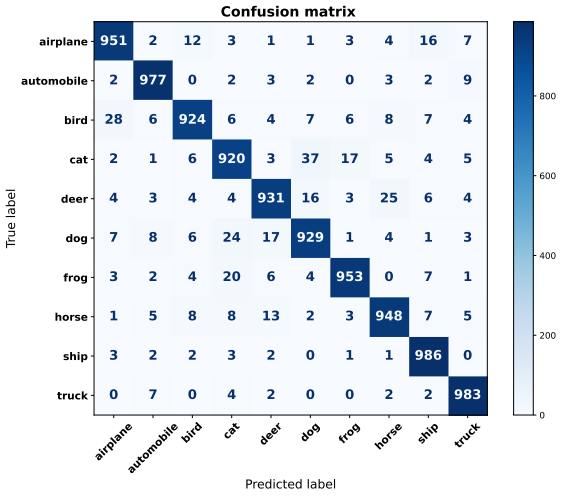
(a) ResNet-20



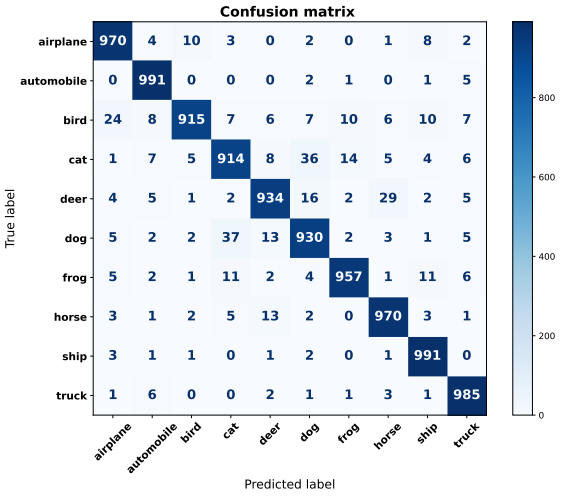
(b) VGG-11



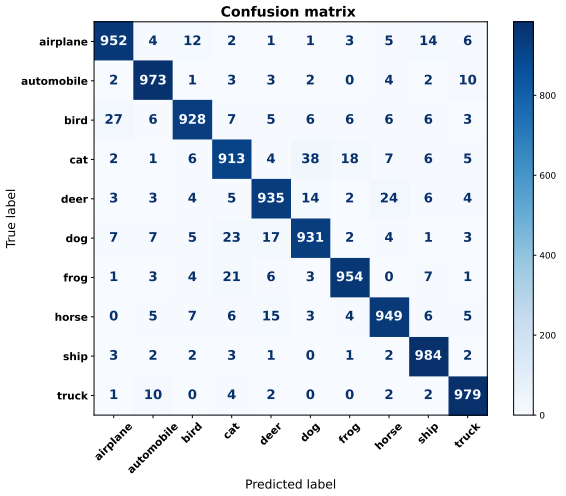
(c) ResNet-44



(d) VGG-13

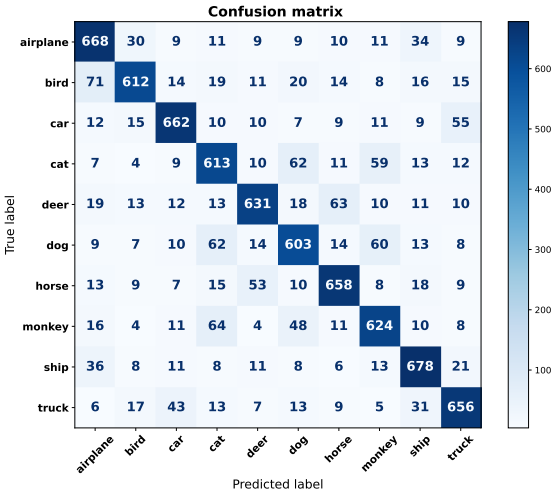


(e) ResNet-56

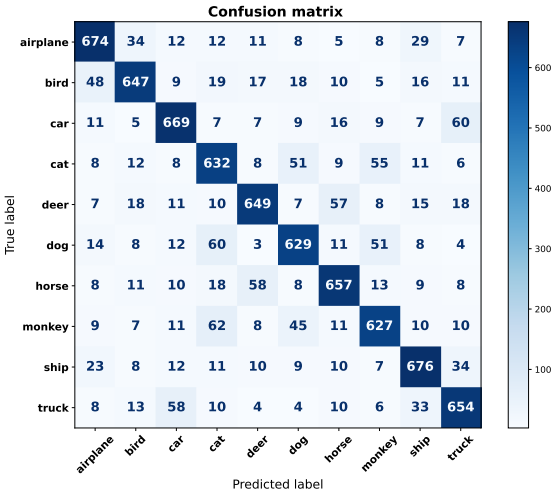


(f) VGG-16

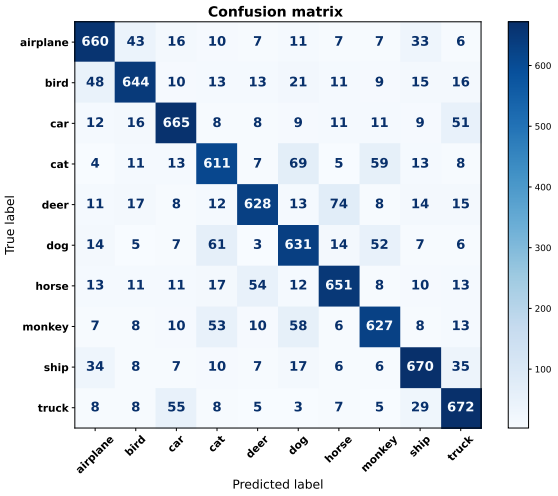
Figure 3.5: Confusion matrices on CIFAR-10 for different models using **RSMDA(R)**.



(a) VGG-11 (RSMDA(R))



(b) VGG-13 (RSMDA(RC))



(c) VGG-16 (RSMDA(R))

Figure 3.6: Confusion matrices on STL-10 for different VGG models using RSMDA variants.

Adversarial Attacks

Deep networks are fooled by adding a small unrecognizable perturbation to the input data, and the perturbed data mislead the network to degrade the performance; this whole mechanism is referred to as an adversarial attack [43, 69]. A simple way to prevent an attack is to generate an unseen input sample [44]. For adversarial attacks, we assume that the attacker has complete information about the model, i.e., a white box attack. We use different pre-trained (all models trained by us) ResNet models for CIFAR10, CIFAR100, and FashionMNIST. In this section, we evaluate the robustness of the proposed approach due to directly dealing with the input data following the previous methods [14]. In order to check the robustness, we compare our three strategies with the baseline and random erasing performance against two adversarial attacks, namely, the fast gradient sign method (FGSM) [43] and the FGSM variant, fast gradient magnitude (FGM) [43, 70], was proposed to alleviate the issue of noise perceptibility [70]. These two attacks, FGSM and FGM, were used to check robustness using different datasets and models. In all adversarial experiments, we used the baseline, RE, and the model trained by the three proposed strategies against these attacks using different values of epsilon [43], i.e., [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]. In Figure 3.7–3.9, the x-axis values and y-axis values show epsilon and accuracy, respectively. For the CIFAR10 dataset, we check the robustness against three strategies and compare it with the baseline and random erasing using different trained models of ResNet. In Figure 3.7, it can be seen that in the case of ResNet20 and ResNet44, interestingly, RSMDA(C) was more robust against both attacks than the others, while in the case of ResNet20 and ResNet56, RSMDA(R) was the winner. Overall, the proposed approach beats the baseline and random erasing.

For the CIFAR100 dataset, we repeat the same experiment pattern to check the behavior of the robustness using a dataset with a large number of classes. We checked the robustness using the CIFAR100 dataset. The pattern was very much different than what we discussed in the CIFAR10 case. As shown in Figure 3.8, RSMDA(RC)

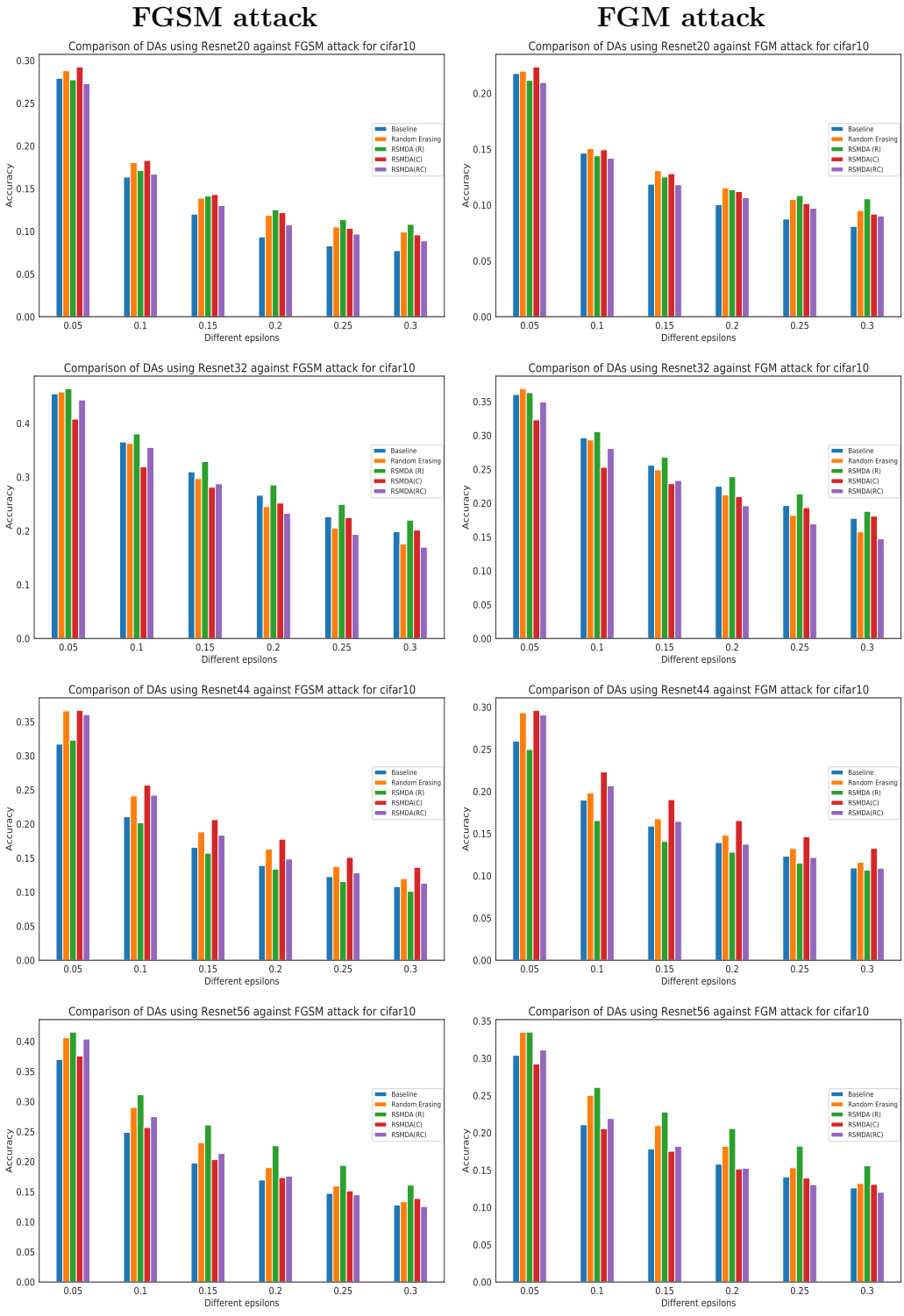


Figure 3.7: Comparison of DAs against different adversarial attacks for CIFAR10 dataset using different models.

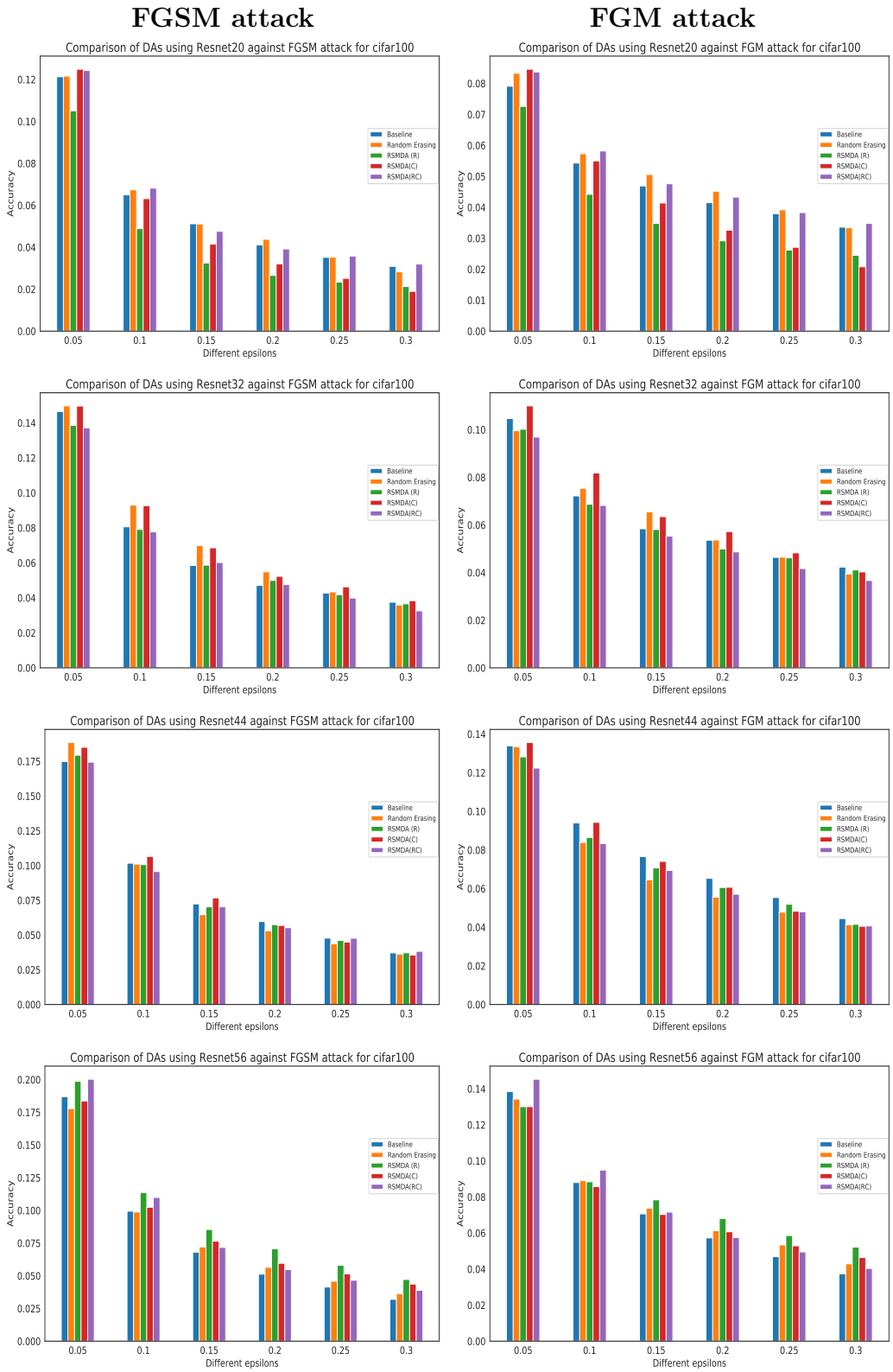


Figure 3.8: Comparison of DAs against different adversarial attacks for CIFAR100 dataset using different models.

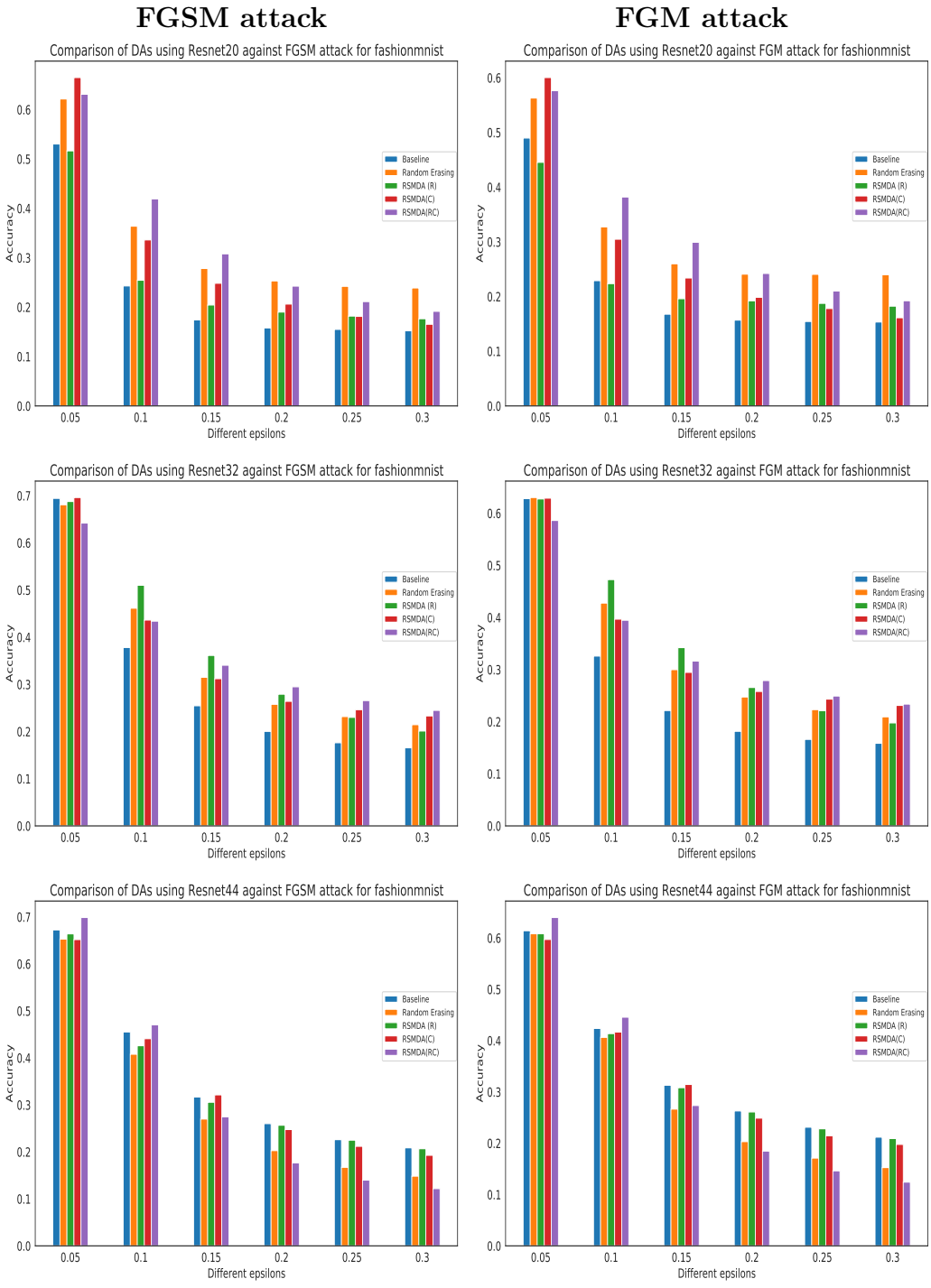


Figure 3.9: Comparison of DAs against different adversarial attacks for fashionMNIST dataset using different models.

was more robust using the ResNet20 and ResNet56 models, and RSMDA(C) was more successful using ResNet32 and ResNet44. In all of the models' cases, the proposed approach shows more robustness compared to the baseline and random erasing.

To check the behavior of the robustness of the grayscale dataset, we use the trained models on fashionMNIST against these adversarial attacks. We repeated the same stream of experiments, as shown in Figure 3.9. Surprisingly, the fashionMNIST case showed very different behavior from what was discussed in the CIFAR10 and cifar100 cases. In Figure 3.9, the overall RSMDA(RC) is more robust using the ResNet20 and ResNet44 models, and RSMDA(R) is more robust with ResNet32. The proposed approach also showed more robustness with the grayscale dataset.

Overall, the proposed approach is more robust. In rare cases, random erasing showed more robustness, such as with ResNet44 at an epsilon value of 0.05 in the cifar100 case (as shown in Figure 3.8), but, with an increase in the value of epsilon, it becomes less robust. The proposed approach's robustness not only checks in the case of different numbers of classes but also checks in different color domain datasets. In both the grayscale and color datasets, it significantly improves the robustness against adversarial attacks, except in some very rare cases.

Class Activation Map (CAM)

Class activation maps (`https://github.com/chaeyoung-lee/pytorch-CAM` [71, 72, 73], accessed on 10/09/2022) highlight different object regions of interest. They are plotted from the final layer of the convolutional neural network [72]. They help us to know where the model focuses more to ascertain whether the model is actually discriminating features or not. We compare this with relevant data augmentations, including CutOut, CutMix, and MixUp, to check whether RSMDA is really learning the discriminating features for two objects from their respective incomplete views. For this purpose, first, we take two images, i.e., the cat and dog shown in the first row of Figure 3.10. In the second row of Figure 3.10, we prepare the augmented

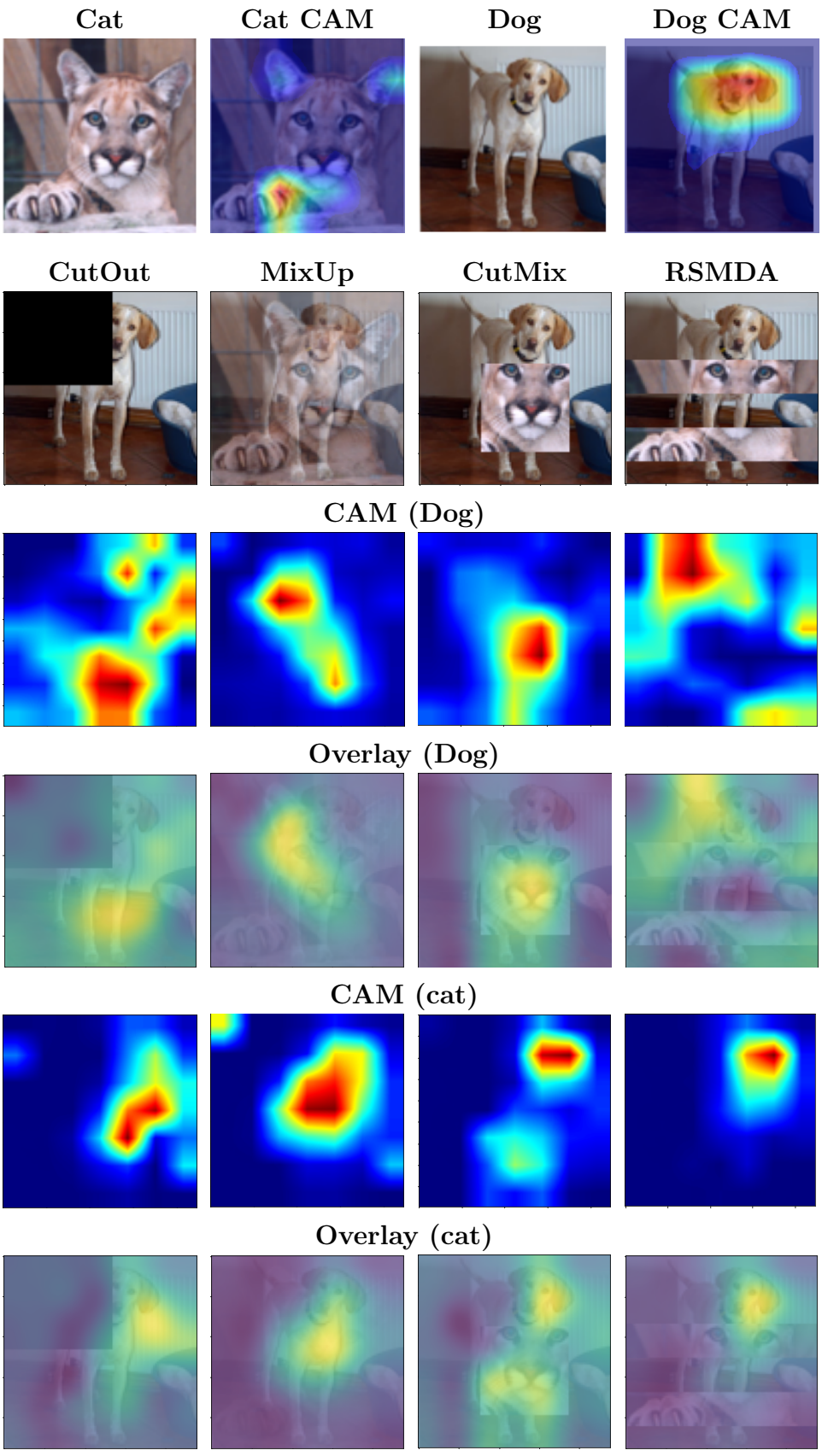


Figure 3.10: Data augmentation CAM visualisation comparison with Pre-trained resnet50 model.

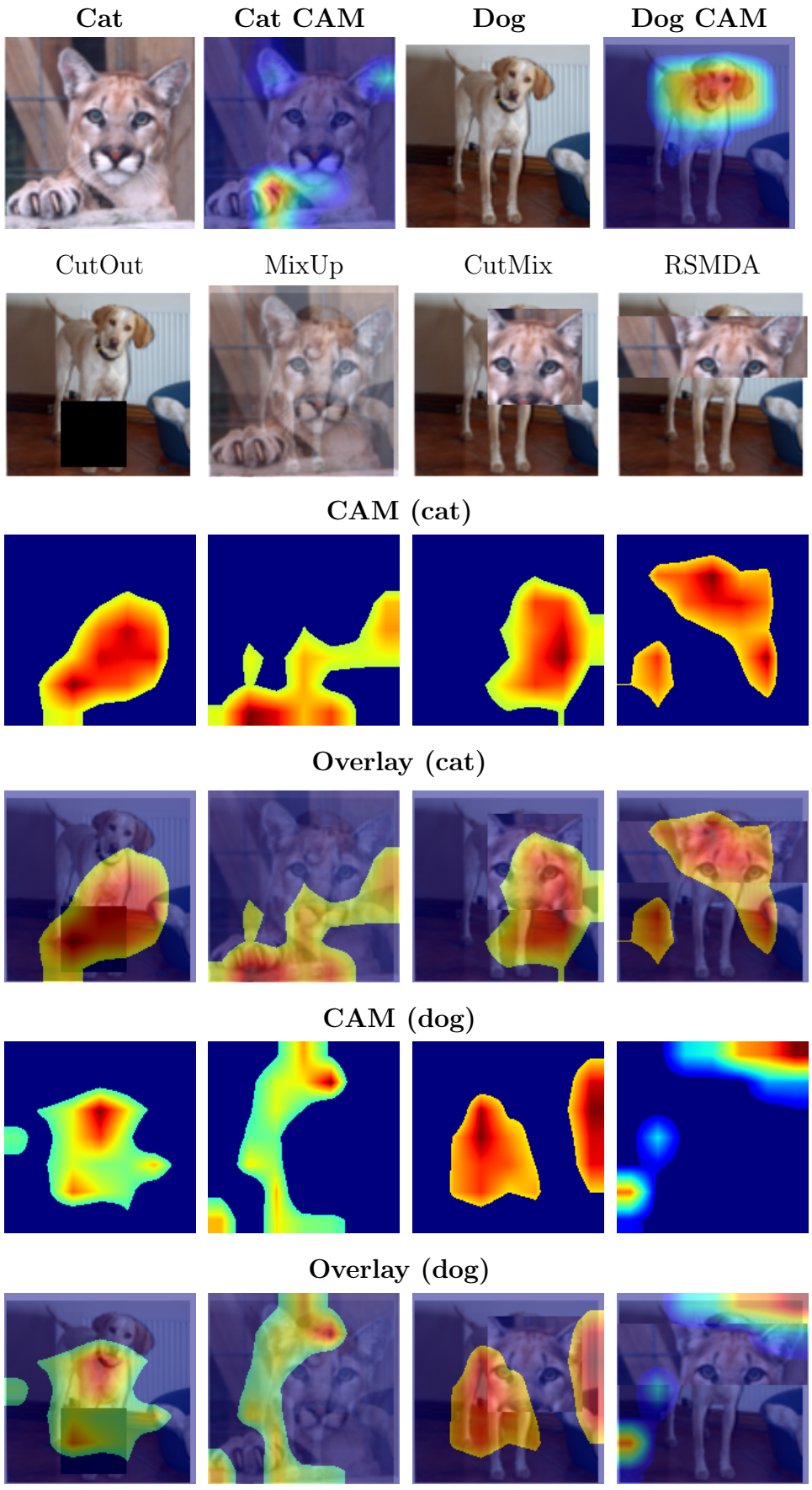


Figure 3.11: Comparison of Grad-CAM visualisations for different augmentation strategies.

output as an input for the model. Then, we use ResNet50 (`https://pytorch.org/hub/nvidia\_deeplearningexamples\_resnet50/`, accessed on 10/09/2022), a pre-trained model, to obtain the CAM for each augmented input. In the third and fourth rows of Figure 3.10, the "CAM (dog) " and the "Overlay (dog) " are shown, respectively, to clearly show where the model focuses more. The same is repeated for the cat class in the fifth and sixth rows. RSMDA suggests that it learns features that are clear to the model, i.e., the tail of the dog where the model focuses more on dog classification. We believe such tiny features are quite helpful for models to recognise objects. Among these four augmentations, CutMix has a strong capability to capture features, as shown in the third column of Figure 3.10. From the experiments, it seems that RSMDA has the capability to learn tiny or small features that are quite helpful for models in recognising objects.

To further understand how each augmentation influences feature learning, we also visualise CAMs using ResNet-50 models trained on STL10 with each augmentation strategy (CutOut, MixUp, CutMix, and RSMDA), as shown in Figure 3.11. Unlike the pre-trained ImageNet model in Figure 3.10, these CAMs reflect representations that have been optimised under the corresponding augmentation. For the cat class, RSMDA leads the model to focus on coherent and compact discriminative regions (e.g., head and torso), while CutOut and MixUp sometimes produce more fragmented attention. A similar pattern is observed for the dog class, where RSMDA maintains strong responses on semantically meaningful parts of the object, even when large regions are occluded. Overall, these results suggest that training with RSMDA encourages models to learn more robust and concentrated object-centric features than the other augmentation strategies.

Figure 3.12 compares Grad-CAM visualisations for the dog and cat images using models trained with different augmentation strategies. For the models trained with Cutout, Mixup, the high-activation regions tend to be either fragmented or partly spread into the background (e.g. around the radiator and floor for the dog image), indicating that the classifiers rely on a mixture of object and context cues. In

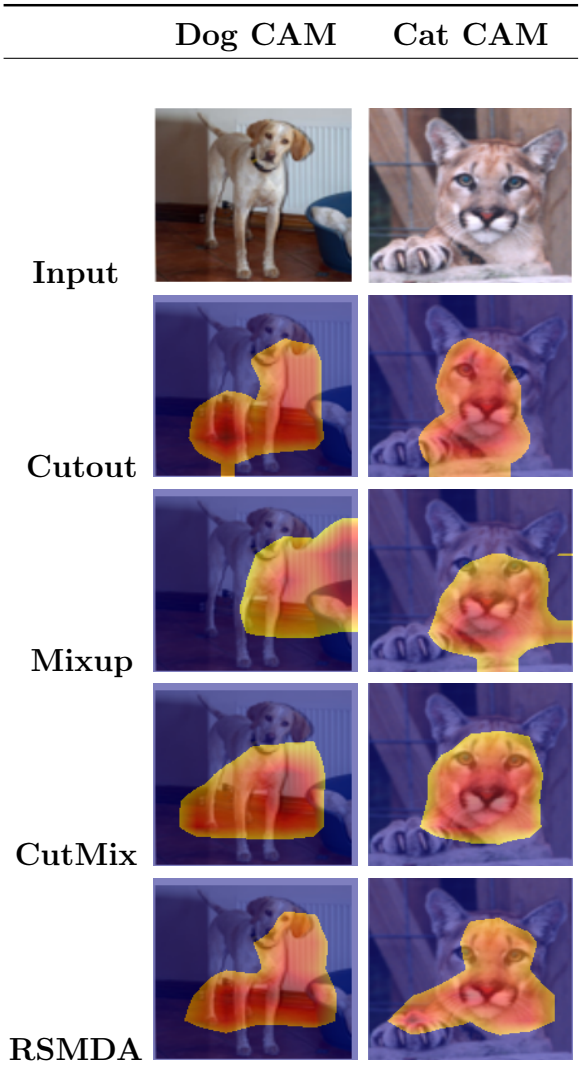


Figure 3.12: Grad-CAM visualizations for the original images and models trained with different augmentations.

contrast, the model trained with the proposed RSMDA focuses more consistently on semantically meaningful regions of the target objects, such as the head and upper body of both the dog and the cat. This suggests that RSMDA encourages the network to learn features that are more object-centric and discriminative.

RSMDA on Images of Different Resolutions

Analytically, the proposed formulation does not apply a larger mixing rate to low-resolution images: for a fixed choice of λ and Beta prior, the expected mixed pixel fraction is the same across resolutions. To illustrate this behaviour, we provide qualitative examples of RSMDA applied to CIFAR-10 (32×32), STL-10 (96×96) and Oxford-IIIT Pet (224×224) in Figure. A.1, A.2 and A.3, respectively, using the same λ and sampling strategy. In later experiments (on higher-resolution inputs), we use the same RSMDA configuration to empirically assess the impact on model performance at high resolution 224×224, in line with the standard ImageNet images.

3.4 Conclusion

In this chapter, we addressed Research Question 1 by proposing a novel data augmentation technique, Random Slices Mixing Data Augmentation (RSMDA), to mitigate feature loss in single-image data augmentation methods. RSMDA introduces a unique sliced mixing approach, in which two images are combined in a structured manner. We explored three variants of RSMDA: row-wise (horizontal), column-wise (vertical), and row-column-wise based on binary randomness. Among these, RSMDA row-wise demonstrated superior performance in terms of both accuracy and robustness.

We also identified the optimal parameters for RSMDA, including slice size and probability, and validated its effectiveness across multiple datasets and models. Results indicate that RSMDA outperforms both single- and multi-image augmentation techniques. Furthermore, robustness analysis showed that RSMDA enhances model

resilience compared to baseline methods and random erasing. CAM visualisations revealed that models trained with RSMDA focus on fine-grained and meaningful features, which are crucial for object recognition.

Moving forward, we aim to explore alternative slice shapes, such as triangular, circular, and elliptical, instead of only rectangular divisions. Additionally, we plan to extend this approach by applying sliced mixing specifically to salient image regions, further refining the augmentation process.

This chapter proposed RSMDA to recover features that single-image erasing loses and that standard multi-image mixing doesn't structure. Its findings (fine-grained feature focus, stronger robustness) set up the next step: use saliency to decide what to erase and where, so we can preserve essential content while still breaking shortcuts in Chapter 4.

Chapter 4

Robust Saliency-driven Erasing: Reducing Overfitting while Maintaining Contextual Relevance

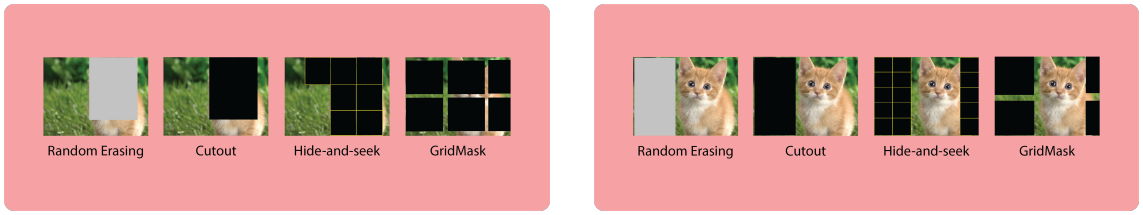
4.1 Introduction and Motivation

Building on the findings of Chapter 3, which introduced Random Slices Mixing Data Augmentation (RSMDA) to recover information lost by single-image erasing and showed that structured slice-based mixing encourages fine-grained, robust feature learning, this chapter moves from *structure mixing* to *how to decide what to erase and where*. While RSMDA operates on geometric slices and treats all pixels within a slice equally, the analysis in Chapter 3 suggested that further gains could be achieved by targeting augmentation to the most informative regions of an image.

This chapter therefore focuses on combining saliency and image erasing to design a new data augmentation strategy that can maximise the trade-off between feature loss and contextual information loss in data augmentation. This is the key issue addressed by RQ2 (1.3.2). As part of this chapter, we discuss how important features can be preserved even in occluded images, and demonstrate how contextual information is retained while mitigating overfitting. Before delving into the methodology

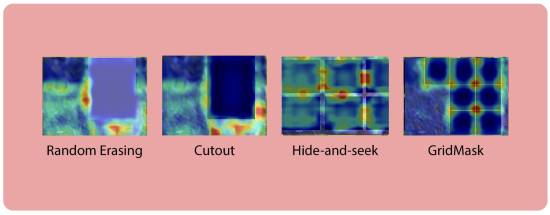
and experimental evaluation of the proposed approach, in the remainder of this Section we elaborate on the motivation that led from RSMDA to a saliency-guided erasing strategy.

Motivation. When it comes to data augmentation strategies and their effectiveness in improving the performance of deep learning models, it is clear that information-erasing data augmentation techniques promote diversity by providing occlusion perspectives in different ways. However, there are high chances of either completely erasing targeted objects (as illustrated in Figure 4.1a), leading to noisy image data (as shown in the corresponding class activation maps in Figure 4.2a), or erasing contextual information (as shown in Figure 4.1b) which might force the model to learn only the most important information (as indicated in the corresponding class activation maps in Figure 4.2b) which in turns results into overfitting. In order to find a better trade-off between these erasing issues, and give occlusion perspective at the same time, we propose a simple yet effective data augmentation strategy called RandSaliencyAug, which detects salient region in the image and applies any of the selected erasing strategies (Row Slice Erasing, Column Slice erasing, Row-Column Saliency Erasing, Partial Saliency Erasing, Horizontal Half Saliency Erasing and Vertical Half Saliency Erasing) from the search space either randomly or based on the performance of the model. The proposed approach neither removes the complete object in images like RE and Cutout nor masks the number of squares like HaS and GM. The overall process of the RandSaliencyAug approach is shown in Figure 4.3.

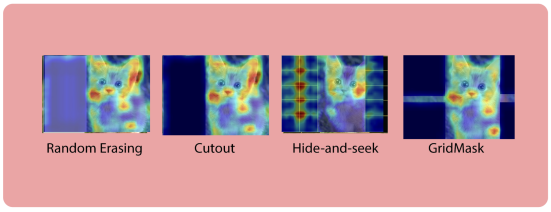


(a) Is data helpful for model? (b) Is data helpful for model generalisation?

Figure 4.1: Can we trade-off between complete object erasing and non-object erasing?



(a) Class activation map of complete object erasing



(b) Class activation map visualisation of erasing less important information

Figure 4.2: Class activation map visualisation of complete object erasing and erasing less important information.

4.2 Methodology

In this section, we present six proposed data augmentation strategies and the overall approach that is based on these six strategies.

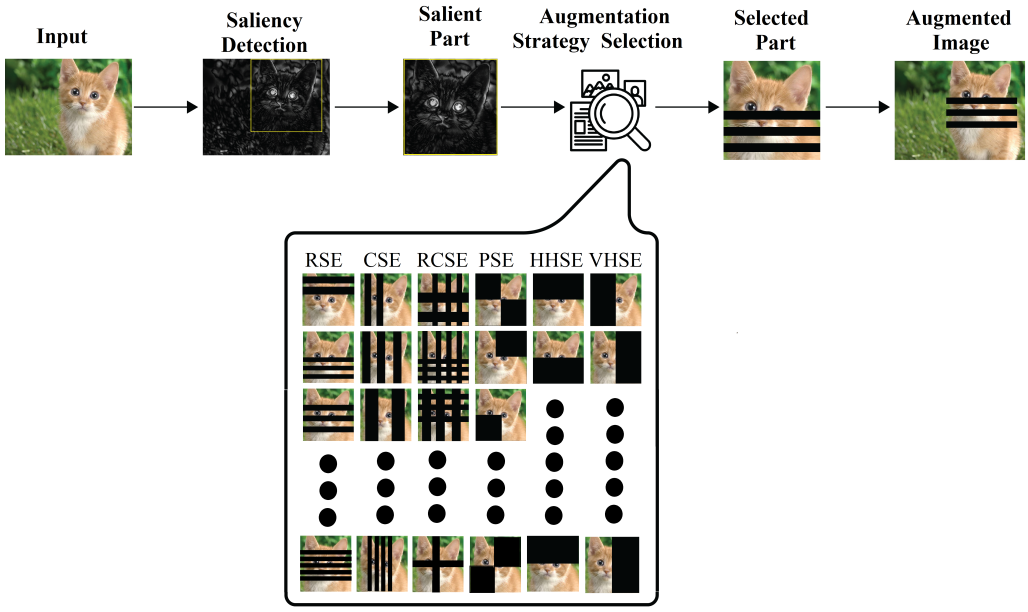


Figure 4.3: RandSaliencyAug: Proposed approach to balance between complete object erasing and contextual information erasing, where RSE, CSE, RCSE, PSE, HHSE and VHSE represent row slice erasing, column slice erasing, row-column saliency erasing, partial saliency erasing, horizontal half saliency erasing and vertical half saliency erasing, respectively.

4.2.1 Approaches in the Search Space

The search space consists of the six proposed data augmentation approaches, each using one of six erasing strategies. Each of the six erasing strategies is discussed in the remainder of this section.

It is important to note that, as illustated in Figure 4.3, all these augmentation approaches receive as input the salient region of the image and then are applied. The salient region is detected using the method in [74, 29].

Row Slice Erasing (RSE)

In this strategy, the salient region $x \in \mathbf{R}^{W \times H \times C}$ of the image I is given. The augmented salient part can be defined as:

$$\tilde{x} = M \odot x \tag{4.1}$$

where binary mask M, is a matrix of size $W \times H$, where each element of the matrix takes on a value of either 0 or 1. A value of 0 indicates that the corresponding pixel in the image should be excluded, while a value of 1 indicates that the pixel should be included. The symbol $\odot$ shows element-wise multiplication

In order to sample the binary mask M, we randomly select slices of size S from a predetermined range. The total number of slices required is determined by dividing the height H of the binary mask by the slice size S given by the below equation 4.2:

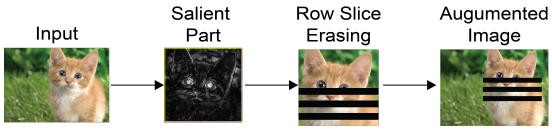
$$TotalSlices = \lfloor H/S \rfloor \tag{4.2}$$

Alternative horizontal slices of M are filled with 0's and 1's. Moreover, a demonstration of Row Slice Erasing (RSE) is given in Figure 4.4a.

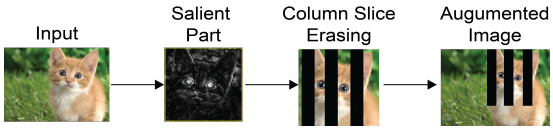
Column Slice Erasing (CSE)

In this strategy, we perform all the defined steps in the RSE 4.2.1 except, the total number of slices is calculated by dividing the width W of the binary mask by the slice size S as shown in the equation 4.3.

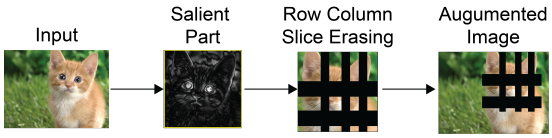
$$TotalSlices = \lfloor W/S \rfloor \tag{4.3}$$



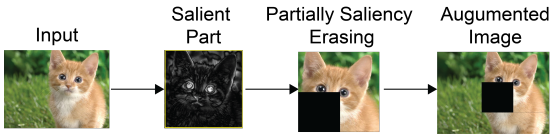
(a) Sample of Row Slice Erasing (RSE) visualisation



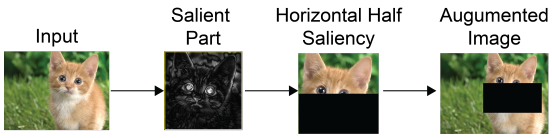
(b) Sample of Column Slice Erasing (CSE) visualisation



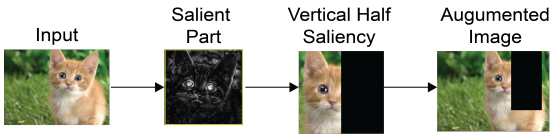
(c) Sample of Row Column Slice Erasing (RCSE) visualisation



(d) Sample of Partially Slice Erasing (PSE) visualisation



(e) Sample of Horizontal Half Slice Erasing (HHSE) visualisation



(f) Sample of Vertical Half Slice Erasing (VHSE) visualisation

Figure 4.4: Samples of the proposed augmentation strategies used in the search space

Alternative vertical slices of M are filled with 0's and 1's. Column slice erasing is illustrated in Figure 4.4b.

Row Column Slice Erasing (RCSE)

Row Column Slice Erasing is a combination of RSE 4.2.1 and CSE 4.2.1. RSE and CSE are performed in sequential order. RCSE is shown in Figure 4.4c.

Partially Saliency Erasing (PSE)

In this strategy, the salient region is divided into four parts, then a randomly random number of square(s) are erased, as shown in Figure 4.4d. Mathematically, mask M is divided into four parts of equal size and each part is filled with either 0's or 1's. One part or diagonally two parts are filled with any of 0's or 1's randomly. Then element-wise multiplication is performed on the salient part to generate augmented image $\tilde{x}$ as shown in equation 4.1.

Horizontal Half Saliency Erasing (HHSE)

In this strategy, the salient region is horizontally divided into two parts. One of them is randomly erased as demonstrated in the Figure 4.4e. The mask M is partitioned horizontally into two equal-sized segments, with one segment filled with 0's and the other with 1's. This partitioning allows for the creation of the augmented image $\tilde{x}$ by performing an element-wise multiplication on the salient part of $\tilde{x}$, as shown in equation 4.1.

Vertical Half Saliency Erasing (VHSE)

Similarly to HHSE, the salient region is vertically divided into two parts. One of them is randomly erased as shown in the Figure 4.4f. The mask M is divided vertically into two equal parts of the same size. One part is filled with 0's and the other with 1's, according to mathematical principles. To create the augmented image $\tilde{x}$, the salient part of $\tilde{x}$ is obtained through an element-wise multiplication process, as described in equation 4.1.

4.2.2 RandSaliencyAug

RandSaliencyAug data augmentation selects one data augmentation from the search space comprised of the six erasing strategies proposed above in Section 4.2.1, and it applies one of two versions of RandSaliencyAug as defined below.

Weighted RandSaliencyAug (W-RSA): in the weighted approach, the RandSaliencyAug method assigns weights to each data augmentation technique based on their performance in terms of accuracy after training. Specifically, the weight of a particular augmentation is calculated as follows. First the baseline accuracy (A_b) and the accuracy with the augmentation (A_{a_i}) are calculated. Then difference d_i is calculated by subtracting A_b from A_{a_i}.

$$d_i = |A_{a_i} - A_b| \tag{4.4}$$

Then the sum of all differences is calculated:

$$D = \sum_{i=1}^{n} d_i \tag{4.5}$$

where n is the number of data augmentation strategies, in our case, it is 6. Then the weight of the augmentation is calculated:

$$w_i = d_i / D \tag{4.6}$$

These weights are used as probabilities to randomly select an augmentation from the given list of augmentations.

Non-Weighted RandSaliencyAug (N-RSA): in non-weighted RandSaliencyAug, weights are uniformly calculated irrespective of accuracy performance. Weights are calculated as follow:

$$w_i = 1/n \tag{4.7}$$

where n is the number of data augmentation strategies, which is 6 in our case.

After calculating the weights, these weights are used as probabilities with an equal chance of selecting each data augmentation from the list in the search space. The process is similar to the one in RandAug [52].

4.3 Experiments

In this section, we discuss the experimental we have conducted to validate our apporach, including details of the training setup and results.

4.3.1 Training set up

Image Classification

For image classification tasks, we follow the same training setup as in the RE [18] for a fair comparison using Fashion-MNIST [64], CIFAR10 and CIFAR100 [62] datasets

except stated the baseline results. We used 300 epochs, 128 batch size, 0.1 initial learning rate, and the learning rate is divided by 10 at 90, 180, and 240 epochs in multistep learning rate. For Table 4.5, CIFAR-10 dataset, models were trained for 80 epochs, while for CIFAR-100, they were trained for 120 epochs [75]. In both cases, the batch size was 32, the learning rate was 1e-3, weights were initialised using the Xavier Normal technique, and weight decay was set at 1e-5. For the ImageNet [13] dataset, we follow the same training setup as in the gridmask [25]. We use different neural network architectures such as ResNet flavors [1], wide-ResNet-28-10 [76] and shake-shake-26-32 [77], along with different state-of-the-art data augmentation approaches. For TinyImageNet [78] dataset, we used the same training setting as in KeepOriginalAugment [79]. For image classification , we use accuracy and error rate as performance measurement matrices, respectively. The higher the accuracy, the better and the lower the error rate, the better.

Object Detection

For object detection tasks, we used PASCAL VOC 2007 [80] dataset and we follow the same training setup as in the fasterRCNN [81, 18]. We use VGG16 architecture as the backbone and apply SGD for 80K to train all models. The learning rate starts at 0.001 and falls to 0.0001 at 60K iterations. For object detection, we use mean average precision (mAP) as performance measurement matrices. A higher mAP is better.

4.3.2 Hyperparamter Study

We first determine the optimal probability for each of the six augmentation strategies. The probability [18] refers to the likelihood of applying a specific data augmentation. These probabilities are optimised using ResNet-18 on CIFAR-10. We use the ResNet-18 architecture because it is relatively small, allowing us to run a large number of experiments quickly and applied across all subsequent experiments, as detailed below.

Table 4.1: Difference of each strategy with baseline, where A% and Δ represent accuracy and accuracy difference, respectively.

Strategy	**Baseline**	**Strategy A%**	**Δ**	**Weight**
PSE	95.28	96.48	1.20	0.24
HHSE	95.28	96.31	1.03	0.21
RSE	95.28	96.02	0.74	0.15
CSE	95.28	95.70	0.42	0.08
RCSE	95.28	95.90	0.62	0.12
VHSE	95.28	96.27	0.99	0.20

Table 4.2: Accuracy (%) of each saliency-based erasing strategy at different erasing probabilities. RSE, CSE, RCSE, PSE, HHSE and VHSE denote row slice erasing, column slice erasing, row–column saliency erasing, partial saliency erasing, horizontal half saliency erasing and vertical half saliency erasing, respectively.

Probability	**RSE**	**PSE**	**VHSE**	**HHSE**	**RCSE**	**CSE**
0.1	95.52	95.62	95.77	95.60	95.62	95.66
0.2	95.48	95.77	95.78	96.04	95.53	95.54
0.3	95.92	95.68	95.96	95.97	95.51	95.23
0.4	95.75	95.95	96.11	96.28	95.48	95.52
0.5	95.82	96.17	96.27	96.31	95.90	95.70
0.6	95.79	95.86	95.76	96.05	95.44	95.67
0.7	95.86	96.48	96.26	96.18	95.46	95.61
0.8	96.02	96.15	96.07	96.00	95.61	95.69
0.9	95.73	96.09	95.99	96.12	95.46	95.64
1.0	95.30	96.01	96.20	95.90	95.83	95.56

Finding the best probability

When providing augmented examples to a model, there is a concern that the model may not have access to the original data, potentially causing a shift in the data distribution. To address this, a well-known technique involves finding a balance between the augmented and original samples [18, 27, 82, 14]. This is typically achieved by assigning probabilities to each data augmentation technique in the search space. In our study, we propose and utilise six data augmentation strategies. But before deploying these data augmentation in the search space, it is important to find the best probability for each data augmentation strategy. To do so, we use a probability ranging from 0.1 to 1.0 with an interval of 0.1 for each strategy as shown in Table 4.2. After that, we record the best probability and fix that in the search space.

Regarding predetermined ranges of slice size for RSE and CSE, it depends on

height and width of the image like RSE range is from 1 to H/2, it means maximum slice size is H/2. In case of CSE, it ranges from 1 to W/2, Moreover, it is important to note, on each epoch of training, slice size is continuously changing and randomly selected. As RCSE is mixture of RSE and CSE, so its range is same as of RSE and CASE. Other three strategies (PSE, HHSE, VHSE) have no such ranges.

Calculating the performance-based weights

We propose and investigate two variations of our approach: Weighted RandSaliencyAug (W-RSA) and Non-Weighted RandSaliencyAug (N-RSA). In W-RSA, we determine weights based on performance. To accomplish this, we compute the difference between the performance of each strategy and the baseline performance, as outlined in Table 4.1. It is important to note that these performance calculations are achieved using ResNet18 model on the CIFAR10 dataset. Next, we calculate the sum of all the differences, which amounts to a total of 5.0. We then proceed to divide each strategy's difference by the sum of all the differences, as presented in Table 4.1. This step allows us to obtain the final weights for each strategy. The weights presented in Table 4.1 are specifically employed in the Weighted RandSaliencyAug (W-RSA) approach for training the ResNet18 model on the CIFAR10 dataset. These best weights are used for all the remaining experiments except fashion-MNIST and CIFAR100. On the other hand, in the Non-Weighted RandSaliencyAug (N-RSA) approach, the weights are considered uniform, following the principle of 1/n, where n represents the number of data augmentation strategies. In our case, since there are 6 augmentation strategies, each strategy is assigned a weight of 1/6. This process aligns with the methodology described in [52].

Why CIFAR10 weights for other datasets? Furthermore, we computed the weights for CIFAR-100, in the same way as illustrated for CIFAR-10. Moreover, we conducted a comparative analysis of the accuracy of various models on CIFAR-100 using both CIFAR-100 and CIFAR-10 weights, as depicted in Table 4.3. Overall, our findings suggest that CIFAR-10 weights are applicable across different datasets,

leading to improved accuracy performance.

Table 4.3: Accuracy performance comparison of the proposed approaches with existing methods on CIFAR-10 and CIFAR-100 datasets. The best performance is highlighted in blue.

CIFAR-10 Results					
Methods	**ResNet-32**	**ResNet-44**	**ResNet-56**	**WRNet-28-10**	**S-S-26-32**
Baseline	93.59 ± 0.06	94.47 ± 0.08	94.69 ± 0.07	96.13	96.42
Dropout [17]	-	-	93.74 ± 0.11	-	-
S-Dropout [83]	-	-	94.28 ± 0.20	-	-
DropBlock [84]	-	-	94.01 ± 0.08	-	-
F-Dropout [83]	-	-	94.67 ± 0.07	-	-
RE	94.34 ± 0.10	94.87 ± 0.09	95.11 ± 0.07	96.92	96.46
HaS	-	94.97 ± 0.00	**95.41 ± 0.00**	96.94	96.89
GM	94.60 ± 0.12	94.72 ± 0.13	95.11 ± 0.21	97.24	97.20
AutoAugment [85]	-	-	-	97.01	96.96
GridMask [25]	-	-	-	97.24	97.20
AugMix [86]	-	-	-	97.5	97.5
Attentive CutMix [75]	-	-	-	97.3	97.4
KeepAugment [31]	-	-	-	**97.8**	**97.9**
N-RSA (ours)	94.65 ± 0.21	94.69 ± 0.05	95.14 ± 0.10	96.96	96.76
W-RSA (ours)	**94.69 ± 0.21**	**94.96 ± 0.12**	95.22 ± 0.11	96.98	96.73
CIFAR-100 Results					
Methods	**ResNet-32**	**ResNet-44**	**ResNet-56**	**WRNet-28-10**	**S-S-26-32**
Baseline	71.50 ± 0.37	74.73 ± 0.21	73.71 ± 0.28	-	-
Dropout [17]	-	-	73.81 ± 0.27	-	-
S-Dropout [83]	-	-	74.97 ± 0.16	-	-
DropBlock [84]	-	-	73.92 ± 0.10	-	-
F-Dropout [83]	-	-	74.97 ± 0.16	-	-
RE	72.82 ± 0.32	75.71 ± 0.16	76.31 ± 0.33	-	-
HaS	-	75.82 ± 0.00	76.47 ± 0.00	-	-
GM	72.01 ± 0.12	75.13 ± 0.12	75.43 ± 0.21	-	-
N-RSA (ours)	73.37 ± 0.20	75.82 ± 0.21	76.84 ± 0.10	-	-
W-RSA (ours)	**73.74 ± 0.10**	76.32 ± 0.17	77.64 ± 0.77	-	-
W-RSA (ours - CIFAR10 weights)	-	**76.53 ± 0.11**	**77.81 ± 0.71**	-	-

4.3.3 Image Classification Results

We evaluate the effectiveness of the proposed approach on fashion-MNIST, CIFAR10, CIFAR100, TinyImageNet and ImageNet. For the fashion-MNIST dataset, we used various CNN architecture, and the proposed approach especially, W-RSA has outperformed all the existing relevant approaches.

FashionMNIST

Across all CNNs, W-RSA shows an improvement of absolute 2% over the baseline while it outperformed the existing methods by a slight margin as shown in Table 4.4. The reported standard deviations in Table 4.4 are computed by training each model configuration over five independent runs, each with a different random seed. For each method and architecture, we calculate the mean accuracy and its corresponding standard deviation. **Interpretation of standard deviations** as the larger standard deviations indicate that a method is more sensitive to random initialization as defined in RE [18], and different data augmentations along with W-RSA. Methods with small deviations exhibit more stable and consistent performance across runs. For instance, methods such as RSMDA(RC) and W-RSA show

Table 4.4: Accuracy performance comparison of the proposed approaches with the existing and relevant approaches on fashionMNIST. Highlighted blue is the best performance

Methods	**ResNet20**	**ResNet32**	**ResNet44**	**ResNet56**
Fashion-MNIST				
Baseline	93.79± 0.11	93.96± 0.13	93.92± 0.16	93.22± 0.16
RE	94.96± 0.10	95.15± 0.12	95.13± 0.10	94.98± 0.11
RSMDA(R)	95.09± 0.12	95.19± 0.17	**95.93± 0.14**	95.00± 0.19
RSMDA(C)	95.28± 0.13	95.35± 0.15	95.22± 0.01	95.00± 0.20
RSMDA(RC)	95.24± 0.06	95.19± 0.12	95.10± 0.25	94.91± 0.59
RSE(ours)	94.02 ± 1.35	94.34 ± 0.90	94.69 ± 0.81	94.72 ± 1.05
CSE(ours)	94.02 ± 1.35	94.66 ± 0.47	95.03 ± 0.32	94.45 ± 0.62
RCSE(ours)	94.63 ± 0.55	94.68 ± 0.60	94.57 ± 0.92	94.62 ± 0.65
HHSE(ours)	94.65 ± 0.79	95.34 ± 0.13	95.49 ± 0.16	**95.24 ± 0.74**
VHSE(ours)	94.52 ± 0.72	94.49 ± 0.78	94.31 ± 1.03	94.43 ± 0.97
PSE(ours)	94.71 ± 0.66	94.62 ± 0.69	94.68 ± 0.73	94.57 ± 0.73
N-RSA (ours)	95.34 ± 0.59	95.36 ± 0.16	95.34 ± 0.06	95.01 ± 0.12
W-RSA(ours)	**95.35 ± 0.12**	**95.37 ± 0.06**	95.27 ± 0.20	95.18 ± 0.06

low variance, suggesting strong stability, whereas higher deviations observed in RSE and CSE indicate that their training outcomes vary more across runs, which may point to sensitivity in the optimization process or weaker regularization behavior.

Table 4.5: Accuracy performance comparison with saliency and image mixing based augmentation methods, where CutMix (Att:) refers to Attentive CutMix and gain is gain over baseline.

CIFAR10 (%)						
Method	**ResNet-18**	**ResNet-34**	**ResNet-50**	**DenseNet-121**	**DenseNet-169**	**EfficientNet - B0**
Baseline	84.67	87.12	95.02	85.65	87.67	87.45
Mixup	88.52	88.70	-	87.56	89.12	88.07
CutMix	87.92	88.75	90.84	87.98	89.23	88.67
CutMix (Att:)	88.94	90.40	-	88.34	90.45	88.94
SaliencyMix	96.53	-	93.19	-	-	-
PuzzleMix	97.10	-	-	-	-	-
CoMixup	97.15	-	-	-	-	-
AutoMix	**97.34**	-	-	-	-	-
Ours(N-RSA)	91.12	90.03	96.31	91.39	**92.09**	88.88
Ours (W-RSA)	91.61	**90.43**	**96.33**	**91.74**	91.81	**88.98**
Gain	+6.94	+3.31	+1.31	+6.09	+4.14	+1.53
CIFAR100 (%)						
Baseline	63.14	65.54	63.52	65.12	66.42	75.67
Mixup	64.40	67.83	-	66.84	68.24	77.21
CutMix	65.90	68.32	68.35	67.62	69.58	77.57
CutMix (Att:)	67.16	70.03	-	**69.23**	**71.34**	**78.52**
SaliencyMix	79.12	-	75.11	-	-	-
PuzzleMix	81.13	-	-	-	-	-
CoMixup	81.17	-	-	-	-	-
AutoMix	**82.04**	-	-	-	-	-
Ours(N-RSA)	68.02	70.08	69.91	67.84	69.21	77.45
Ours (W-RSA)	68.01	**70.51**	69.99	67.94	70.60	78.01
Gain	+4.87	+4.97	+6.47	+2.82	+4.18	+2.34

Table 4.6: Accuracy of each strategy over CIFAR10 and CIFAR100 using different models. This accuracy acts as weights in the proposed strategy.

Methods	**CIFAR10**			**CIFAR100**		
	Resnet34	**Resnet44**	**Resnet56**	**Resnet34**	**Resnet44**	**Resnet56**
Basline	93.59 ± 0.06	94.47 ± 0.08	94.69 ± 0.07	71.50 ± 0.37	74.73 ± 0.21	73.71 ± 0.28
RSE(ours)	94.21 ± 0.21	94.78 ± 0.18	95.12 ± 0.02	72.57 ± 0.40	75.92 ± 0.142	76.28 ± 0.18
CSE(ours)	94.01 ± 0.21	94.49 ± 0.17	94.84 ± 0.04	73.37 ± 0.20	75.54 ± 0.21	75.70 ± 0.28
RCSE(ours)	94.12 ± 0.32	94.7 ± 0.18	94.81 ± 0.21	72.30 ± 0.15	75.88 ± 0.27	76.21 ± 0.21
HHSE(ours)	94.45 ± 0.23	95.03 ± 0.11	95.11 ± 0.03	73.83 ± 0.19	76.02 ± 0.28	76.58 ± 0.29
VHSE(ours)	94.21± 0.17	94.53 ± 0.21	95.17 ± 0.31	73.37 ± 0.08	75.28 ± 0.15	75.95 ± 1.13
PSE(ours)	94.32 ± 0.18	94.95 ± 0.01	95.13 ± 0.12	72.30 ± 0.08	75.65 ± 0.33	76.19 ± 0.30

CIFAR datasets results

Similarly for CIFAR10 and CIFAR100 datasets, the proposed approach W-RSA outperforms all CNN architectures except Wide-ResNet and shake-shake-26-32 (S-S-26-32) architecture. Though W-RSA showed competitive performance on wide-ResNet-28-10 and S-S-26-32 architectures. Overall, W-RSA showed the best performance and our proposed approaches (Row Slice Erasing, Column Slice erasing, Row-Column Saliency Erasing, Partial Saliency Erasing, Horizontal Half Saliency Erasing and Vertical Half Saliency Erasing) also showed competitive performance on CIFAR10 and CIFAR100 as shown in Table 4.3. For Table 4.3, required weights are calculated from Table 4.6, which acts as probability in search space of W-RSA. Furthermore, we compare both flavours of the proposed approach with saliency based and mixing data augmentation. The proposed approach almost outperformed in most cases and showed competitive performance as shown in Table 4.5.

Table 4.7: Results on ImageNet using different network architecture and comparison with existing approaches, where Acc(%) is accuracy(%). Highlighted blue is the best performance.

Model	**Acc(%)**	**Model**	**Acc(%)**	**Model**	**Acc(%)**
ResNet50 [84, 1]	76.5	ResNet101 [27, 1]	78.0	ResNet152 [1]	78.3
+Dropout [17, 31]	76.8	+Dropout [17, 31]	77.7	-	-
+DropPath [87, 31]	77.1	-	-	-	-
+DropBlock [84, 31]	78.3	+DropBlock [84, 31]	79.0	-	-
+Cutout [26]	77.1	-	-	-	-
+HaS [24]	77.2	-	-	-	-
+Mixup [27, 31]	77.9	+Mixup [27, 31]	79.2	-	-
+AutoAugment [85]	77.6	-	-	-	-
+RandAugment [52]	77.6	+RandAugment [52]	79.2	-	-
+RandomErasing [18]	77.3	+RandomErasing [18]	79.6	-	-
+GridMask [25]	77.9	+GridMask [25]	79.1	+GridMask [25]	79.7
+AutoAugment [85]	77.6	+AutoAugment [85]	79.3	-	-
+KeepAutoAug [31]	78.0	+KeepAutoAug [31]	79.7	-	-
+SaliencyMix [29]	78.46	+SaliencyMix [29]	80.45	-	-
+FMix [88]	78.51	+FMix [88]	80.20	-	-
+PuzzleMix [89]	78.86	+PuzzleMix [89]	80.67	-	-
+AutoMix [90]	**79.25**	+AutoMix [90]	**80.98**	-	-
+N-RSA (Ours)	77.9	+N-RSA (ours)	79.2	+N-RSA (ours)	79.5
+W-RSA (Ours)	78.1	+W-RSA (Ours)	79.4	+ W-RSA (Ours)	**79.8**

For ImageNet classification, we used the same weights (probabilities) for W-RSA as used in CIFAR10 and CIFAR100, as training individual strategies on Ima-

Table 4.8: Test Error rate (%) on TinyImageNet [78] using various models.

Model	PreActResNet-18	PreActResNet-50
TinyImageNet Dataset		
Baseline	42.33 ± 0.21	38.58 ± 0.24
+Cutout [26, 30]	42.04 ± 0.31	38.36 ± 0.21
+SalfMix [30]	40.28 ± 0.28	35.92 ± 0.07
+Mixup [30, 67]	40.22 ± 0.2	35.51 ± 0.15
+SaliencyMix [29, 30]	37.76 ± 0.05	**32.83 ± 0.47**
+CutMix [14, 30]	38.11 ± 0.32	33.54 ± 0.19
+ResizeMix [30, 91]	38.47 ± 0.25	33.25 ± 0.12
KeepOriginalAugment [79]	35.1 ± 0.20	35.6 ± 0.12
N-RSA (Ours)	34.97 ± 0.10	33.10 ± 0.10
W-RSA (Ours)	**34.91 ± 0.20**	32.92 ± 0.23

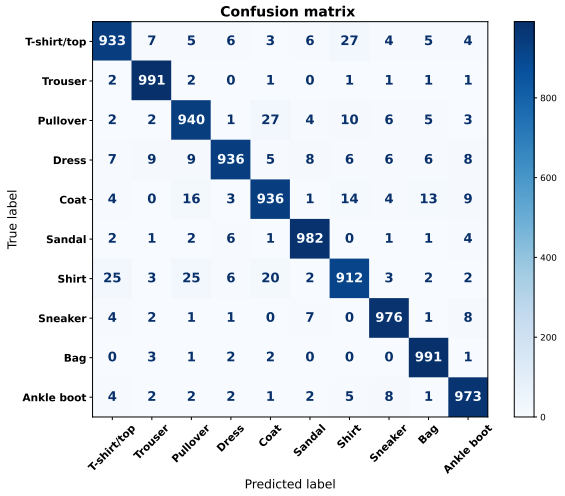
geNet requires more training time, computational resources, and cumbersome tasks. We check RSA and W-RSA using various CNN architectures. Interestingly both showed competitive performance over the existing data augmentation methods. W-RSA showed the better performance among all the approaches using resnet50 and resnet152 as shown in Table 4.7.

For tinyImagenet, we used PreActResnet18 and PreActResnet50 models, our approach has showed lower error rate as compare to other methods for PreActResnet18 while it showed the second best in the case of PreActResNet50 as shown in Table 4.8.

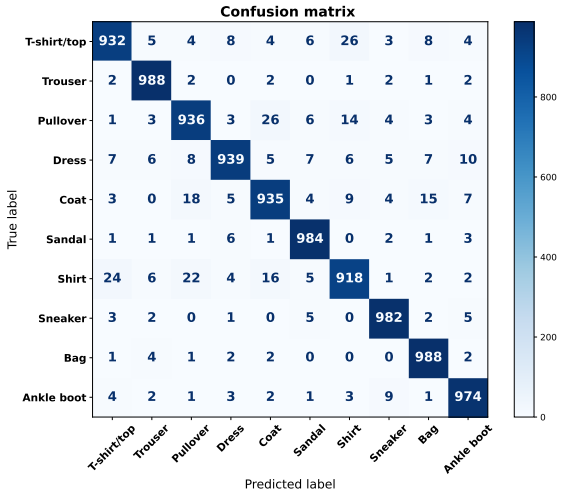
Overall, our proposed approaches specially W-RSA have shown better performance for diverse classification tasks on various datasets using different CNN architectures.

,

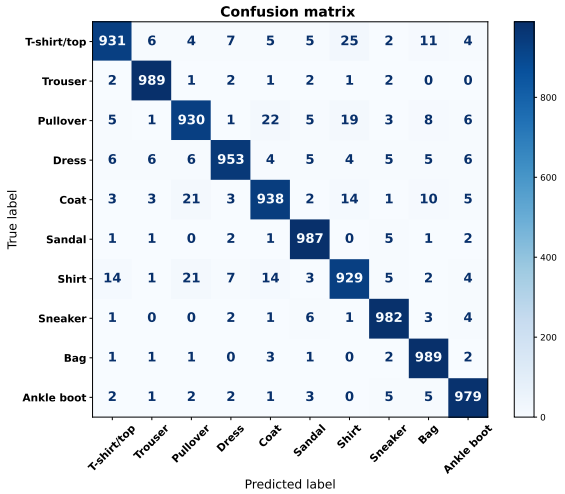
To further analyse performance across individual classes, we additionally report precision, recall, F1 score, and ROC AUC. The detailed class-wise results for the different datasets and models are presented in Table 4.9 for FashionMNIST, Tables 4.10 and 4.11 for CIFAR10, Table 4.12 for CIFAR100, and Table 4.13 for TinyImageNet. CIFAR100 and TinyImageNet have 100 and 200 classes, respectively, so it is not feasible to show all per-class results in the main text. However, we provide the full set of metrics, including confusion matrices for each class of these two datasets, at `https://github.com/kmr2017/ThesisPhDCode/tree/main/`



(a) ResNet-20 (WRSA)



(b) ResNet-32 (WRSA)

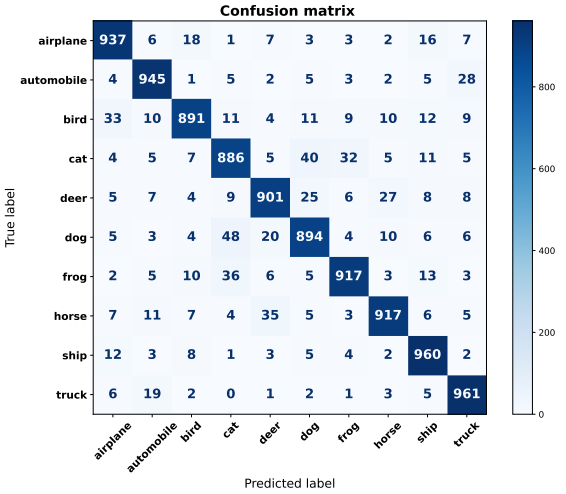


(c) ResNet-56 (HSSE)

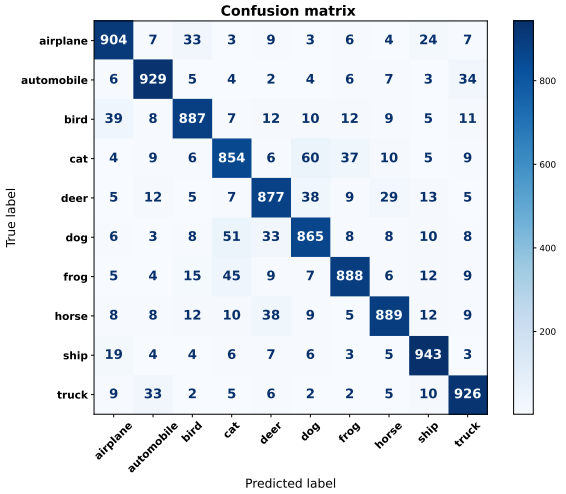
Figure 4.5: Confusion matrices on Fashion-MNIST for different ResNet models using WRSA and HSSE.

Table 4.9: Class-wise metrics on FashionMNIST for different models with RSA

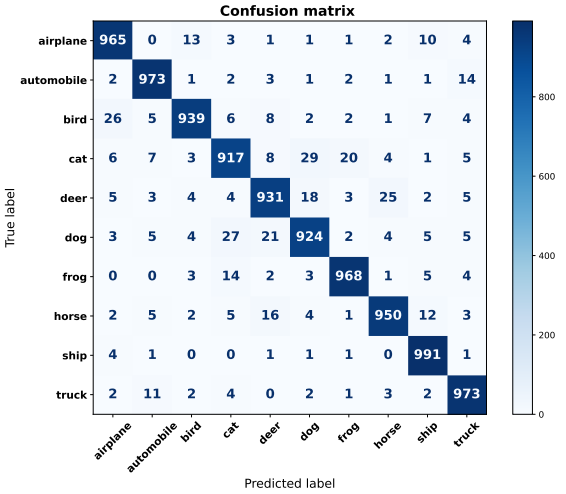
Class	Precision	Recall	F1	ROC AUC
ResNet20 with - WRSA				
T-shirt/top	0.9491	0.9330	0.9410	0.9928
Trouser	0.9716	0.9910	0.9812	0.9979
Pullover	0.9372	0.9400	0.9386	0.9951
Dress	0.9720	0.9360	0.9536	0.9938
Coat	0.9398	0.9360	0.9379	0.9938
Sandal	0.9704	0.9820	0.9761	0.9975
Shirt	0.9354	0.9120	0.9235	0.9910
Sneaker	0.9673	0.9760	0.9716	0.9957
Bag	0.9659	0.9910	0.9783	0.9974
Ankle boot	0.9605	0.9730	0.9667	0.9939
Mean	**0.9569**	**0.9570**	**0.9569**	**0.9949**
ResNet32 with - WRSA				
T-shirt/top	0.9530	0.9320	0.9424	0.9881
Trouser	0.9715	0.9880	0.9797	0.9977
Pullover	0.9426	0.9360	0.9393	0.9908
Dress	0.9670	0.9390	0.9528	0.9903
Coat	0.9416	0.9350	0.9383	0.9902
Sandal	0.9666	0.9840	0.9752	0.9974
Shirt	0.9396	0.9180	0.9287	0.9907
Sneaker	0.9704	0.9820	0.9761	0.9951
Bag	0.9611	0.9880	0.9744	0.9983
Ankle boot	0.9615	0.9740	0.9677	0.9952
Mean	**0.9575**	**0.9576**	**0.9575**	**0.9934**
ResNet56 with - HSSE				
T-shirt/top	0.9638	0.9310	0.9471	0.9922
Trouser	0.9802	0.9890	0.9846	0.9985
Pullover	0.9432	0.9300	0.9366	0.9907
Dress	0.9734	0.9530	0.9631	0.9936
Coat	0.9475	0.9380	0.9427	0.9914
Sandal	0.9686	0.9870	0.9777	0.9968
Shirt	0.9355	0.9290	0.9323	0.9899
Sneaker	0.9704	0.9820	0.9761	0.9957
Bag	0.9565	0.9890	0.9725	0.9976
Ankle boot	0.9674	0.9790	0.9732	0.9955
Mean	**0.9606**	**0.9607**	**0.9606**	**0.9942**



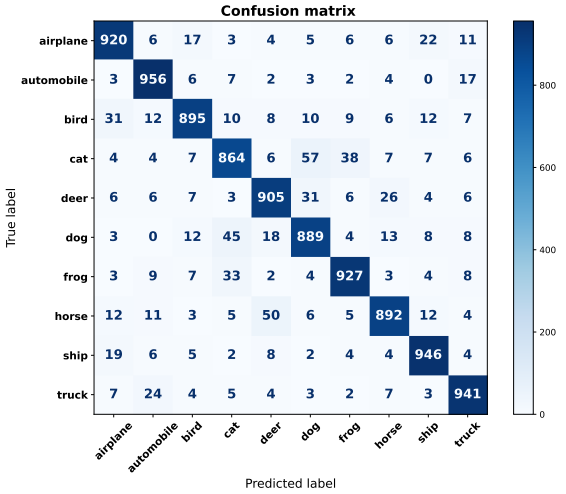
(a) DenseNet-121 (WRSA)



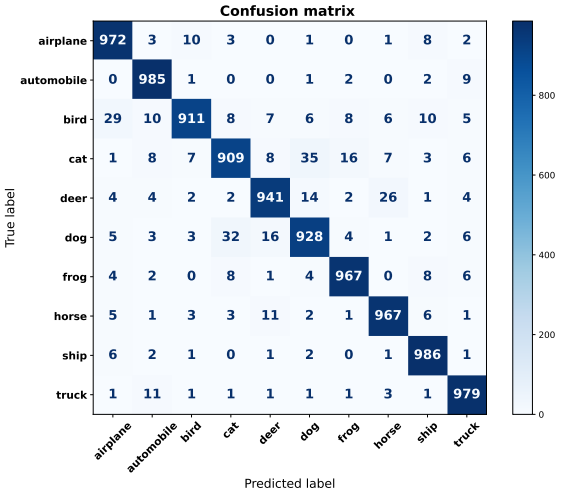
(b) EfficientNet-B0 (WRSA)



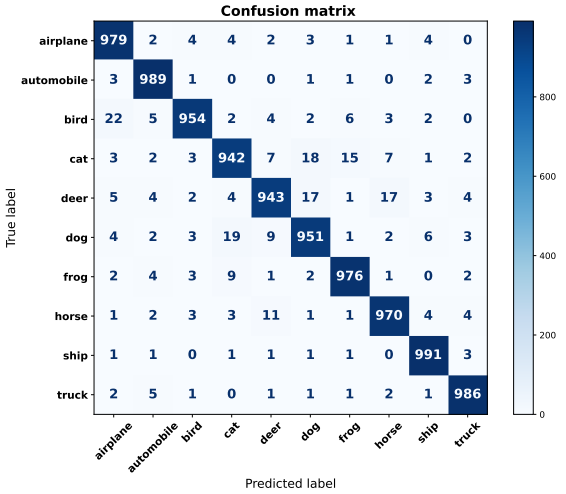
(c) ResNet-32 (WRSA)



(d) ResNet-34 (WRSA)



(e) ResNet-44 (WRSA)



(f) ResNet-50 (WRSA)

Figure 4.6: Confusion matrices on CIFAR-10 for different models using WRSA.

Table 4.10: Class-wise metrics on CIFAR10 for different models with RSA

Class	Precision	Recall	F1	ROC AUC
DenseNet121 with WRSA				
Airplane	0.9232	0.9370	0.9300	0.9927
Automobile	0.9320	0.9450	0.9384	0.9940
Bird	0.9359	0.8910	0.9129	0.9916
Cat	0.8851	0.8860	0.8856	0.9860
Deer	0.9157	0.9010	0.9083	0.9877
Dog	0.8985	0.8940	0.8962	0.9890
Frog	0.9338	0.9170	0.9253	0.9876
Horse	0.9348	0.9170	0.9258	0.9906
Ship	0.9213	0.9600	0.9403	0.9959
Truck	0.9294	0.9610	0.9449	0.9933
Mean	**0.9210**	**0.9209**	**0.9208**	**0.9908**
EfficientNetB0 with WRSA				
Airplane	0.8995	0.9040	0.9017	0.9823
Automobile	0.9135	0.9290	0.9212	0.9844
Bird	0.9079	0.8870	0.8973	0.9858
Cat	0.8609	0.8540	0.8574	0.9650
Deer	0.8779	0.8770	0.8774	0.9801
Dog	0.8616	0.8650	0.8633	0.9620
Frog	0.9098	0.8880	0.8988	0.9750
Horse	0.9146	0.8890	0.9016	0.9830
Ship	0.9094	0.9430	0.9259	0.9897
Truck	0.9070	0.9260	0.9164	0.9849
Mean	**0.8962**	**0.8962**	**0.8961**	**0.9792**

`RevisionExperiments/RQ2Files`. Moreover, we show confusion matrices for different models across these datasets using the RandSaliencyAug approach in Figures 4.5 and 4.6.

4.3.4 Object Detection

For object detection, we used PASVAL VOC 2007 dataset using VGG16 as the backbone in Faster-RCNN. We compare our approach with existing and relevant methods. Our approach W-RSA showed overall the highest mAP among all the methods. Though for individual classes, W-RSA has not shown the highest mAP but it showed consistent AP across all the classes, which indicates that our proposed is more stable and suitable for the object detection task as shown in Table 4.14. Dealing with multiple important objects in an image while keeping them relevant to the image label is a known challenge. To tackle this, our proposed method checks its effectiveness by verifying if any object is completely removed. If so, we eliminate that object's coordinate, making sure the model focuses on learning information from the remaining coordinates. To avoid the model consistently missing the erased

Table 4.11: Class-wise metrics on CIFAR10 for different ResNet models with RSA.

Class	Precision	Recall	F1	ROC AUC
ResNet32 with WRSA				
Airplane	0.9507	0.9650	0.9578	0.9935
Automobile	0.9634	0.9730	0.9682	0.9962
Bird	0.9670	0.9390	0.9528	0.9949
Cat	0.9338	0.9170	0.9253	0.9884
Deer	0.9395	0.9310	0.9352	0.9942
Dog	0.9381	0.9240	0.9310	0.9883
Frog	0.9670	0.9680	0.9675	0.9950
Horse	0.9586	0.9500	0.9543	0.9968
Ship	0.9566	0.9910	0.9735	0.9983
Truck	0.9558	0.9730	0.9643	0.9944
Mean	**0.9531**	**0.9531**	**0.9530**	**0.9940**
ResNet34 with WRSA				
Airplane	0.9127	0.9200	0.9163	0.9855
Automobile	0.9246	0.9560	0.9400	0.9914
Bird	0.9294	0.8950	0.9119	0.9805
Cat	0.8843	0.8640	0.8741	0.9637
Deer	0.8987	0.9050	0.9018	0.9810
Dog	0.8802	0.8890	0.8846	0.9626
Frog	0.9242	0.9270	0.9256	0.9802
Horse	0.9215	0.8920	0.9065	0.9811
Ship	0.9293	0.9460	0.9376	0.9919
Truck	0.9298	0.9410	0.9354	0.9922
Mean	**0.9135**	**0.9135**	**0.9134**	**0.9810**
ResNet44 with WRSA				
Airplane	0.9464	0.9720	0.9591	0.9957
Automobile	0.9572	0.9850	0.9709	0.9977
Bird	0.9702	0.9110	0.9397	0.9903
Cat	0.9410	0.9090	0.9247	0.9796
Deer	0.9544	0.9410	0.9476	0.9912
Dog	0.9336	0.9280	0.9308	0.9802
Frog	0.9660	0.9670	0.9665	0.9926
Horse	0.9555	0.9670	0.9612	0.9953
Ship	0.9601	0.9860	0.9729	0.9977
Truck	0.9607	0.9790	0.9698	0.9960
Mean	**0.9545**	**0.9545**	**0.9543**	**0.9916**
ResNet50 with WRSA				
Airplane	0.9579	0.9790	0.9683	0.9984
Automobile	0.9734	0.9890	0.9812	0.9994
Bird	0.9795	0.9540	0.9666	0.9980
Cat	0.9573	0.9420	0.9496	0.9953
Deer	0.9632	0.9430	0.9530	0.9974
Dog	0.9539	0.9510	0.9524	0.9979
Frog	0.9721	0.9760	0.9741	0.9986
Horse	0.9671	0.9700	0.9685	0.9989
Ship	0.9773	0.9910	0.9841	0.9992
Truck	0.9791	0.9860	0.9826	0.9981
Mean	**0.9681**	**0.9681**	**0.9680**	**0.9981**

Table 4.12: Mean of metrics for different models on CIFAR100.

Class	Precision	Recall	F1	ROC AUC
ResNet32 with - WRSA				
Mean	**0.746**	**0.745**	**0.745**	**0.943**
ResNet34 with - WRSA				
Mean	**0.721**	**0.719**	**0.719**	**0.934**
ResNet44 with - WRSA				
Mean	**0.771**	**0.771**	**0.770**	**0.945**
ResNet56 with - WRSA				
Mean	**0.783**	**0.783**	**0.782**	**0.947**

Table 4.13: Mean of metrics on TinyImageNet.

Class	Precision	Recall	F1	ROC AUC
PreActResNet18 with WRSA				
Mean	**0.658**	**0.655**	**0.655**	**0.908**

object, we introduce a 0.5 probability of applying augmentation. This way, the model has exposure to both original and augmented examples for diversity. When an object is only partially erased, we investigate whether our method can still learn hidden important information.

Overall, findings of the proposed method across different datasets and different achitectures is shown in Table 4.15.

4.3.5 Class Activation Maps (CAMs)

Class Activation Maps[1] [71, 72] (CAMs) emphasise regions of interest within objects, where a model directs its attention to recognise an object. To achieve this, CAMs are extracted from the model to visualise its focus. CAMs facilitate an understanding of whether the model is effectively learning discriminative features. We utilise a the pretrained ResNet-50 model and feed augmented images to extract CAMs. We compare our proposed approach with pertinent data augmentation methods, including cutout, hide-and-seek, and autoaugment. Figure 4.7 shows the results, where the first column portrays original input data, the subsequent four columns exhibit four distinct data augmentation methods, and the last column showcases our

[1] `https://github.com/chaeyoung-lee/pytorch-CAMs`

Table 4.14: VOC 2007 test detection average precision (%). FRCN ⋆ refers to FRCN with training schedule in [81], and mAP is calculated accross all 20 classes. Highlighted blue is the best performance

Method	**trainSet**	**mAP**	**aero**	**bike**	**bird**	**boat**	**bottle**	**bus**	**car**	**cat**	**chair**	**cow**
FRCN [92]	07	66.9	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7
FRCN ⋆ [81]	07	69.1	75.4	80.8	67.3	59.9	37.6	81.9	80.0	84.5	50.0	77.1
ASDN [81]	07	71.0	74.4	81.3	67.6	57.0	46.6	81.0	79.3	86.0	52.9	75.9
IRE [18]	07	70.5	75.9	78.9	69.0	57.7	46.4	81.7	79.5	82.9	49.3	76.9
ORE [18]	07	71.0	75.1	79.8	69.7	60.8	46.0	80.4	79.0	83.8	51.6	76.2
I+ORE [18]	07	71.5	76.1	81.6	69.5	60.1	45.6	82.2	79.2	84.5	52.5	78.7
Ours												
N-RSA	07	72.47	76.3	83.0	62.7	55.9	53.4	80.6	87.9	82.2	50.0	74.9
W-RSA	07	**72.92**	74.1	85.7	68.6	55.8	52.7	84.4	86.5	85.4	51.8	75.6
Method	**trainSet**	**mAP**	**table**	**dog**	**horse**	**mbike**	**person**	**plant**	**sheep**	**sofa**	**train**	**tv**
FRCN [92]	07	-	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8
FRCN ⋆ [81]	07	-	68.2	81.0	82.5	74.3	69.9	28.4	71.1	70.2	75.8	66.6
ASDN [81]	07	-	73.7	82.6	83.2	77.7	72.7	37.4	66.3	71.2	78.2	74.3
IRE [18]	07	-	67.9	81.5	83.3	76.7	73.2	40.7	72.8	66.9	75.4	74.2
ORE [18]	07	-	67.8	81.2	83.7	76.8	73.8	43.1	70.8	67.4	78.3	75.6
I+ORE [18]	07	-	71.6	80.4	83.3	76.7	73.9	39.4	68.9	69.8	79.2	77.4
	Ours											
N-RSA	07	-	71.9	81.1	85.4	83.6	81.8	44.3	67.2	70.7	81.0	75.7
W-RSA	07	-	71.8	80.5	88.2	83.1	81.5	44.6	65.8	67.4	80.0	74.8

Table 4.15: Summary of the proposed RandSaliencyAug performance across all datasets using the full set of architectures evaluated for each dataset. Baseline, Non-Weighted RSA (N-RSA), and Weighted RSA (W-RSA) results are reported along with the performance metric used. In most settings, W-RSA achieves the best performance.

Dataset	**Baseline**	**N-RSA**	**W-RSA**	**Metric**	**Architectures Used**
Fashion-MNIST	93.22–93.96	95.01–95.36	**95.18–95.37**	Accuracy (%)	ResNet-20, ResNet-32, ResNet-44, ResNet-56
CIFAR-10	84.67–95.02	90.03–96.96	**90.43–96.98**	Accuracy (%)	ResNet-18, ResNet-34, ResNet-50, DenseNet-121, DenseNet-169, EfficientNet-B0
CIFAR-100	63.14–75.67	67.84–77.45	**68.01–78.01**	Accuracy (%)	ResNet-18, ResNet-34, ResNet-50, DenseNet-121, DenseNet-169, EfficientNet-B0
TinyImageNet	38.58–42.33 (error↓)	34.97–33.10 (error↓)	**34.91–32.92** (error↓)	Error rate (%) ↓	PreActResNet-18, PreActResNet-50
ImageNet	76.5–78.3	77.9–79.5	**78.1–79.8**	Accuracy (%)	ResNet-50, ResNet-101, ResNet-152
VOC 2007	69.1	72.47	**72.92**	mAP (%)	Faster R-CNN (VGG-16)

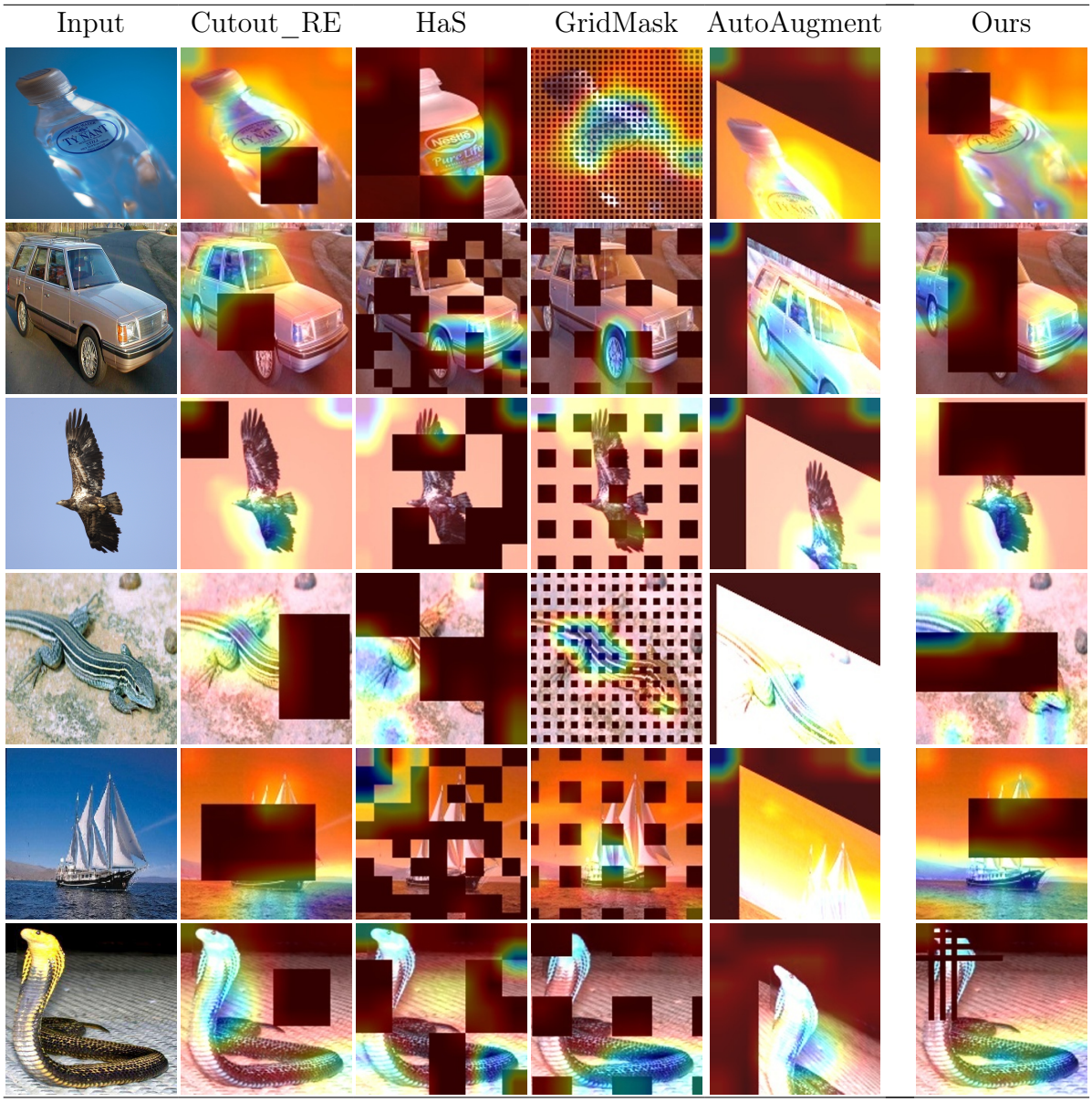


Figure 4.7: Class Activation Map comparison of different data augmentation methods with ours

proposed approach. In the cutout or random erasing method, the random erasure of regions influences the model's focus. For instance, in the case of a bottle, the focus shifts to the upper part of the bottle. Similarly, for an eagle, erasing unimportant regions shifts the focus to the lower part of the eagle. In the case of a snake, even though the snake itself is not erased, the model focuses solely on the head, relegating less important regions that generally contribute to generalisation. With the hide-and-seek data augmentation as shown in third column of Figure 4.7, the random and varied region concealment in each augmented image drastically shifts the focus. For example, a car is identified solely by its headlights, rather than its tires and other features. Similarly, the lizard classification predominantly relies on the focus on the foot, which could belong to any squamate species. This holds true for the other augmented class images as well.

In the GridMask augmentation as shown in fourth column of Figure 4.7, the grid-based erasure causes intriguing patterns. For instance, the focus on a car entirely shifts to its tires, a key feature. In the case of the lizard class, the focus is on its back, with some attention to the head. However, less significant features remain unfocused in all the augmented images.

For autoaugment as shown in fifth column of Figure 4.7, the focus primarily centers on only important regions. In the eagle case, feathers are almost completely disregarded, with the head being the key focus, a crucial region. The same principle applies to the snake image and other instances.

Our proposed approach, as depicted in the last column of Figure 4.7, involves partially occluding salient regions using various strategies. Remarkably, the model maintains focus on all regions, regardless of their significance. In the car image, for instance, important regions are partially erased, yet the model continues to concentrate on the headlights and tires simultaneously, which differentiates it from other methods. Similarly, in the eagle case, although the upper feathers are obscured, the model still directs its focus to the head and other feathers. This applies to the remaining images using our approach as well. In summary, CAMs suggest that

our proposed method is more generalised, capable of recognising objects even with partial occlusion, setting it apart from other methodologies.

Table 4.16: Results on CIFAR10 using various models architectures and various baselines. 'Time' reports the per epoch training time on google colab GPU. 'Accuracy' reports the accuracy on test set. Accuracy of existing methods are taken from KeepAugment [31].

Wide ResNet-28-10	**Accuracy (%)**	**Time (sec)**
GridMask	97.5 ± 0.1	70
AugMix	97.5 ± 0.0	71
Attentive CutMix	97.3 ± 0.1	90
KeepAutoAugment+L	$\mathbf{97.8 \pm 0.1}$	86
AutoMix		262.94
Ours (N-RSA)	97.6 ± 0.1	75
Ours (W-RSA)	97.6 ± 0.1	75
ShakeShake	**Accuracy** (%)	**Time (sec)**
GridMask	97.4 ± 0.1	48
AugMix	97.5 ± 0.0	49
Attentive CutMix	97.4 ± 0.1	110
KeepAutoAugment+L	$\mathbf{97.9 \pm 0.1}$	98
AutoMix		113.78
Ours (N-RSA)	97.4 ± 0.2	55
Ours (W-RSA)	97.3 ± 0.3	55

Table 4.17: Only augmentation time - time requires to perform data augmentation on a single image

Augmentation	**Time (sec)**
Grid Mask	0.05
Random Erasing	0.02
CutOut	0.02
Hide and Seek	0.06
AutoMix	0.87
Ours (N-RSA)	0.07
Ours (W-RSA)	0.07

4.3.6 Computational complexity

We evaluate the computational time required for the proposed data augmentation method and compare it with various existing data augmentation techniques. The training epoch time for the proposed augmentation is contrasted with different augmentation methods in Table 4.16. Among the methods listed in Table 4.16, Atten-

tive CutMix demands more time due to its process of generating a heatmap from the first image using a pretrained model, selecting N patches from that image, and then pasting them onto another image. Subsequently, KeepAutoAugment consumes more time as it identifies salient regions and applies RandomAug [52] exclusively to non-salient regions. In comparison, the proposed method exhibits faster processing times than these approaches. However, it does take slightly more time than GridMask and Augmix, since our proposed approach detects salient regions. Additionally, we examine the time required for augmenting a single image, presented in Table 4.17. Notably, Cutout and random erasing are the quickest augmentation methods in terms of time consumption. On the other hand, our proposed approach consumes more time due to its process of saliency detection followed by selecting the appropriate augmentation strategy. Considering the performance improvements achieved, the time overhead is negligible.

Table 4.18: Performance comparison on adversarial robustness using different data augmentation methods on adversarially perturbed ImageNet validation set.

	BASELINE	**CUTOUT**	**MIXUP**	**CUTMIX**	**SALIENCYMIX**	**Ours (N-RSA)**	**Ours (W-RSA)**
ACC. (%)	8.2	11.5	24.4	31.0	**32.96**	29.3	29.7

4.3.7 Robustness against adversarial attacks

Adversarial attacks reveal deep learning models' vulnerability to subtle perturbations that can deceive models even when barely noticeable or unrecognisable, referred to as adversarial examples [43, 69]. Given that data augmentation inherently involves interacting with examples, it naturally provides a means for generating adversarial instances. Consequently, it becomes crucial to assess the proposed data augmentation's robustness and compare it against alternative methods. Our study employs a trained ResNet-50 model augmented with our techniques, tested on the adversarially ImageNet validation set. Robustness is evaluated through adversarially perturbed examples from the ImageNet validation set using the established Fast Gradient Sign Method (FGSM) [44]. Notably, the proposed method demonstrates heightened robustness compared to other techniques and comparable resilience to

image mixing methods like CutMix and SaliencyMix, as indicated in Table 4.18. This showcases the potential of tailored data augmentation approaches in enhancing model robustness in adversarial contexts.

4.3.8 Discussion on comparison with complex augmentation policies.

As shown in Tables 4.5 and 4.7, RandSaliencyAug does not always outperform the strongest mixing-based approaches such as AutoMix and PuzzleMix. However, these methods operate under substantially more complex augmentation policies: they jointly optimise mixing patterns and often combine colour, intensity, and spatial transformations within large, tuned policy spaces. By contrast, RSA is deliberately constrained to saliency-guided erasing on a single image, without search over augmentation policies or additional mixing operations. This design choice makes RSA closer in spirit to erasing-based methods (e.g., RE, HaS, GridMask), for which our comparisons are more balanced and where W-RSA consistently matches or improves upon their performance across Fashion-MNIST, CIFAR-10/100, TinyImageNet, ImageNet, and VOC 2007. Moreover, the computational analysis in Section 4.3.6, Table 4.16 and Table 4.17 shows that RSA introduces only a modest overhead relative to simple erasing (due to saliency estimation), while remaining more efficient than heavier saliency-based or policy-driven approaches such as Attentive CutMix, KeepAutoAugment and AutoMix. As a result, RSA is more efficient than these.

4.4 Conclusion

In this chapter, we presented the RandSaliencyAug framework, which addresses research question 2 by leveraging six distinct strategies—Row Slice Erasing, Column Slice Erasing, Row-Column Saliency Erasing, Partial Saliency Erasing, Horizontal Half Saliency Erasing, and Vertical Half Saliency Erasing. These strategies provide a diverse search space, allowing the approach to strike a balance between complete

object removal and the preservation of critical contextual information. Additionally, we explored two variants of the framework, namely the weighted and non-weighted versions, to ensure comprehensive exploration and validation.

Beyond performance evaluation, this study delves into key aspects such as time complexity, explainability through Class Activation Maps (CAM), and robustness. The computational efficiency of RandSaliencyAug was thoroughly examined, and its interpretability was enhanced using CAM, which provided insights into the model's focus and contributed to greater transparency. Importantly, the proposed approach demonstrated robustness across multiple tasks and datasets.

Through empirical validation, RandSaliencyAug achieved excellent results in both image classification tasks on Fashion-MNIST, CIFAR10, CIFAR100, tinyImageNet, and ImageNet, as well as object detection tasks on PASCAL VOC, utilising a range of CNN architectures. These findings confirm the efficacy and versatility of RandSaliencyAug across various tasks and benchmarks, marking a significant advancement in enhancing the robustness and performance of computer vision models.

Future work could explore the performance of RandSaliencyAug on occluded datasets, a potential area for further investigation. Additionally, optimising the parameters, such as weights and probabilities, remains a limitation and is left for future research.

This chapter builds directly on RSMDA's lesson that preserving discriminative details matters. In this chapter, we combined saliency with erasing to balance object and context and to curb overfitting without destroying feature fidelity. The outcome will inform Chapter 5, where we extend beyond "keep salient, touch context" to actively combine salient and non-salient regions and vary placement to avoid domain shift.

Chapter 5

Combining Salient and non-Salient Regions in Data Augmentation

5.1 Introduction and Motivation

Building on the RandSaliencyAug framework introduced in Chapter 4, which combined saliency and erasing to balance object and context and curb overfitting without destroying feature fidelity, this chapter moves from *where and what to erase* to *how to recombine and which part to augmnent* salient and non-salient regions. The analysis in Chapter 4 showed that protecting salient content while carefully occluding context is beneficial, but also highlighted an object–context trade-off and the risk of domain shift if salient and non-salient areas are treated too asymmetrically. In this chapter, we address these limitations by designing augmentations that explicitly mix and relocate both types of regions.

This chapter focuses on combining salient and non-salient regions in data augmentation, addressing Research Question 3 (RQ3), as defined in Section 1.3.3.

The field of computer vision has witnessed remarkable advancements in recent years, driven in large part by innovative data augmentation techniques aimed at enhancing model generalisation. These techniques, such as SalfMix and KeepAugment, have shown their efficacy in improving model performance. However, they are not

without their limitations. SalfMix, a technique that duplicates salient features, can inadvertently lead to overfitting (Figure 5.1b), diminishing a model's ability to generalise beyond the training data. Similarly, KeepAugment, which selectively preserves salient regions during data augmentation, can introduce a domain shift between salient and non-salient regions(Figure) 5.1c), obstructing the seamless exchange of vital contextual information and hampering the overall model's understanding.

The need for a more balanced approach that harnesses both diverse salient and diverse non-salient regions for augmentation becomes evident. This balance is crucial to empower models to achieve higher performance improvements without compromising on information preservation. The field of computer vision has witnessed

(a) Input (b) SalfMix (c) KeepAugment (d) Ours

Figure 5.1: Comparison of the relevant data augmentation methods with ours

remarkable advancements in recent years, driven in large part by innovative data augmentation techniques aimed at enhancing model generalisation. These techniques, such as SalfMix and KeepAugment, have shown their efficacy in improving model performance. However, they are not without their limitations. SalfMix, a technique that duplicates salient features, can inadvertently lead to overfitting 5.1b, diminishing a model's ability to generalise beyond the training data. Similarly, KeepAugment, which selectively preserves salient regions during data augmentation, can introduce a domain shift between salient and non-salient regions 5.1c, obstructing the seamless exchange of vital contextual information and hampering the overall model's understanding.

The need for a more balanced approach that harnesses both diverse salient and diverse non-salient regions for augmentation becomes evident. This balance is crucial to empower models to achieve higher performance improvements without compro-

Table 5.1: Comparison of KeepOriginalAugment with SOTA augmentation methods. "Mix" indicates whether images are mixed; "One-image" shows if a single image is used; "Saliency" denotes use of saliency; "SRA" and "NSRA" refer to augmentation of only salient or non-salient regions, respectively; "Blending" indicates whether images are blended.

Method	**Mix ?**	**One-Image?**	**Saliency?**	**SRA?**	**NSRA?**	**Blend?**
Mixup [27]	✗	✗	✗	✗	✗	✓
CutMix [14]	✓	✗	✗	✗	✗	✗
SaliencyMix [29]	✓	✗	✓	✗	✗	✗
ResizeMix [91]	✓	✗	✗	✗	✗	✗
PuzzleMix [89]	✓	✗	✓	✗	✗	✓
Self-Augment [93]	✓	✓	✗	✗	✗	✗
SalfMix [30]	✓	✓	✓	✗	✗	✗
KeepAugment [31]	✓	✓	✓	✗	✓	✗
Ours	✓	✓	✓	✓	✓	✗

mising on information preservation. In light of these challenges, our research introduces a novel data augmentation approach: KeepOriginalAugment. The motivation behind KeepOriginalAugment lies in its ability to strike an optimal balance between data diversity and information preservation. By intelligently incorporating the most salient region within the non-salient region, it enables augmentation to be applied to either or both regions. This approach holds the promise of significantly enhancing model performance, as our experiments on benchmark datasets demonstrate. In a landscape where data augmentation is pivotal to model success, our research seeks to address the limitations of existing techniques, contributing a novel solution that empowers models to leverage the full potential of diverse data while maintaining feature fidelity. More precisely, KeepOriginalAugment has been differentiated from existing SOTA methods in Table 5.1 and overall KeepOriginalAugment process is shown in Figure 5.2. The contributions of our work in this chapter are as follows:

- We propose KeepOriginalAugment as a solution to the limitations of SalfMix and KeepAugment.

- We validate the effectiveness of our approach through experiments on various datasets using different network architectures.

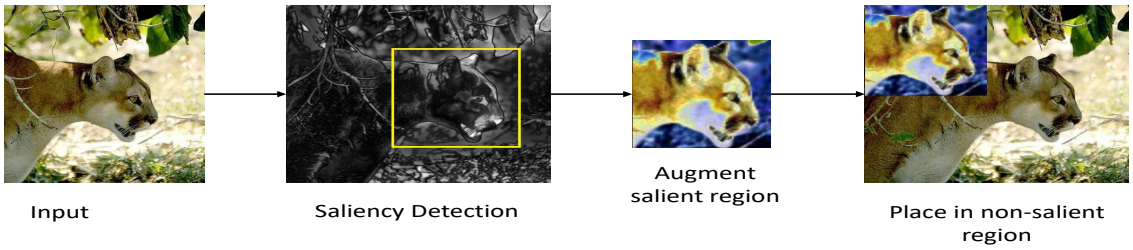


Figure 5.2: Overall KeepOriginalAugment process

5.2 Methodology

Our primary objective is to enhance fidelity by preserving both the original and augmented information within a single image, thereby encouraging the model to learn diverse information. To achieve this, we begin by measuring the importance of different regions in the image using a saliency map. Let $s_{(\mathrm{i,j})}(x, y)$ represent the saliency map of image x with label y at pixel (i, j). We calculate the importance score I for a specific region R as follows:

$$I(R, x, y) = \sum_{(i,j)\in R} s_{(\mathrm{i,j})}(x, y) \tag{5.1}$$

In this equation, we sum up the saliency scores of all pixels within the region R. We employ a standard saliency map based on the vanilla gradient method proposed by Simonyan et al. [94]. Next, we preserve only those regions that possess an importance score I greater than a predefined threshold τ. We used same τ as it is used in KeepAugment [31]. By identifying these important regions, we aim to address the issue of redundant salient features present in SalfMix, as well as tackle the domain shift problem encountered in KeepAugment.

How is grid made?

After saliency detection, a single salient region has been identified in the image. Let the image have height H and width W. The salient region is represented by the coordinates of its top left and bottom right corners,

$$(x_1, y_1) \quad \text{and} \quad (x_2, y_2),$$

with

$$0 \le x_1 < x_2 \le W, \qquad 0 \le y_1 < y_2 \le H.$$

The corresponding salient bounding box is the axis-aligned rectangle

$$R_{\text{sal}} = [x_1, x_2) \times [y_1, y_2).$$

This salient box induces a 3×3 partition (grid) over the entire image. We use the following vertical and horizontal lines:

$$x = 0,\ x_1,\ x_2,\ W \quad \text{and} \quad y = 0,\ y_1,\ y_2,\ H.$$

These lines split the x- and y-axes into three intervals each:

$$X_{\text{left}} = [0, x_1), \quad X_{\text{mid}} = [x_1, x_2), \quad X_{\text{right}} = [x_2, W),$$
$$Y_{\text{top}} = [0, y_1), \quad Y_{\text{mid}} = [y_1, y_2), \quad Y_{\text{bottom}} = [y_2, H).$$

Taking Cartesian products of these intervals yields nine rectangular grid cells $G_{ij} = X_i \times Y_j$. Explicitly, in (x, y) coordinates:

$$\begin{aligned}
G_{\text{TL}} &= [0, x_1) \times [0, y_1) && \text{(top left)},\\
G_{\text{TM}} &= [x_1, x_2) \times [0, y_1) && \text{(top middle)},\\
G_{\text{TR}} &= [x_2, W) \times [0, y_1) && \text{(top right)},\\
G_{\text{ML}} &= [0, x_1) \times [y_1, y_2) && \text{(middle left)},\\
G_{\text{C}} &= [x_1, x_2) \times [y_1, y_2) && \text{(centre, salient region)},\\
G_{\text{MR}} &= [x_2, W) \times [y_1, y_2) && \text{(middle right)},\\
G_{\text{BL}} &= [0, x_1) \times [y_2, H) && \text{(bottom left)},\\
G_{\text{BM}} &= [x_1, x_2) \times [y_2, H) && \text{(bottom middle)},\\
G_{\text{BR}} &= [x_2, W) \times [y_2, H) && \text{(bottom right)}.
\end{aligned}$$

The heights of the three horizontal strips are

$$h_{\text{top}} = y_1, \qquad h_{\text{mid}} = y_2 - y_1, \qquad h_{\text{bottom}} = H - y_2,$$

and the widths of the three vertical strips are

$$w_{\text{left}} = x_1, \qquad w_{\text{mid}} = x_2 - x_1, \qquad w_{\text{right}} = W - x_2.$$

Consequently, each grid cell G has size given by the corresponding pair (h, w), e.g.

$$\text{size}(G_{\text{C}}) = (h_{\text{mid}}, w_{\text{mid}}) = (y_2 - y_1,\ x_2 - x_1),$$

Given a salient bounding box $[x_1, x_2) \times [y_1, y_2)$, the grid is defined by the lines at $x = 0, x_1, x_2, W$ and $y = 0, y_1, y_2, H$, whose Cartesian products form a 3×3 partition of the image with the salient region as the central cell.

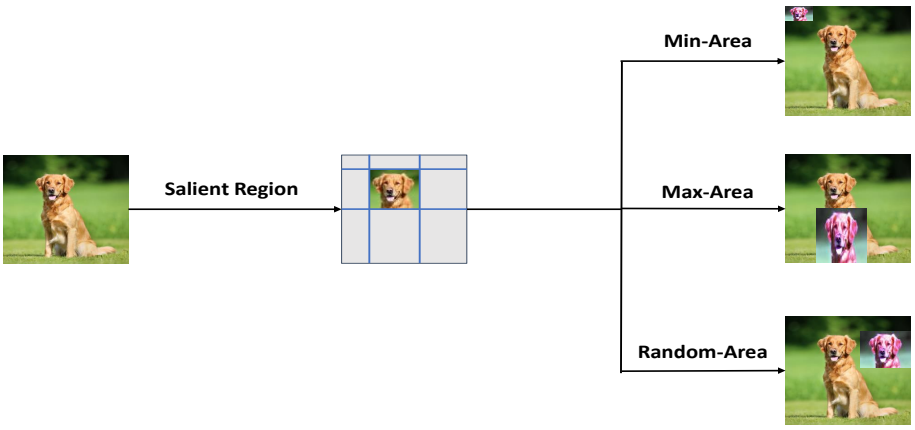


Figure 5.3: Strategies: where to place the salient region?

(a) **Where to place the salient region?** To determine the placement of the salient region within the non-salient region and achieve diversity, we explore three different strategies:

Min-Area: In this strategy, we identify eight regions surrounding the salient region. Among these regions, we select the one with the minimum area. We

then resize the salient region according to the size of that minimum area and place it within that region. This is illustrated in Figure 5.3.

Max-Area: Conversely, in the Max-Area strategy, we identify the eight regions and choose the one with the maximum area. The salient region is resized accordingly and placed within this region. This is demonstrated in Figure 5.3.

Random-Area: To introduce scaling augmentation to the salient region, we adopt a more flexible approach. Instead of limiting the placement to the minimum or maximum area, we randomly select one of the eight regions. The salient region is resized based on the size of the chosen region, and it is placed within that area. This is depicted in Figure 5.3.

Through experimental evaluations, we have observed that the Random-Area strategy provides greater diversity. This is attributed to the random scaling data augmentation applied to the salient region, which introduces varying degrees of augmentation effects.

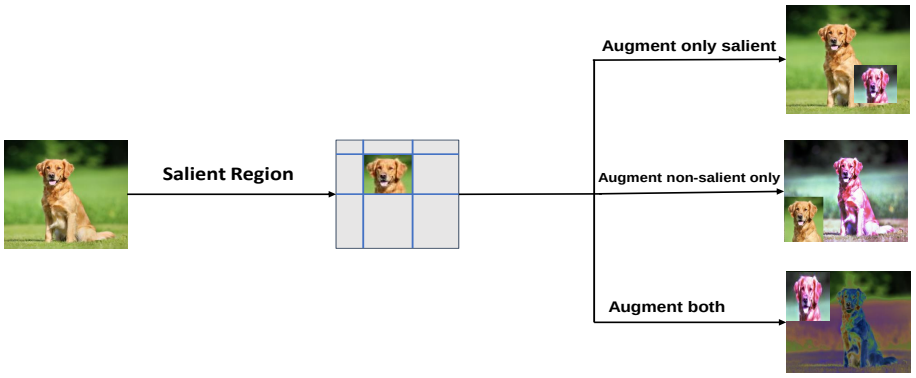


Figure 5.4: Strategies: Which part should be augmented?

(b) **Which part should be augmented?** After identifying the salient region, we propose and investigate three distinct strategies to enhance fidelity and diversity:

Augment only salient: In this strategy, we solely apply random augmentation to the salient region. Subsequently, we paste the augmented salient region

into the non-salient region of the original image. It is crucial to emphasise that augmentation is exclusively performed on the salient region. The application of this strategy is illustrated in Figure 5.4.

Augment non-salient only: In this strategy, we conduct augmentation on the entire image while preserving the original salient region. The augmented image is then combined with the original salient region, which is extracted from the unaltered original image. The overall approach is depicted in Figure 5.4.

Augment both: This strategy involves performing separate augmentations on both the salient region and the entire image. The augmented salient region is integrated with the augmented whole image, as shown in Figure 5.4. By considering a trade-off probability between the original sample and the augment both strategy sample, we observed that the augment both strategy demonstrates greater diversity and fidelity across various computer vision tasks as discussed in Section 5.3. It is important to note that we utilise randAug [52] for augmentation, which offers computational efficiency similar to that employed by KeepAugment.

5.3 Experiments

5.3.1 Training setup

For a fair comparison, we followed the training setups outlined in previous works, specifically KeepAugment [31], SalfMix [30], and random erasing [18]. Our training process consisted of 300 epochs, a batch size of 128, an initial learning rate of 0.1, and the CosineAnnealingLR scheduler. We employed stochastic gradient descent with a momentum of 0.9 and weight decay of 0.0005. To evaluate the generalisation performance, we utilised various neural network architectures such as ResNet, Wide-ResNet, and PreActResNet.

To assess the effectiveness of our proposed methods, we conducted experiments on different datasets, including CIFAR-10 (10 classes), CIFAR-100 (100 classes) [62],

and TinyImageNet (200 classes) [78, 13].

For the PreActResNet architecture, we employed the standard saliency function from the OpenCV library[1] since PreActResNet required more computation time for saliency estimation.

Throughout our experiments, we utilised accuracy and error rate as performance metrics. Higher accuracy values indicate better performance, while lower error rates are desired. For dataset diversity measurement and query retrieval results, we followed the same parameters as discussed in [32].

5.3.2 Hyperparameter

To determine the optimal combination of salient region placement strategy and region augment strategy, we conducted experiments using the ResNet-18 neural network architecture on the CIFAR-10 dataset.

Based on our experiments, we found that the augment both strategy, combined with the Random-Area strategy, yielded the best results in terms of performance. The reasons behind the effectiveness of this combination are discussed in Section 5.3.4.

5.3.3 Results

In Table 5.2, we present the performance of our methods, including KeepOriginalCutout, compared to the baseline and other state-of-the-art techniques, such as Cutout and KeepCutout, on the CIFAR-10 dataset. It is important to mention KeepOriginalCutout is combination of cutout and KeepOriginalAugment, which basically cutout the non-salient region then combine with KeepOriginalAugment as done by KeepAugment [31]. Our method, KeepOriginalCutout, demonstrated superior accuracy, with an absolute 1% improvement over the baseline and a 0.6% improvement over the KeepCutout method, when evaluated using the ResNet-18

[1] *https://docs.opencv.org/3.4/da/dd0/classcv_1_1saliency_1_1StaticSaliency FineGrained.html*

architecture. Additionally, our method showed competitive accuracy performance when evaluated using the ResNet-110 and Wide ResNet architectures.

Table 5.2: Test accuracy (%) on CIFAR10 dataset using various model architectures.

CIFAR10 Dataset			
Model	ResNet-18	ResNet-110	Wide ResNet-28-10
Cutout [26]	95.6 ± 0.1	94.8 ± 0.1	96.9 ± 0.1
KeepCutout [31, 26]	96.1 ± 0.1	95.5 ± 0.1	97.3 ± 0.1
KeepCutout (LR) [31, 26]	96.2 ± 0.1	95.5 ± 0.1	97.3 ± 0.1
KeepCutout (RL) [31, 26]	96.0 ± 0.1	95.3 ± 0.1	97.2 ± 0.1
N-RSA (RandSaliencyAug)	–	–	96.96
W-RSA (RandSaliencyAug)	–	–	96.98
KeepOriginalCutout (Ours)	96.6 ± 0.1	95.1 ± 0.2	97.1 ± 0.2

Model	WideResNet-28-10	PyramidNet
AutoAugment [85, 31]	97.3 ± 0.1	98.5 ± 0.0
KeepAutoAugment [31]	97.8 ± 0.1	98.7 ± 0.0
KeepAutoAugment (LR) [31]	97.8 ± 0.1	98.7 ± 0.0
KeepAutoAugment (EL) [31]	97.8 ± 0.1	98.6 ± 0.0
N-RSA (RandSaliencyAug)	96.96	–
W-RSA (RandSaliencyAug)	96.98	–
KeepOriginalAugment (Ours)	97.9 ± 0.3	98.6 ± 0.1

To assess the generalisation of our proposed approach, we conducted experiments on larger and more diverse datasets using various neural network architectures. The results, presented in Table 5.3, show that our proposed method achieved superior error rate performance compared to all other data augmentation methods. It even outperformed different versions of HybridMix, which combines several image-mixing methods as an ensemble, in most cases. However, there were a few instances where our proposed method did not surpass the performance of HybridMix, such as in the case of PreActResNet-50. Notably, our proposed approach achieved an absolute 2% improvement in error rate for the CIFAR-100 dataset when evaluated using the PreActResNet-101 architecture.

To gain a clearer picture of how KeepOriginalAugment behaves at the class level, we also report class-wise precision, recall, F1 score, and ROC AUC. These results are summarised in Table 5.4 for CIFAR10, Table 5.5 for CIFAR100, and Table 5.6 for TinyImageNet. Because CIFAR100 and TinyImageNet contain 100 and 200 classes, respectively, listing all per-class values in the main text is imprac-

Table 5.3: Test Error rate (%) on different datasets using various model architectures, where PARN represents PreActResNet.

CIFAR10 Dataset			
Model	PARN-18	PARN-50	PARN-101
Baseline	5.17 ± 0.27	4.6 ± 0.2	4.49 ± 0.18
+Cutout [26, 30]	4.3 ± 0.09	3.77 ± 0.08	3.54 ± 0.11
+SalfMix [30]	4.14 ± 0.25	3.61 ± 0.09	3.38 ± 0.11
+Mixup [30, 67]	4.1 ± 0.39	3.56 ± 0.04	3.54 ± 0.08
+SaliencyMix [29, 30]	3.8 ± 0.1	2.98 ± 0.1	2.82 ± 0.08
+CutMix [14, 30]	3.96 ± 0.21	3.07 ± 0.09	2.95 ± 0.06
+ResizeMix [91, 30]	3.74 ± 0.2	3.09 ± 0.11	2.85 ± 0.09
+HybridMix v1 [30]	3.85 ± 0.13	3.22 ± 0.04	3.04 ± 0.14
+HybridMix v2 [30]	3.74 ± 0.05	2.94 ± 0.09	2.78 ± 0.04
+HybridMix v3 [30]	3.38 ± 0.07	$\mathbf{2.89 \pm 0.11}$	2.75 ± 0.07
+Ours- KeepOriginalAugment	$\mathbf{3.30 \pm 0.10}$	3.20 ± 0.30	$\mathbf{2.70 \pm 0.00}$
CIFAR100 Dataset			
Model	PARN-18	PARN-50	PARN-101
Baseline	24.22 ± 0.22	22.02 ± 0.18	21.81 ± 0.24
+Cutout [26, 30]	23.72 ± 0.27	21.64 ± 0.43	21.46 ± 0.25
+SalfMix [30]	22.64 ± 0.13	20.48 ± 0.17	19.89 ± 0.13
+Mixup [30, 67]	21.78 ± 0.4	18.91 ± 0.26	18.82 ± 0.37
+SaliencyMix [29, 30]	20.02 ± 0.13	17.5 ± 0.16	17.33 ± 0.09
+CutMix [14, 30]	20.51 ± 0.17	17.72 ± 0.17	17.61 ± 0.25
+ResizeMix [91, 30]	20.96 ± 0.11	17.56 ± 0.09	17.36 ± 0.19
+HybridMix v1 [30]	21.42 ± 0.17	18.27 ± 0.12	17.45 ± 0.12
+HybridMix v2 [30]	19.88 ± 0.27	17.38 ± 0.27	17.22 ± 0.21
+HybridMix v3 [30]	19.84 ± 0.09	$\mathbf{17.3 \pm 0.25}$	17.25 ± 0.23
+Ours- KeepOriginalAugment	$\mathbf{18.70 \pm 0.30}$	17.80 ± 0.2	$\mathbf{15.90 \pm 0.11}$
TinyImageNet Dataset			
Model	PARN-18	PARN-50	
Baseline	42.33 ± 0.21	38.58 ± 0.24	
+Cutout [26, 30]	42.04 ± 0.31	38.36 ± 0.21	
+SalfMix [30]	40.28 ± 0.28	35.92 ± 0.07	
+Mixup [30, 67]	40.22 ± 0.2	35.51 ± 0.15	
+SaliencyMix [29, 30]	37.76 ± 0.05	$\mathbf{32.83 \pm 0.47}$	
+CutMix [14, 30]	38.11 ± 0.32	33.54 ± 0.19	
+ResizeMix [91, 30]	38.47 ± 0.25	33.25 ± 0.12	
+Ours- KeepOriginalAugment	$\mathbf{35.1 \pm 0.20}$	35.6 ± 0.12	

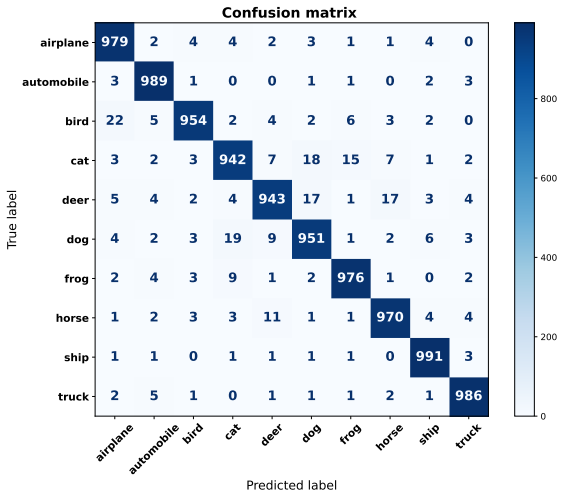
tical. Instead, the complete class-wise metrics, together with the corresponding confusion matrices, are provided in our online repository at `https://github.com/kmr2017/ThesisPhDCode/tree/main/RevisionExperiments/RQ3Files`. In addition, Figure 5.5 shows representative confusion matrices for the different models on CIFAR10, while the remaining confusion matrices for the other datasets are also available in the same repository.

Table 5.4: Class-wise metrics for different models on CIFAR10 with KeepOriginalAugment.

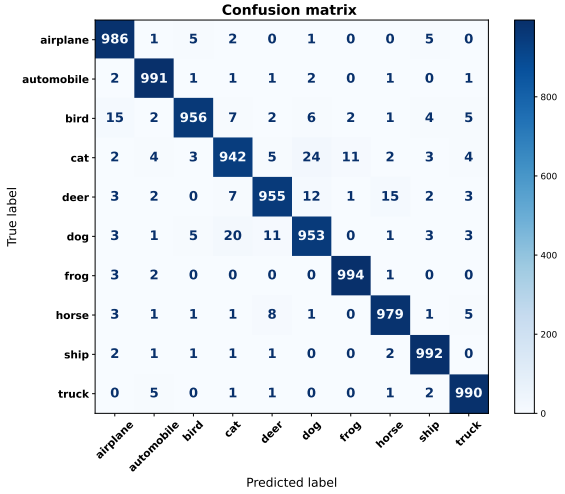
Class	Precision	Recall	F1	ROC AUC
PreActResNet101 with KeepOriginalAugment				
Airplane	0.9676	0.9860	0.9767	0.9972
Automobile	0.9812	0.9910	0.9861	0.9973
Bird	0.9835	0.9560	0.9696	0.9952
Cat	0.9593	0.9420	0.9506	0.9898
Deer	0.9705	0.9550	0.9627	0.9931
Dog	0.9540	0.9530	0.9535	0.9908
Frog	0.9861	0.9940	0.9900	0.9990
Horse	0.9761	0.9790	0.9775	0.9985
Ship	0.9802	0.9920	0.9861	0.9988
Truck	0.9792	0.9900	0.9846	0.9986
Mean	**0.9738**	**0.9738**	**0.9737**	**0.9958**
PreActResNet18 with KeepOriginalAugment				
Airplane	0.9579	0.9790	0.9683	0.9944
Automobile	0.9734	0.9890	0.9812	0.9973
Bird	0.9795	0.9540	0.9666	0.9950
Cat	0.9573	0.9420	0.9496	0.9833
Deer	0.9632	0.9430	0.9530	0.9921
Dog	0.9539	0.9510	0.9524	0.9844
Frog	0.9721	0.9760	0.9741	0.9947
Horse	0.9671	0.9700	0.9685	0.9930
Ship	0.9773	0.9910	0.9841	0.9987
Truck	0.9791	0.9860	0.9826	0.9970
Mean	**0.9681**	**0.9681**	**0.9680**	**0.9930**
ResNet18 with KeepOriginalCutout				
Airplane	0.9626	0.9780	0.9702	0.9961
Automobile	0.9753	0.9890	0.9821	0.9980
Bird	0.9795	0.9570	0.9681	0.9945
Cat	0.9599	0.9340	0.9468	0.9894
Deer	0.9664	0.9500	0.9581	0.9958
Dog	0.9513	0.9580	0.9547	0.9920
Frog	0.9702	0.9770	0.9736	0.9941
Horse	0.9692	0.9740	0.9716	0.9966
Ship	0.9812	0.9910	0.9861	0.9981
Truck	0.9792	0.9870	0.9831	0.9971
Mean	**0.9695**	**0.9695**	**0.9694**	**0.9952**

5.3.4 Discussion

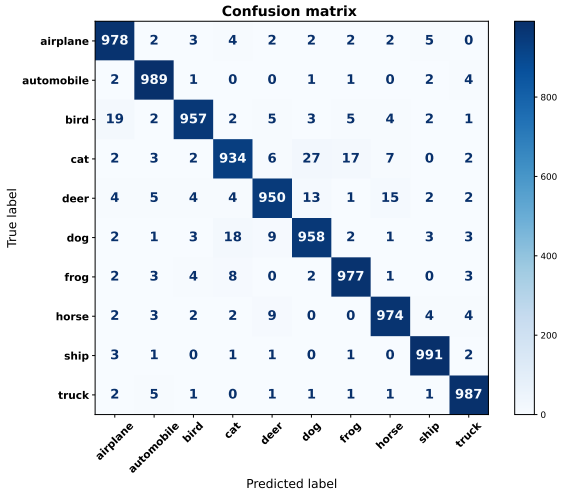
In this study, we aimed to answer two important questions while addressing the limitations of SalfMix and KeepAugment. The first question was **Where to place**



(a) PreActResNet-18 (KeepOriginalAugment)



(b) PreActResNet-101 (KeepOriginalAugment)



(c) ResNet-18 (KeepOriginalCutout)

Figure 5.5: Confusion matrices on CIFAR-10 for different models using KeepOriginal-based augmentations.

Table 5.5: Mean of the metrics for different models on CIFAR100.

Class	Precision	Recall	F1	ROC AUC
PreActResNet101 with KeepOriginalAugment				
Mean	**0.848**	**0.847**	**0.847**	**0.955**
PreActResNet18 with KeepOriginalAugment				
Mean	**0.818**	**0.817**	**0.817**	**0.949**

Table 5.6: Mean of the metrics for different models on TinyImageNet.

Class	Precision	Recall	F1	ROC AUC
PreActResNet101 with - KeepOriginalAugment				
Mean	**0.844**	**0.843**	**0.843**	**0.954**
PreActResNet18 with - KeepOriginalAugment				
Mean	**0.818**	**0.817**	**0.817**	**0.947**

the salient region? Experimental results indicated that the salient region should be placed randomly in any non-salient region. The random placement of the salient region provides scaling data augmentation to that specific region. For example, on epoch 1, it may be placed in the region with the minimum area, while on the next epoch, it may be placed in a region with a different area. This random placement strategy ensures that scaling data augmentation is applied to various regions, promoting diversity in the training process.

The second question we addressed was **Which parts should be augmented?** Based on our experiments, both the salient region and the whole image should be augmented. This approach allows the neural network to access both the original data and diverse augmented data during training. By providing only augmented examples to the model, there is a concern that the model may not have access to the original data, which can lead to a shift in the data distribution. To mitigate this, we adopted a well-known technique of finding a balance between the augmented and original samples. We used a probability of 0.5 as a trade-off, meaning that during training, the neural network was fed with original and augmented data in equal proportions as it is investigated by random erasing [18]. This dynamic combination of original and augmented data ensures that the model learns from both sources, enhancing diversity and fidelity in the learned features.

5.4 Conclusion

In this chapter, we introduced KeepOriginalAugment, a novel method designed to address the challenges of feature redundancy and domain shift in data augmentation, thereby answering research question 3. Our method provides an effective solution for enhancing the performance and generalisation of deep learning models by optimising strategies for placing salient regions and selecting augmented parts.

Through comprehensive experiments across diverse datasets and various network architectures, KeepOriginalAugment demonstrated superior performance, consistently outperforming existing state-of-the-art (SOTA) methods. Our approach achieved higher accuracy and lower error rates across different datasets, proving its effectiveness in improving model performance. The method's key strength lies in its ability to provide scaling data augmentation to randomly chosen non-salient regions, ensuring diversity and richness in the training data. Additionally, by augmenting both the salient regions and the entire image, we maintained a balance between original and augmented samples, which helped prevent potential data distribution shifts. In the next chapter, we will evaluate the proposed approach and its impact on bias detection and mitigation, particularly by analysing dataset diversity from gender and professional perspectives.

This chapter generalises the saliency principle: instead of only protecting salient parts (risking domain shift), KeepOriginalAugment mixes what is augmented (salient, non-salient, or both) with where it is placed, provide scaling and diversity while preserving information. These design choices and their hyper-parameters directly seed Chapter 6, where the same mechanisms (RandSaliencyAug and KeepOriginalAugment) are adapted to face data to measure and mitigate gender and geographic stereotypes.

Chapter 6

Assessing and Mitigating Gender Bias using Saliency in Data Augmentation

6.1 Introduction and Motivation

Building on the saliency-based augmentation methods developed in Chapters 4 and 5 (RandSaliencyAug and KeepOriginalAugment), this chapter shifts the focus from improving general performance and information preservation to assessing and mitigating bias. Chapter 5 generalised the saliency principle by mixing what is augmented (salient, non-salient, or both) with where it is placed, providing scaling and diversity while preserving information. in this chapter, we adapt the same mechanisms to fairness-sensitive facial datasets, using them to analyse and reduce gender-related and professional stereotypes rather than solely to improve accuracy.

This chapter presents and discusses data augmentation techniques that focus on assessing and mitigating gender bias using saliency on facial regions, while also exploring dataset diversity. The investigation is carried out to address Research Question 4 (RQ4), as defined in Section 1.3.4. In addition, we propose a novel saliency-based metric for measuring bias and fairness, and introduce face-specific variants of

our previous augmentations (e.g., FaceRandSaliencyAug and FaceKeepOriginalAugment) to evaluate and reduce gender bias across different datasets and professional categories.

Computer vision models have shown various biases, including racial bias [34, 35], gender bias [34, 36], and geographical bias [32, 33]. For example, system detects women with white skin more accurately than those with darker skin tones in facial recognition accuracy [34]. Another illustration is oxygen treatment, which is a widely used medical procedure that is tracked by a pulse oximeter, which uses infrared light to assess blood oxygen levels. However, this method tends to overestimate oxygen saturation in non-white patients, leading to under-detection of low oxygen levels, particularly among black patients, who are three times more likely to experience this discrepancy compared to white patients [38]. One of the most common source of bias is the dataset used to train the model, which can carry that bias into real-time deployment. For example, a common method of image collection for training datasets involves collecting images from online sources such as search engines like Google or image hosting platforms such as Flickr, as demonstrated in the case of FFHQ [39]. However, this approach is prone to bias and often results in the devolvement of biased datasets. Audits of social bias in visual datasets have primarily focused on two key attributes: race particularly skin tone, and gender [32, 34, 35]. This highlights the importance of facial region when aiming to mitigate bias in such datasets. To deal with these biases, several methods have been proposed [40, 41, 42]. Work by [40] investigates debiasing in image classification task by generating adversarial examples to solve bias in training data distribution thus to increase model fairness. Work by Kim et al. [41] introduces BiaSwap, which debiases deep neural networks without prior knowledge of bias types, using unsupervised sorting and style transfer to swap bias attributes between images. Work by Lee et al. [42] introduced DiverseBias, a feature-level data augmentation technique for improving the debiasing of image classification models by synthesizing diverse bias-conflicting samples through disentangled representation learning.

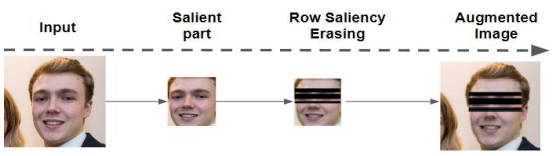
This chapter address RQ4 by investigating RandSaliencyAug and KeepOriginalAugment, previously introduced in Chapter 4 and Chapter 5, respectively, from a bias perspective. We refer to these approaches as FaceSaliencyAug and FaceKeepOriginalAugment, as they are utilised to explore various biases using facial region.

6.2 Methodology

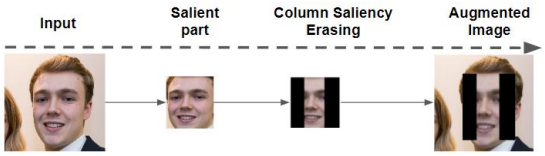
In this section, we discuss the FaceSaliencyAug and FaceKeepOriginalAugment methods.

6.2.1 FaceSaliencyAug

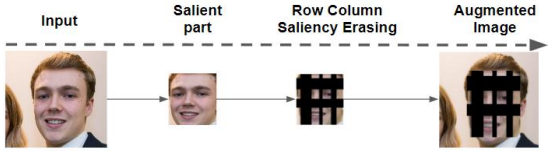
We present six data augmentation strategies that define the search space, as previously discussed in relation to Research Question 2, and outline our proposed approach based on this search space.



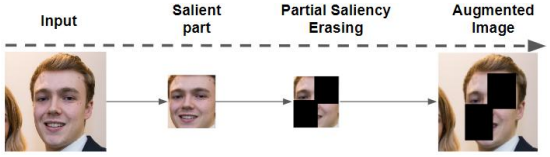
(a) Sample of Row Slice Erasing (RSE) visualisation



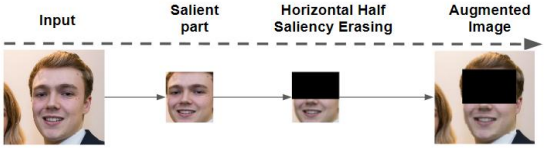
(b) Sample of Column Slice Erasing (CSE) visualisation



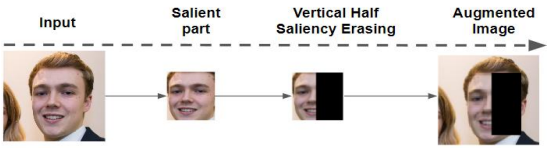
(c) Sample of Row Column Slice Erasing (RCSE) visualisation



(d) Sample of Partially Slice Erasing (PSE) visualisation



(e) Sample of Horizontal Half Slice Erasing (HHSE) visualization



(f) Sample of Vertical Half Slice Erasing (VHSE) visualisation

Figure 6.1: Visualisation of the proposed augmentation strategies for the search space.

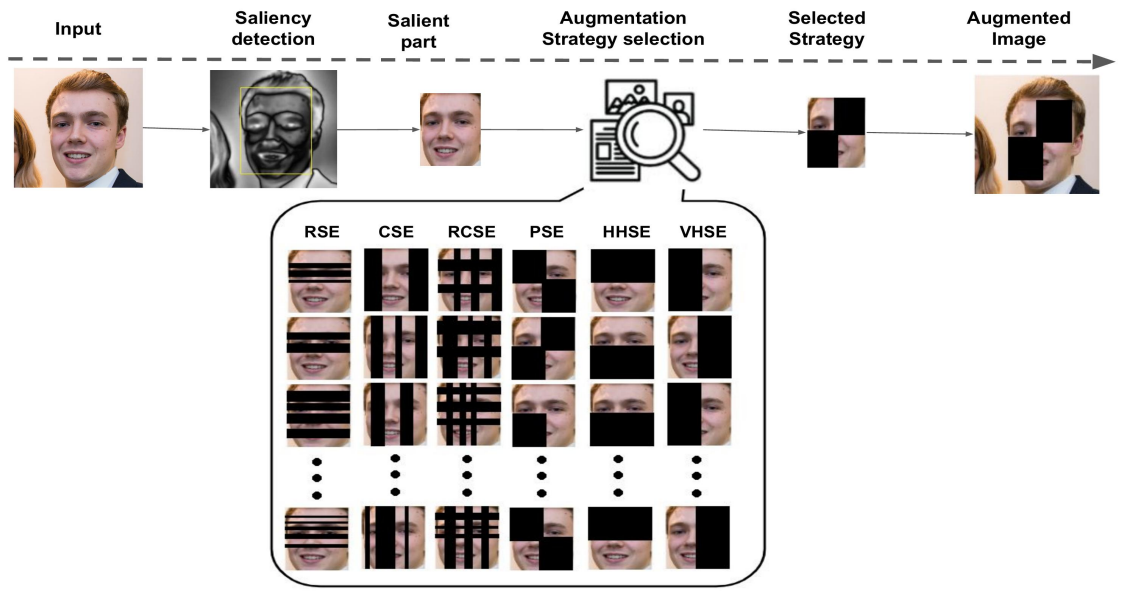


Figure 6.2: FaceSaliencyAug: Proposed approach to balance between complete object erasing and contextual information erasing, where RSE, CSE, RCSE, PSE, HHSE and VHSE represent row slice erasing, column slice erasing, row-column saliency erasing, partial saliency erasing, horizontal half saliency erasing and vertical half saliency erasing, respectively.

- **Search Space - Data Augmentation**

  The search space encompasses six proposed data augmentation techniques, each detailed below. It's worth noting that these augmentation methods operate on the salient region of the image, which is detected using methodologies outlined in previous works.

  Row Slice Erasing (RSE) RSE technique augments the salient region x of image I by element-wise multiplication with a binary mask M. The mask M contains values of 0 or 1, denoting exclusion or inclusion of pixels, respectively. Slices of size S are randomly selected from the salient region to generate the binary mask. Horizontal slices of the mask are filled alternately with 0's and 1's. Refer to Figure 6.1a for visualisation.

$$\tilde{x} = x \odot M \tag{6.1}$$

  Column Slice Erasing (CSE) Similar to RSE, CSE involves applying aug-

mentation to the salient region x of the image I. The augmented salient part $\tilde{x}$ is obtained using a binary mask M generated by selecting slices of size S from the salient region. However, in this strategy, vertical slices of the mask are alternately filled with 0's and 1's. See Figure 6.1b for a visualisation.

Row Column Slice Erasing (RCSE) RCSE combines RSE and CSE techniques. RSE and CSE are sequentially applied. RCSE is illustrated in Figure 6.1c. **Partially Saliency Erasing (PSE)** PSE divides the salient region into four parts and randomly erases one or more squares (Figure 6.1d). Mathematically, a binary mask M is split into four equal parts, with a random number of parts filled randomly with 0's or 1's. Element-wise multiplication generates the augmented image $\tilde{x}$ (Equation 6.1).

Horizontal Half Saliency Erasing (HHSE) HHSE horizontally divides the salient region into two parts and randomly erases one part (Figure 6.1e). The mask M is partitioned into two segments, with one filled with 0's and the other with 1's. This process generates the augmented image $\tilde{x}$ through element-wise multiplication (Equation 6.1).

Vertical Half Saliency Erasing (VHSE) VHSE vertically divides the salient region into two parts and randomly erases one part (Figure 4.4f). The mask M is split vertically into two equal-sized segments, with one filled with 0's and the other with 1's. This process generates the augmented image $\tilde{x}$ via element-wise multiplication (Equation 6.1).

- **FaceRandAug:**

  Given an input image I and a salient region x detected within I, FaceRandAug randomly selects one of the following erasing strategies from the given data augmentation list with equal probability: RSE, CSE, RCSE, PSE, HHSE and VHSE.

  Let S denote the selected erasing strategy. The augmented salient region $\tilde{x}$ is obtained as follows:

$$\tilde{x} = \begin{cases} \text{Erased Salient Region using } S, & \text{if } S \text{ is selected} \\ x, & \text{otherwise} \end{cases}$$

where $\tilde{x}$ represents the augmented salient region. The selection process of FaceRandAug ensures that each erasing strategy has an equal chance of being selected, similar to the searching process of RandAug [52]. This method provides a flexible way to incorporate various erasing strategies into the data augmentation pipeline for facial mask selection.

6.2.2 FaceKeepOriginalAugment

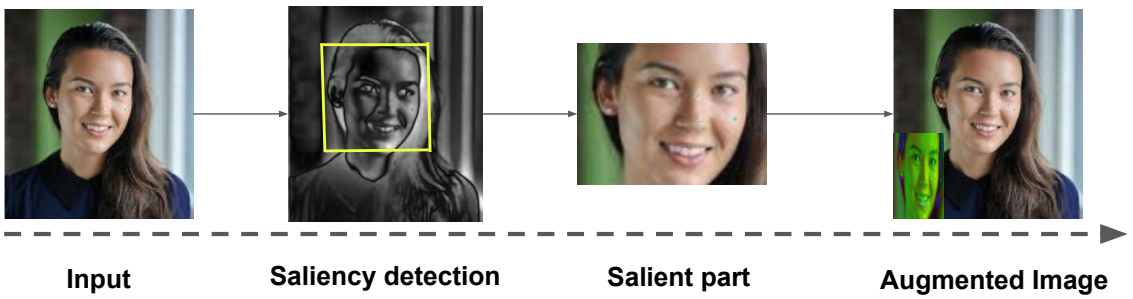


Figure 6.3: Overall architecture of the proposed approach

In this section, we present an extended approach of KeepOriginalAugment, as introduced in Chapter 5. This extended version, referred to as FaceKeepOriginalAugment, is explored in the context of biases using face regions.

Our primary objective is to enhance diversity by preserving both original and augmented information within a single image, thereby promoting fairness to encourage the model to learn diverse information. To achieve this, we begin by detecting the salient region in the image [1]. In our method, we utilise the saliency detection technique proposed by Montabone et al. [74], which has demonstrated superior performance compared to other methods [29]. After finding the salient region, we discuss two important questions i) Where to place the salient region? and ii) Which part should be augmented?.

[1] https://docs.opencv.org/3.4/da/dd0/classcv_1_1saliency_1_1StaticSaliencyFineGrained.html

By identifying the important region, we aim to address the issue of redundant salient features present in SalfMix, as well as tackle the domain shift problem encountered in KeepAugment.

To determine the placement of the salient region within the non-salient region and achieve diversity, we explore three different strategies:

- **Where to place the salient region?**

  Min-Area: We identify eight regions surrounding the salient region and select the one with the minimum area. The salient region is resized according to the size of that minimum area and placed within it, as shown in the output of the first row of Figure 6.4.

  Max-Area: Conversely, we choose the region with the maximum area among the eight surrounding regions and resize the salient region accordingly, as shown in the output of the second row of Figure 6.4.

  Random-Area: We adopt a more flexible approach by randomly selecting one of the eight regions. The salient region is resized based on the size of the chosen region and placed within it as shown in the output of the third row of Figure 6.4.

- **Which part should be augmented?**

  After identifying the salient region, we propose and investigate three distinct strategies to enhance diversity:

  Augment only salient: We solely apply random augmentation to the salient region and then paste the augmented salient region into the non-salient region of the original image as shown in output of first row of Figure 6.5.

  Augment non-salient only: We perform augmentation on the entire image while preserving the original salient region. The augmented image is then combined with the original salient region, extracted from the unaltered original image, as shown in the output of the second row of Figure 6.5.

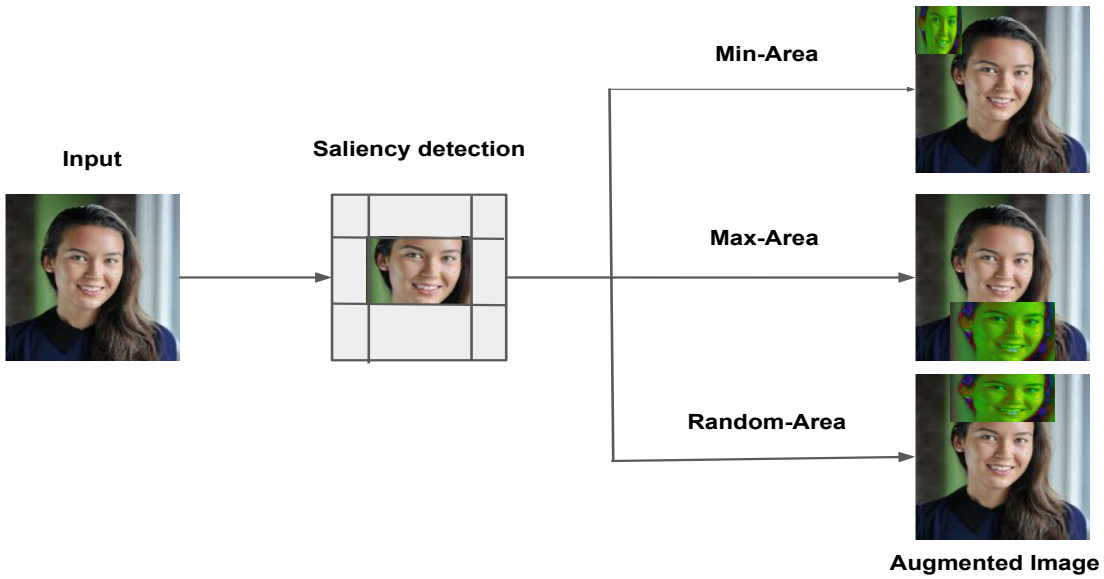


Figure 6.4: Where to place the salient region?

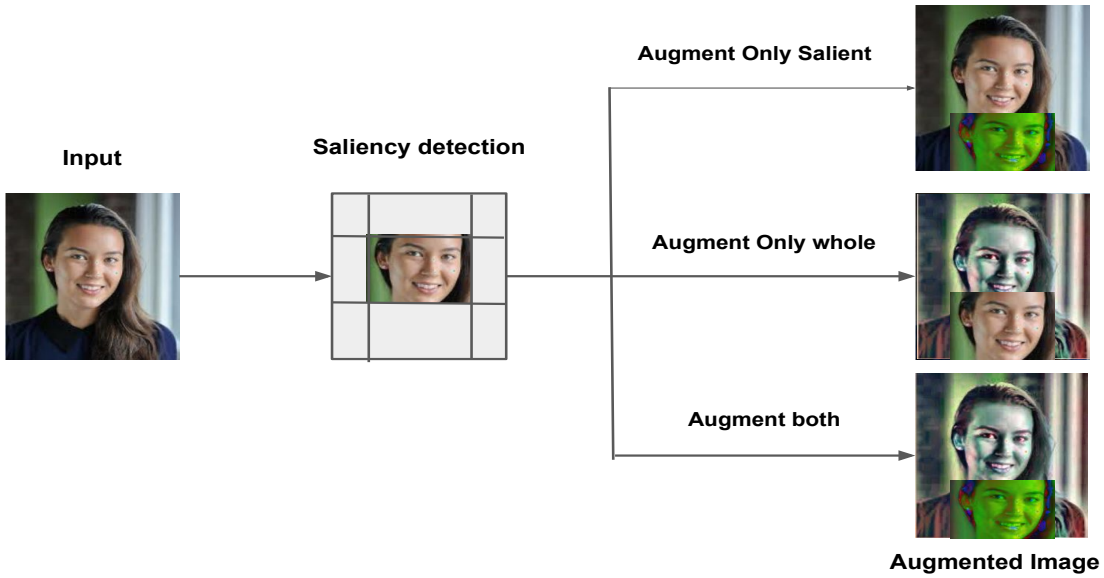


Figure 6.5: Which part should be augmented?

Augment both: This strategy involves performing separate augmentations on both the salient region and the entire image. The augmented salient region is integrated with the whole augmented image as shown in the output of the third row of Figure 6.5.

We observed that the augment both strategy demonstrates greater diversity across various computer vision tasks, detailed discussion is given in the hyper-parameter Section 6.3.4.

It is important to note that we utilise randAug [52] for augmentation, offering computational efficiency similar to that employed by KeepAugment.

6.2.3 Saliency-Based Diversity and Fairness Metric

Additionally, we propose the novel Saliency-Based Diversity and Fairness Metric. The metric is designed to account for both within-group diversity and inter-group fairness. Within-group diversity measures the variation among samples within a single group, capturing the richness of diversity in each class. Inter-group fairness evaluates the differences between groups, ensuring equitable representation. The proposed metric integrates both aspects, weighting them appropriately to handle imbalanced dataset while maintaining fairness and diversity across the groups. Mandal et al.[32] used cosine similarity, which measures the angle between two vectors, focusing on the direction rather than the magnitude. It works well for comparing the similarity between two feature vectors in terms of orientation, but for data diversity, we generally need to capture more than just the direction.

We refer to the feature vectors as X. First, we perform saliency detection on the images and then pass these saliency image to the pretrained VGG16 model [3], created by the Visual Geometry Group at the University of Oxford, to obtain the feature vector. Before calculating the diversity metrics, the feature vectors are normalised to ensure consistency and comparability across groups. The normalisation

of the feature vectors is performed as follows:

$$X' = \frac{X}{\|X\|} \tag{6.2}$$

where X' is the normalised feature vector and $\|X\|$ is its Euclidean norm. By normalising the feature vectors, we ensure that the diversity metrics are scale-invariant and comparable across different groups. The normalised feature vectors are then used to compute Euclidean distances. This ensures that both within-group and inter-group diversity measures are normalised and lie within a comparable range.

Within-Group Diversity

Let X'_i represent the set of normalised saliency-based features for group i (e.g., Male or Female). The within-group diversity for group i is computed using the Euclidean distance between pairs of feature vectors within that group:

$$D_{\text{within}}(X'_i) = \frac{1}{N_i(N_i-1)} \sum_{j=1}^{N_i} \sum_{k=j+1}^{N_i} \text{dist}(X'_{i,j}, X'_{i,k}) \tag{6.3}$$

where $\text{dist}(X'_{i,j}, X'_{i,k})$ is the Euclidean distance between two normalised feature vectors $X'_{i,j}$ and $X'_{i,k}$, and N_i is the number of feature vectors in group i.

Inter-Group Diversity and Fairness

The inter-group diversity measures the average pairwise distances between feature vectors from different groups (e.g., between Male and Female groups). This aspect is crucial for ensuring that the metric captures inter-group fairness. Let X'_i and X'_j represent the feature sets of two different groups. The inter-group diversity is computed as:

$$D_{\text{inter}}(X'_i, X'_j) = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \text{dist}(X'_{i,k}, X'_{j,l}) \tag{6.4}$$

where N_i and N_j are the number of feature vectors in groups i and j, respectively. The Euclidean distance between normalised feature vectors ensures that the inter-

group diversity reflects the actual distance between different groups.

Combined Metric: Fairness and Diversity

In scenarios where the dataset is imbalanced (i.e., the number of samples across groups varies significantly), it is important to ensure that larger groups do not dominate the overall diversity metric. To handle this, the metric weights the within-group and inter-group diversity terms by the size of each group, allowing the metric to reflect the actual contributions of both minority and majority groups.

The final metric, which combines both diversity within groups and diversity across groups, is given in equation. 6.5. The feature normalisation and the use of group sizes in weighting ensure that the final metric is balanced and reflects the contributions of both smaller and larger groups while also keeping the metric bounded between 0 and 1 (assuming both α and β are less than 0.5 individually).

$$M_{\text{fairness-diversity}} = \alpha \cdot \frac{1}{N} \sum_{i=1}^{K} N_i \cdot D_{\text{within}}(X'_i) + \beta \cdot \frac{1}{N(N-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} N_i \cdot N_j \cdot D_{\text{inter}}(X'_i, X'_j) \tag{6.5}$$

where: K is the total number of groups, N_i is the number of samples in group i, $N = \sum_{i=1}^{K} N_i$ is the total number of samples across all groups, α is the weight for within-group diversity (focusing on intra-group diversity), β is the weight for inter-group diversity (focusing on fairness across groups).

6.2.4 Additional Data Augmentations proposed for debiasing

In this section, we explore some additional methodologies. Initially, we employ facial recognition on the input image using the well-established and highly efficient face recognition algorithm, Single-shot Detection (SSD) [95]. To perform this task, we utilise a pre-trained model [2] and detect faces using OpenCV [3]. Once the facial region

[2] https://raw.githubusercontent.com/opencv/opencv_3rdparty/dnn_samples_face_detector_20180205_fp16/re
[3] https://docs.opencv.org/3.4/d6/d0f/group__dnn.html

has been successfully detected within the original image, x, we proceed to apply the newly proposed data augmentation techniques as follows:

1. **Partial Mixing (PM) :** In this approach, the facial regions x_m and x_f , corresponding to male and female subjects respectively, are considered.. Each is divided into four equal parts, and a random selection of squares from both facial regions is mixed. A mask, M, is partitioned into four segments, each filled with either 0's or 1's to include or exclude those squares, respectively. Subsequently, an element-wise multiplication is conducted between the mask, M, and the male facial region, x_m, and $1 - M$ and female facial region, x_f, then both are added, resulting in the generation of the augmented image, $\tilde{x_a}$, as illustrated in Equation 6.6. Finally, the augmented facial region $\tilde{x_a}$ is reinserted into the original image. The overall process is depicted in Figure 6.6.

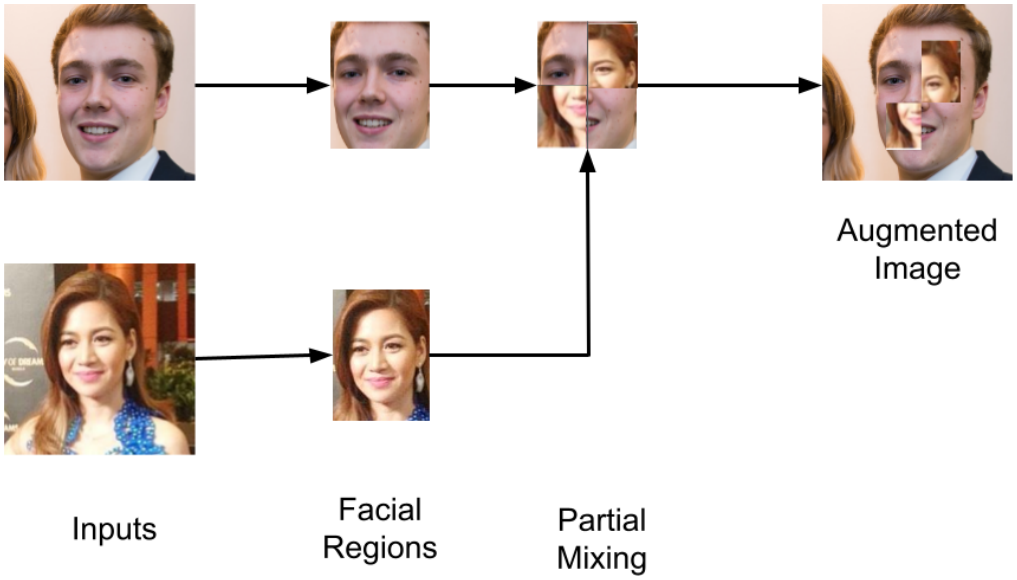


Figure 6.6: Partial Mixing Data Augmentation Process

$$\tilde{x_a} = M \odot x_m + (1 - M) \odot x_f \tag{6.6}$$

2. **Noise addition (NA):** In this strategy, we incorporate uniformly distributed noise, generated within the range of 0 to 1, as expressed in Equation 6.7. This

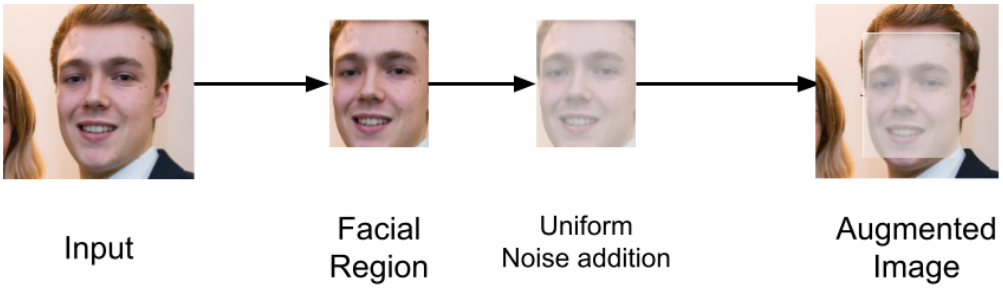


Figure 6.7: Noise Addition Data Augmentation Process

randomly generated noise, denoted as n_r, is then added to the facial region, x_m or x_f. Consequently, an augmented facial region, $\tilde{x}_a$, is produced, as outlined in Equation 6.8.

$$n_r = uniform(0, 1) \tag{6.7}$$

$$\tilde{x}_a = x_f + n_r \tag{6.8}$$

Then $\tilde{x}_a$ is placed back to its position in the original images. The overall process is shown in Figure 6.7.

6.3 Experiments

To evaluate data diversity and gender bias, we employ two approaches: FaceSaliencyAug and FaceKeepOriginalAugment. Both approaches are evaluated on five datasets: Flickr Faces HQ (FFHQ), WIKI, IMDB, Labelled Faces in the Wild (LFW), UTK Faces, and Diverse Dataset, and four occupation datasets are utilised, following the experimental settings from [96, 32]. An important term used in this chapter is the language–location pair [32], which is the combination of the language used for the query and the geographic region from which that query is issued. The world is divided into nine regions, each with an associated "lingua franca", giving pairs such as Arabic–NAWA, English–NA, English–WE, Hindi–SA, Indonesian–SEA, Mandarin–EA, Russian–EE, Spanish–LA, and Swahili–SSA. For each query term (CEO, Politician, Engineer, Nurse, School Teacher), they use the same nine language–location pairs and retrieve one image set per pair, where each image

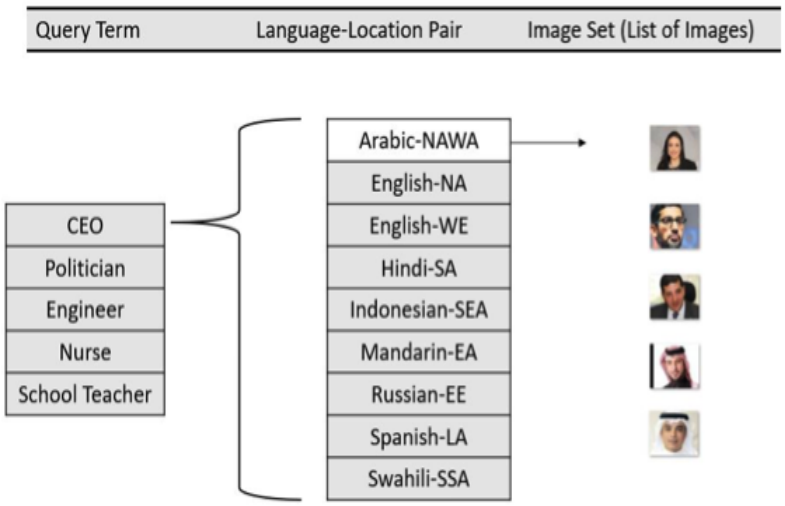


Figure 6.8: Structure of the image sets for each query term and language–location pair.

set is a list of 20 images. Thus, every query term is associated with nine image sets (one for each language–location pair), and Figure 6.8 illustrates this arrangement for the query term "CEO" and the pair "Arabic–NAWA"

6.3.1 Data Diversity Evaluation

We use two variants of the Image Similarity Score (ISS), namely ISS_{Intra} and ISS_{Cross}, to measure diversity. ISS$Intra$ quantifies intra-dataset diversity, whereas ISS_{Cross} evaluates diversity across different datasets [32]. A higher diversity score indicates reduced bias [32].

For two images, I_1 and I_2, with extracted features ν_1 and ν_2 respectively, the image similarity [32] is calculated as:

$$\begin{gathered} \text{sim}\,(I_1, I_2) = 1 - \frac{v_1 \cdot v_2}{\|v_1\|_2 \cdot \|v_2\|_2} \\ \text{sim}\,(I_1, I_2) \in [0, 2] \end{gathered} \tag{6.9}$$

An image similarity score of 0 indicates that the two images are identical or highly similar, with a cosine similarity of 1 ($\theta = 0$). On the other hand, a similarity score of 2 corresponds to visually opposite images, where the cosine similarity is -1 ($\theta = \pi$). In this case, the images are maximally dissimilar.

Image–Image Association Score (IIAS). Let A and B be two sets of images representing two gender categories (e.g., male and female), and let W be a set of target images (e.g., images generated for a particular occupation). For each target image $w \in W$, we compute

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b), \tag{6.10}$$

where $\cos(\cdot, \cdot)$ is the cosine similarity between the image embeddings. The Image–Image Association Score (IIAS) is then defined as

$$\text{IIAS} = \text{mean}_{w \in W}\, s(w, A, B). \tag{6.11}$$

A positive IIAS indicates that the target images are, on average, more similar to the attribute images in A than those in B, while a negative IIAS indicates the opposite direction of association, as this metric is introduced by this work [55].

6.3.2 Gender Bias Evaluation

To assess gender bias in Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), we utilise the Image-Image Association Score (IIAS) [97, 96]. IIAS computes bias based on gender-representing attributes and images depicting specific concepts.

Our evaluation considers four CNN models (VGG16, ResNet152, Inceptionv3, and Xception) and four ViT models (ViT B/16, B/32, L/16, and L/32). Training settings align with [33]: for CNNs, layers are frozen initially, followed by custom layer training for 50 epochs, then partial unfreezing and training for another 50 epochs. For ViTs, layers are frozen and trained for 100 epochs, then fully trained for 50 epochs at a lower learning rate.

A gender bias study from [33, 96] collected images from Google searches using job terms, forming two training sets: one balanced and one biased. The test set remains balanced. The training set consists of 7,200 images (900 per category), while the test

set comprises 1,200 images (300 per category, 150 per gender). Separate datasets for men and women are also used for evaluation.

Table 6.1: ISS_{intra} of Datasets and baseline result originate from [32].

Dataset	Baseline	RSMDA [82]	FSA [98]
FFHQ [39]	0.9940	0.9935	**0.9939**
Diverse Dataset [32]	0.9895	0.9900	**0.9905**
WIKI [99]	0.9786	0.9800	**0.9832**
IMDB [99]	0.9661	0.9700	**0.9717**
LFW [100]	0.9536	**0.9550**	0.9539
UTK [101]	0.9418	0.9440	**0.9481**

6.3.3 Results for FaceSaliencyAug

We performed experiments on two tasks: measuring the data diversity on five datasets and evaluating the diversity of genders on four occupation data sets. For data set diversity, we calculated the intra-dataset image similarity score (ISS_{intra}), demonstrating improved diversity in all data sets (Table 6.1), and the proposed also shown superior performance than RSMDA [82] except for the LFW data set. For a more detailed analysis of query combinations with language-location pairs, Table 6.2 presents ISS_{intra} for various queries in different language-location pairs. We compare the baseline ISS (intra and across) values with our proposed approach. The queries include CEO, Engineer, Nurse, Politician, and School Teacher, each evaluated with multiple language-location pairs. Overall, our proposed approach demonstrates higher diversity, as measured by ISS scores, as shown in Table 6.3. Additionally, we also analysed Inter-dataset Image Similarity Score (ISS_{inter}) and ISS_{intra} for five occupation datasets. Our approach generally showed greater diversity, except for the school teacher dataset as shown in Table 6.3, potentially due to a female bias identified by Mandal et al. (2023) [33]. Specifically, our approach consistently outperformed the baseline for "CEO" and "Engineer" occupations in both ISS_{intra} and ISS_{cross}. In additional, across various datasets, our approach achieved slightly better results compared to the baseline. Moreover, our approach yielded

Table 6.2: Image Similarity score across all possible queries, baseline results are taken from [32] and where FSA is FaceSaliencyAug.

Query	Language Location Pair	$\mathbf{ISS}_{Intra}$		$\mathbf{ISS}_{cross}$	
		Baseline	**FSA [98]**	**Baseline**	**FSA [98]**
CEO	Arabic-West Asia & North Africa	0.899012	0.906985	0.984683	0.987136
	English-North America	0.968974	0.976045		
	English-West Europe	0.929469	0.957394		
	Hindi-South Asia	0.997845	0.994985		
	Indonesian-SE Asia	0.983675	0.985087		
	Mandarin-East Asia	0.989452	0.994088		
	Russian-East Europe	0.959661	0.964493		
	Spanish-Latin America	0.974743	0.974597		
	Swahili-Sub Saharan Africa	0.977119	0.975551		
Engineer	Arabic-West Asia & North Africa	0.98639	0.982757	0.993904	0.994201
	English-North America	0.988344	0.991277		
	English-West Europe	1.000911	1.001757		
	Hindi-South Asia	1.003149	0.994307		
	Indonesian-SE Asia	0.987191	0.990045		
	Mandarin-East Asia	0.991146	0.987163		
	Russian-East Europe	1.007155	1.000498		
	Spanish-Latin America	0.984955	0.986797		
	Swahili-Sub Saharan Africa	0.983727	0.986656		
Nurse	Arabic-West Asia & North Africa	1.002607	0.995654	0.989952	0.990378
	English-North America	0.971564	0.973874		
	English-West Europe	0.99561	0.992313		
	Hindi-South Asia	0.984535	0.987963		
	Indonesian-SE Asia	0.975914	0.979673		
	Mandarin-East Asia	0.98904	0.993333		
	Russian-East Europe	0.997979	0.996215		
	Spanish-Latin America	1.000587	0.993549		
	Swahili-Sub Saharan Africa	0.958532	0.972825		
Politician	Arabic-West Asia & North Africa	0.977348	0.979589	0.983637	0.987169
	English-North America	0.995927	0.999067		
	English-West Europe	0.979358	0.981664		
	Hindi-South Asia	0.979915	0.980589		
	Indonesian-SE Asia	0.972307	0.992733		
	Mandarin-East Asia	0.976251	0.973902		
	Russian-East Europe	0.93835	0.965788		
	Spanish-Latin America	0.988452	0.989963		
	Swahili-Sub Saharan Africa	0.943626	0.952849		
School Teacher	Arabic-West Asia & North Africa	1.014298	1.013607	0.990403	0.989746
	English-North America	0.997715	0.988912		
	English-West Europe	0.940142	0.956337		
	Hindi-South Asia	1.000047	0.995507		
	Indonesian-SE Asia	0.985991	0.990466		
	Mandarin-East Asia	1.00862	1.004755		
	Russian-East Europe	0.976169	0.962497		
	Spanish-Latin America	0.965902	0.961019		
	Swahili-Sub Saharan Africa	0.985919	0.990466		

Table 6.3: Image Similarity score across all possible queries. Baseline results are taken from [32].

	ISS_{intra}		ISS_{cross}	
Query	Baseline	FSA [98]	Baseline	FSA [98]
CEO	0.9644	**0.9699**	0.9846	**0.9871**
Engineer	**0.9925**	0.9913	0.9939	**0.9942**
Nurse	0.9862	**0.9873**	0.990	**0.9904**
Politician	0.9724	**0.9796**	0.9836	0.9872
School Teacher	**0.9860**	0.9848	**0.9904**	0.9897
Mean Value	0.9803	**0.9826**	0.9885	**0.9897**

higher mean ISS scores across all queries, indicating its effectiveness in enhancing diversity. Overall, our approach achieved impressive diversity scores. In addition,

Table 6.4: Average Image-Image Association Scores (IIAS) for CNNs and ViTs. Positive values indicate bias towards men, negative towards women. Total absolute IIAS reflects bias magnitude. Our approach reduces gender bias, highlighted in red. Baseline result and dataset were taken from [96, 33], where FSA is FaceSaliencyAug in this table.

Masked								
Class	Biased				Unbiased			
	CNN Baseline	CNN FSA [98]	ViT Baseline	ViT FSA [98]	CNN Baseline	CNN FSA [98]	ViT Baseline	ViT FSA [98]
CEO	0.059	0.005	0.1	0.010	0.26	0.007	0.02	0.007
Engineer	0.23	0.018	0.14	0.019	0.36	0.005	0.17	0.021
Nurse	-0.14	-0.021	-0.35	-0.0214	-0.05	-0.007	-0.2	-0.018
School Teacher	-0.17	-0.004	-0.15	-0.027	-0.12	-0.004	-0.05	-0.013
Total IIAS (abs)	0.599	0.048	0.74	0.077	0.79	0.023	0.44	0.059
Bias reduction	∼ 12 times ↓		∼ 10 times ↓		∼ 34 times↓		∼ 8 times↓	
Unmasked								
Class	Biased				Unbiased			
	CNN Baseline	CNN FSA [98]	ViT Baseline	ViT FSA [98]	CNN Baseline	CNN FSA [98]	ViT Baseline	ViT FSA [98]
CEO	0.050	0.022	0.17	0.004	0.07	0.003	0.06	0.002
Engineer	0.180	0.028	0.19	0.008	0.04	0.016	0.21	0.003
Nurse	-0.21	-0.054	-0.21	-0.008	-0.06	-0.002	-0.17	-0.005
School Teacher	-0.02	-0.025	-0.4	-0.006	-0.04	-0.0002	-0.14	-0.001
Total IIAS (abs)	0.46	0.129	0.97	0.026	0.21	0.0212	0.58	0.011
Bias reduction	∼ 4 times ↓		∼ 37 times↓		∼ 10 times↓		∼ 53 times↓	

we assessed gender bias using the Image-Image Association Scores (IIAS) on both masked (obscured) and unmasked (non-obscured) data, utilising different CNNs and

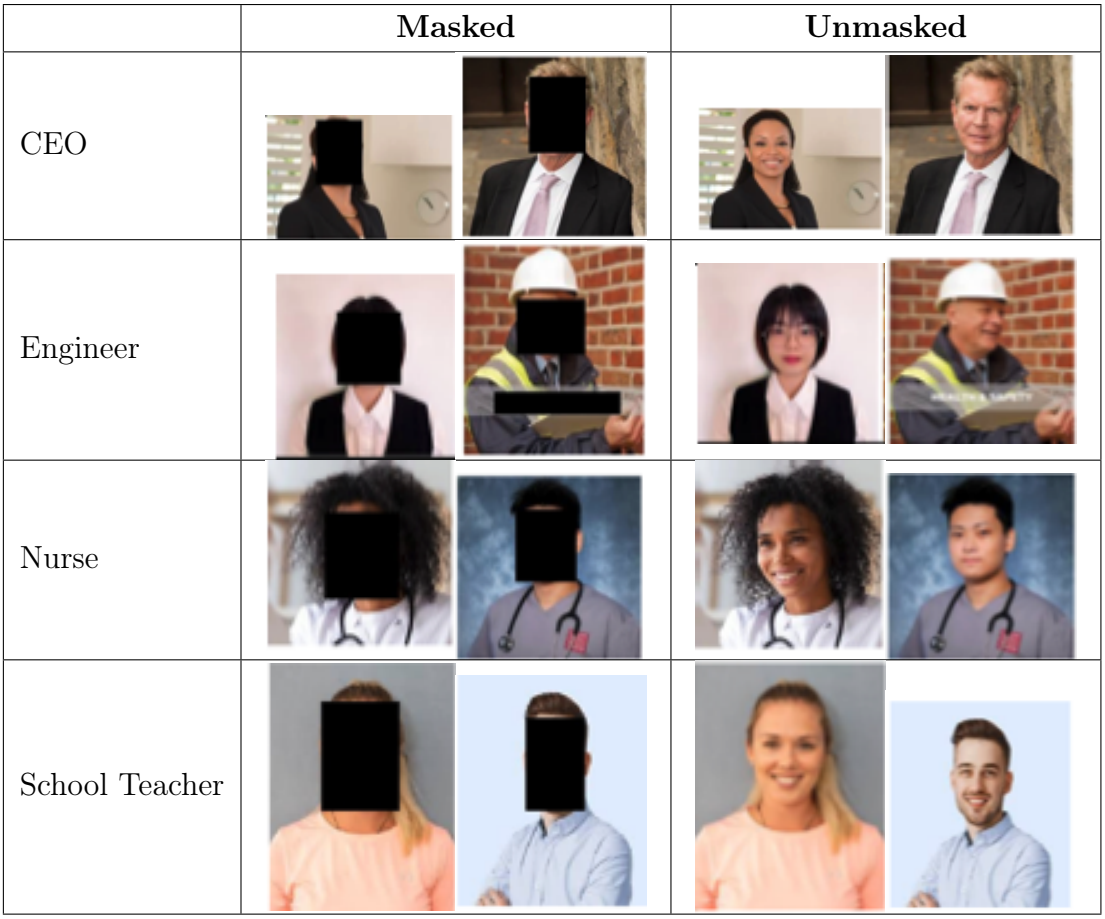


Figure 6.9: Examples of masked and unmasked images for each occupation. Masked images have the facial region covered with a solid black rectangle, removing identifiable facial attributes, while unmasked images show the original, non-obscured faces.

ViTs. Following Mandal et al. [32], we define masked images as those in which the facial region is covered with a solid black rectangle to hide identifiable facial attributes, whereas unmasked images retain the original faces (see Figure 6.9). Our proposed approach consistently reduced bias across all scenarios, with particularly significant reductions observed in the unmasked dataset. Notably, when the data was well-balanced (unbiased) in terms of gender, our approach achieved an approximately 53-fold reduction in bias. Furthermore, our approach demonstrated superior performance across all classes compared to the baseline, as indicated by the larger reductions in bias for the "FSA" columns compared to the "Baseline" columns. In addition, while our approach exhibited substantial reductions in bias across all occupations compared to the baseline, the magnitude of reduction varied across occupations and data types. For instance, the reduction in bias for the "Nurse"

occupation was more pronounced in the unmasked dataset compared to the "School Teacher" occupation. Furthermore, our approach achieved particularly pronounced bias reduction for ViTs compared to CNNs, highlighting its effectiveness across different model architectures. While the reduction in bias was slightly less pronounced in the masked dataset compared to the unmasked dataset, our approach still exhibited substantial improvements. Overall, our approach showed highly effective in reducing gender bias, as highlighted in red in Table 6.4.

In Figure 6.10, the results reveal a clear difference in bias reduction between models for the masked scenario, where gender-specific features are hidden. For occupations such as CEO and Engineer, the Baseline CNN (Masked Biased) and Baseline ViT (Masked Biased) models display positive IIAS scores, indicating the presence of gender bias. The baseline ViT (Masked Biased) model, in particular, exhibits higher bias in these occupations, with significantly positive scores compared to the CNN models. When applying our augmentation method, there is a noticeable reduction in bias. The CNN + Our Augmentation (Masked Biased) and ViT + Our Augmentation (Masked Biased) models show lower IIAS scores, indicating that the proposed augmentation strategy effectively reduces the bias in these occupations. Specifically, the ViT + Our Augmentation (Masked Biased) model significantly outperforms the Baseline ViT, reducing bias in both CEO and Engineer occupations. For the Nurse and School Teacher occupations, which typically reflect higher stereotypical associations, the bias is much more pronounced in the Baseline ViT (Masked Biased) model. However, with our augmentation, particularly in the ViT + Our Augmentation (Masked Biased) model, the bias is substantially reduced, bringing the IIAS scores closer to zero. This demonstrates that our augmentation method is effective in masking gendered features and reducing stereotypical associations in these biased roles. The CNN + Our Augmentation (Masked Biased) model similarly reduces bias, though the effect is more moderate compared to the ViT-based model.

In Figure 6.11, where the unmasked scenario allows for gender-specific visual

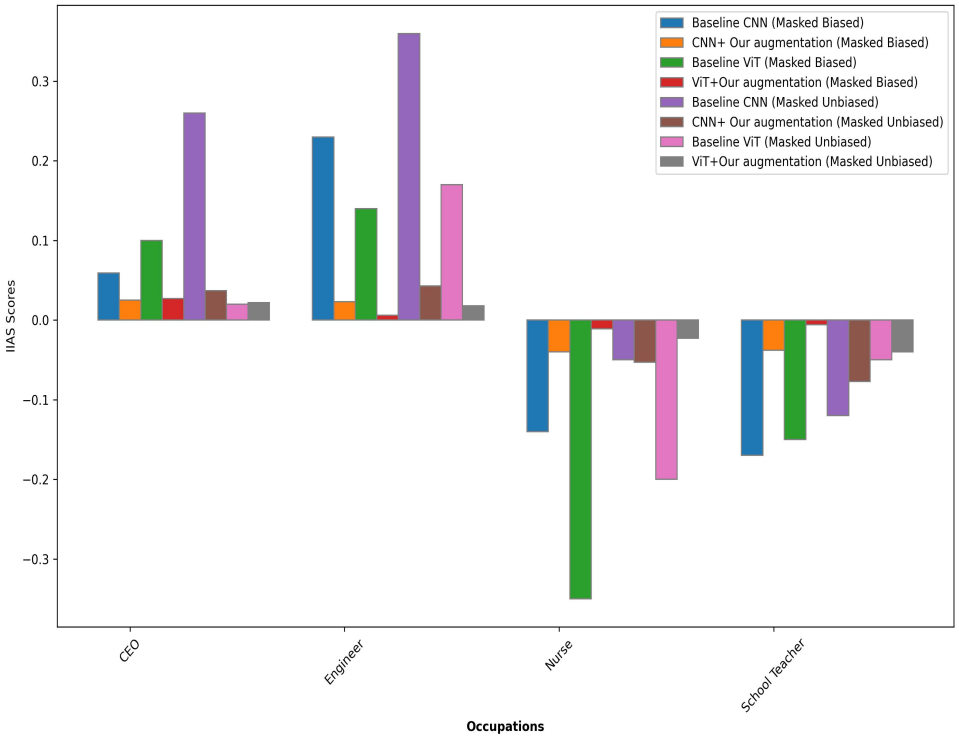


Figure 6.10: Comparison of our approach for gender bias reduction in CNN and ViT- Masked Scienario.

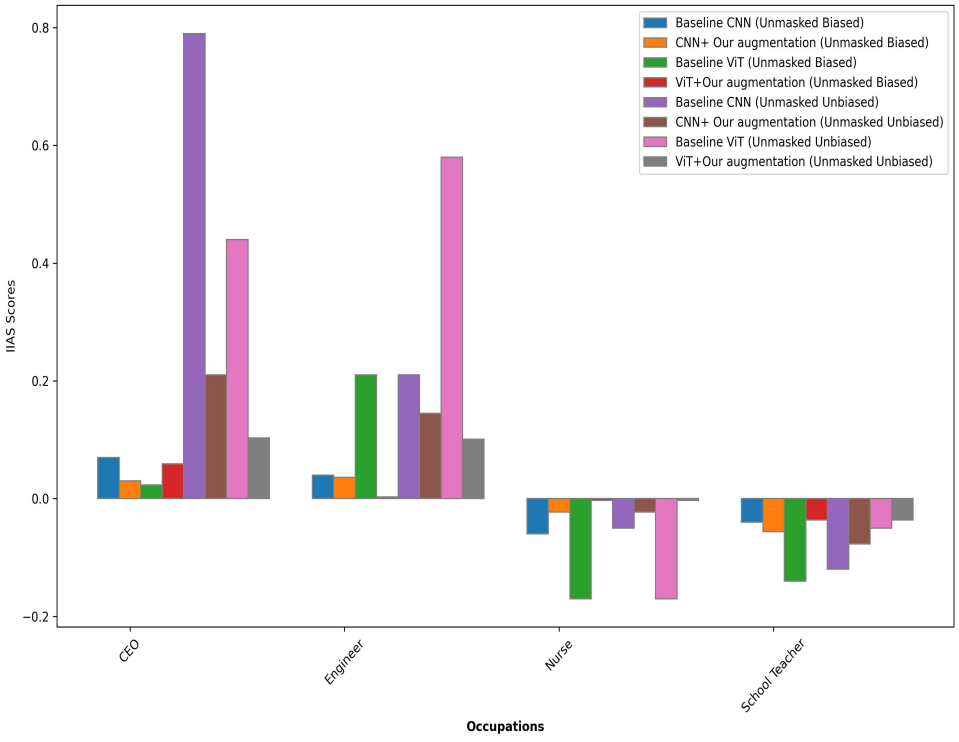


Figure 6.11: Comparison of our approach for gender bias reduction in CNN and ViT- Unmasked Scienario.

cues, the IIAS scores across all occupations are generally higher, reflecting stronger gender biases when the models have access to these cues. For CEO and Engineer, the Baseline CNN (Unmasked Biased) and Baseline ViT (Unmasked Biased) models show significant positive scores, indicating that bias is more evident when gender-related features are not masked. Once again, our augmentation method shows substantial improvements. The CNN + Our Augmentation (Unmasked Biased) and ViT + Our Augmentation (Unmasked Biased) models reduce the biases for CEO and Engineer, with ViT + Our Augmentation (Unmasked Biased) being particularly effective in lowering the IIAS scores, though the scores are higher than in the masked scenario. This indicates that while the augmentation reduces bias, the effect is slightly diminished in the unmasked scenario compared to the masked one. For the Nurse and School Teacher occupations, the Baseline ViT (Unmasked Biased) model shows the highest bias, with very negative IIAS scores, reflecting strong stereotypical associations. However, the ViT + Our Augmentation (Unmasked Biased) model effectively reduces these biases, significantly lowering the IIAS scores. The CNN + Our Augmentation (Unmasked Biased) model also reduces bias but does not perform as well as the ViT-based model. Overall, these results suggest that the augmentation method is effective in reducing bias in the unmasked scenario, though masking gender-specific features (as seen in Figure 6.10) yields stronger bias mitigation.

6.3.4 Results for FaceKeepOriginalAugment

Hyperparameter

We conducted a comparative hyperparameters experiments to evaluate the algorithm's performance across five professional datasets—CEO, Engineer, Nurse, Politician, and School Teacher—using two data diversity metrics: ISS_{Intra} and ISS_{Cross}. The methodology involved systematically testing all combinations of augmentation strategies (*only original*, *only salient*, *both*) and placement strategies (*max*, *min*, *random*) across these datasets to identify optimal parameter settings.

The results, presented in Table 6.5, demonstrate that the *both strategy* paired with the *random area strategy* consistently yielded the highest scores for ISS*intra* and ISS*cross* across all professions. This combination effectively balances scaling data augmentation and enhanced diversity, as highlighted in previous studies [79]. The findings underscore the robustness of the *both strategy* and the flexibility of the *random area strategy* in addressing diverse datasets.

Table 6.5: Comparison of strategies across professions for finding optimal hyperparameters - augmentation strategy and area strategy

Profession		ISS_{intra}			ISS_{cross}		
	Part strategy	**Placement strategy**			**Placement strategy**		
		max	**min**	**random**	**max**	**min**	**random**
CEO	Only Original	0.9928	0.9935	0.9887	0.9932	0.9959	0.9945
	Only Salient	0.9933	0.9913	0.9915	0.9960	0.9935	0.9926
	Both	0.9901	0.9928	**1.0701**	0.9946	0.9949	**1.0809**
Nurse	Only Original	0.9944	0.9947	0.9943	0.9969	0.9953	0.9974
	Only Salient	0.9959	0.9955	0.9932	0.9948	0.9945	0.9947
	Both	0.9943	0.9934	**1.0521**	0.9943	0.9944	**1.0057**
Engineer	Only Original	0.9960	0.9970	0.9999	0.9982	0.9974	0.9965
	Only Salient	1.0000	0.9921	0.9999	0.9958	0.9972	0.9965
	Both	0.9964	0.9971	**1.0170**	0.9978	0.9963	**1.0304**
School Teacher	Only Original	0.9963	0.9934	0.9968	0.9971	0.9968	0.9968
	Only Salient	0.9946	0.9974	0.9967	0.9972	0.9961	0.9981
	Both	0.9948	0.9969	**1.0558**	0.9962	0.9977	**1.0552**
Politician	Only Original	0.9952	0.9963	0.9927	0.9981	0.9952	0.9942
	Only Salient	0.9943	0.9940	0.9930	0.9963	0.9947	0.9962
	Both	0.9939	0.9943	**1.0832**	0.9967	0.9959	**1.0353**

Why these algorithm work?

Our hyperparameter experiments, shown in Table 6.5, and previous work [79], suggest that the salient region should be randomly placed within any non-salient region (the *random-placement strategy*, as in Table 6.5). This random placement ensures scaling data augmentation across different regions, promoting diversity in training. For example, the salient region may vary in size between images, providing varied scale data augmentation.

Additionally, both the salient region and the entire image should be augmented (the *Both-Part strategy*, as in Table 6.5). This enables the model to learn from both

original and augmented data. To balance this, we used a 0.5 probability, giving the model equal exposure to both types, as explored in random erasing [18]. This approach enhances feature diversity while maintaining data distribution integrity.

Table 6.6: ISS_{intra} of Datasets and baseline result originate from [32], FKOA is FaceKeepOriginalAugment.

Dataset	**Baseline**	**SalfMix**	**KeepAugment**	**FKOA [102]**
FFHQ [39]	0.9940	0.9944	0.9946	**0.9951**
Diverse Dataset [32]	0.9895	0.9900	0.9918	**0.9965**
WIKI [99]	0.9786	0.9825	0.9867	**0.9980**
IMDB [99]	0.9661	0.9721	0.9780	**0.9971**
LFW [100]	0.9536	0.9622	0.9704	**0.9956**
UTK [101]	0.9418	0.9530	0.9603	**0.9903**

Table 6.7: Image Similarity score across all possible queries, baseline results are taken from [32] , FKOA is FaceKeepOriginalAugment

Language Location	**CEO**		**Engineer**		**Nurse**	
	Baseline	**FKOA [102]**	**Baseline**	**FKOA [102]**	**Baseline**	**FKOA [102]**
Arabic-West Asia & North Africa	0.899012	**0.9988**	0.98639	**0.9959**	**1.002607**	0.9916
English-North America	0.968974	**0.9985**	0.988344	**1.0031**	0.971564	**0.9933**
English-West Europe	0.929469	**0.9964**	1.000911	1.0004	0.99561	**0.9982**
Hindi-South Asia	**0.997845**	0.9898	**1.003149**	0.9966	0.984535	**0.9968**
Indonesian-SE Asia	**0.983675**	0.9912	**0.987191**	0.9897	0.975914	**0.9937**
Mandarin-East Asia	**0.989452**	0.9927	**0.991146**	0.9923	0.98904	**0.9983**
Russian-East Europe	0.959661	**0.9957**	**1.007155**	0.9976	**0.997979**	0.9940
Spanish-Latin America	0.974743	**0.9929**	0.984955	**0.9992**	**1.000587**	0.9964
Swahili-Sub Saharan Africa	0.977119	**0.9936**	0.983727	**0.9957**	0.958532	**0.9950**
Language Location	**Politician**		**School Teacher**			
	Baseline	**FKOA [102]**	**Baseline**	**FKOA [102]**		
Arabic-West Asia & North Africa	0.977348	**0.9951**	1.014298	1.0000		
English-North America	0.995927	**0.9987**	0.997715	0.9978		
English-West Europe	0.979358	**0.9968**	0.940142	**0.9994**		
Hindi-South Asia	0.979915	**0.9936**	1.000047	0.9972		
Indonesian-SE Asia	0.972307	**0.9921**	0.985991	**1.0017**		
Mandarin-East Asia	0.976251	**0.9957**	1.00862	0.9984		
Russian-East Europe	0.93835	**0.9992**	0.976169	**0.9952**		
Spanish-Latin America	0.988452	**0.9943**	0.965902	**0.9980**		
Swahili-Sub Saharan Africa	0.943626	**0.9983**	0.985919	**1.0033**		

To assess dataset diversity, we employ the Intra-dataset Image Similarity Score (ISS_{intra}) across various datasets, showcasing substantial improvements compared to the baseline as shown in Table 6.6. Note that all experiments were repeated five times, unless otherwise specified. To anaylse further, the Table 6.7 presents the ISS_{intra} for various queries across different language-location pairs. The baseline ISS values are compared with ours. The queries include CEO, Engineer, Nurse, Politician, and School Teacher, each with multiple language-location pairs. It is observed

that the ISS values for most queries and language-location pairs have improved compared to the baseline values. This improvement indicates the effectiveness of the proposed method in enhancing image similarity in cross-cultural contexts. Notably, certain queries show more significant improvements, such as CEO and School Teacher in Arabic-West Asia & North Africa, Mandarin-East Asia, and Spanish-Latin America language locations as shown in Table 6.7.

Table 6.8: Image Similarity Score across all queries,FKOA is FaceKeepOriginalAugment

Query	ISS_{intra}				ISS_{cross}			
	Baseline [32]	SalfMix [30]	KeepAugment [31]	FKOA [102]	Baseline [32]	SalfMix [30]	KeepAugment [31]	FKOA [102]
CEO	0.9644	0.9750	0.9800	**0.9944**	0.9846	0.9905	0.9920	**0.9956**
Engineer	0.9925	0.9938	0.9945	**0.9967**	0.9939	0.9950	0.9960	**0.9972**
Nurse	0.9862	0.9900	0.9925	**0.9952**	0.9900	0.9928	0.9950	**0.9961**
Politician	0.9724	0.9810	0.9855	**0.9960**	0.9836	0.9885	0.9910	**0.9964**
School Teacher	0.9860	0.9920	0.9938	**0.9990**	0.9904	0.9945	0.9955	**0.9977**
Mean Value	0.9803	0.9864	0.9892	**0.9963**	0.9885	0.9923	0.9940	**0.9966**

Table 6.9: Average Image-Image Association Scores (IIAS) for CNNs and ViTs (Positive values indicate bias towards men, negative towards women. Total absolute IIAS reflects bias magnitude. Baseline results from [96]). All the experiments were repeated three times and FKOA results are from [102], FKOA is FaceKeepOriginalAugment

Masked								
Class	Biased				Unbiased			
	$CNN_{Baseline}$	CNN_{FKOA}	$ViT_{Baseline}$	ViT_{FKOA}	$CNN_{Baseline}$	CNN_{FKOA}	$ViT_{Baseline}$	ViT_{FKOA}
CEO	0.059	**0.025**	0.1	**0.027**	0.26	**0.037**	0.02	**0.022**
Engineer	0.23	**0.023**	0.14	**0.006**	0.36	**0.043**	0.17	**0.018**
Nurse	−0.14	**-0.040**	−0.35	**-0.011**	−0.05	**-0.053**	−0.2	**-0.023**
School Teacher	−0.17	**-0.038**	−0.15	**-0.006**	−0.12	**-0.077**	−0.05	**-0.04**
Total IIAS (abs)	0.599	**0.126**	0.74	**0.05**	0.79	**0.21**	0.44	**0.103**
Bias reduction	∼ 5 times ↓		∼ 15 times ↓		∼ 4 times ↓		∼ 4 times ↓	
Unmasked								
Class	Biased				Unbiased			
	$CNN_{Baseline}$	CNN_{FKOA}	$ViT_{Baseline}$	ViT_{FKOA}	$CNN_{Baseline}$	CNN_{FKOA}	$ViT_{Baseline}$	ViT_{FKOA}
CEO	0.050	**0.023**	0.17	**0.003**	0.07	**0.03**	0.023	**0.059**
Engineer	0.180	**0.016**	0.19	**0.008**	0.04	**0.036**	0.21	**0.003**
Nurse	−0.21	**-0.039**	−0.21	**-0.035**	−0.06	**-0.023**	−0.17	**-0.003**
School Teacher	−0.02	**-0.035**	−0.4	**-0.001**	−0.04	**-0.056**	−0.14	**-0.036**
Total IIAS (abs)	0.46	**0.113**	0.97	**0.047**	0.21	**0.145**	0.58	**0.101**
Bias reduction	∼ 4 times ↓		∼ 21 times ↓		∼ 2 times ↓		∼ 6 times ↓	

Furthermore, we investigate the Inter-dataset Image Similarity Score (ISS_{cross}) and ISS_{intra} for five occupation datasets. Our approach consistently demonstrates enhanced diversity, with notable exceptions observed in the school teacher dataset, potentially influenced by underlying biases identified by Mandal et al. (2023) [33].

Specifically, our method outperforms the baseline across different datasets, notably achieving significant improvements for occupations such as "Politician" and "Nurse". While both approaches exhibit comparable performance for "Politician" in ISS_{cross}, our method showcases superior results across different occupations, emphasising its effectiveness in promoting dataset diversity as shown in Table 6.8. Moreover, our approach yields higher mean ISS scores across all queries, highlighting its efficacy in enhancing diversity. Overall, our approach presents promising advancements in dataset diversity assessment as shown in Table 6.8.

Table 6.9 presents the average Image-Image Association Scores (IIAS) for Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) across various classes and masking scenarios. Positive IIAS values indicate biases towards men, while negative values signify biases towards women. Notably, in the biased dataset, biases towards men are observed for the CEO and Engineer classes, as evidenced by positive IIAS values, while biases towards women are observed for the Nurse and School Teacher classes, indicated by negative IIAS values. This nuanced analysis reveals the gender biases inherent in the dataset and underscores the importance of addressing these biases in computer vision models. FaceKeepOriginalAugment approach consistently demonstrates a remarkable reduction in gender bias compared to baseline models. Specifically, our method achieves reductions of approximately 5 to 15 times for CNNs and ViTs in masked scenarios as shown in Figure 6.12, and 4 to 21 times in unmasked scenarios as shown in Figure 6.13 across various occupation classes. These substantial reductions, highlighted in red, underscore the effectiveness of our approach in promoting fairness and inclusivity within computer vision models. Moreover, it's crucial to note the total absolute IIAS values, which reflect the overall magnitude of gender bias within the models. Our approach consistently yields lower total absolute IIAS values compared to baseline results taken from [96], indicating a substantial reduction in the overall magnitude of gender bias. This comprehensive view of bias reduction further reinforces the robustness and efficacy of FaceKeepOriginalAugment in mitigating gender bias within computer

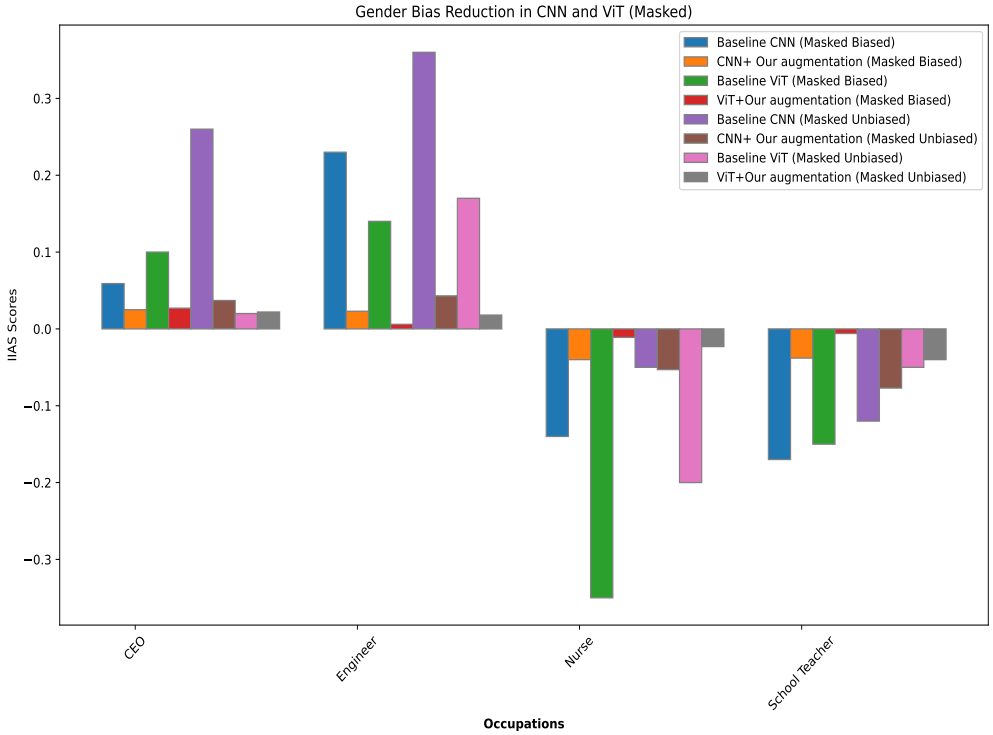


Figure 6.12: Comparison of our approach for gender bias reduction in CNN and ViT- Masked Scienario.

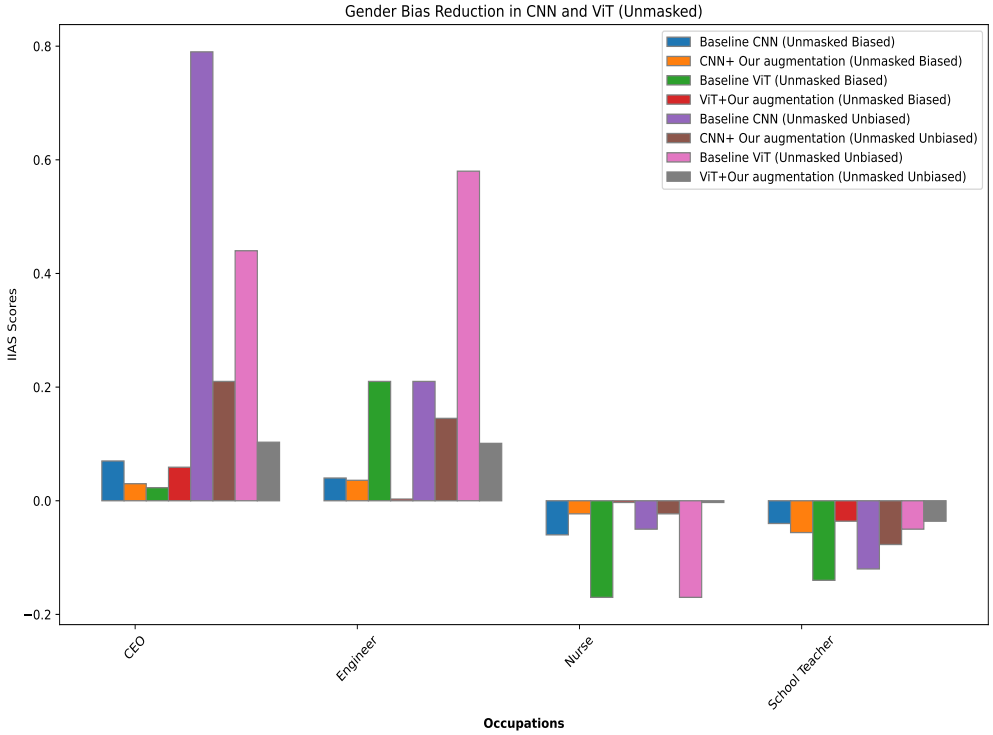


Figure 6.13: Comparison of our approach for gender bias reduction in CNN and ViT- Unmasked Scienario.

vision models. Our findings not only highlight the nuanced gender biases present across different occupation classes but also demonstrate the significant effectiveness of our approach in addressing these biases. By reducing gender bias and promoting fairness within computer vision models, our FaceKeepOriginalAugment approach contributes towards building more bias-free system.

6.3.5 Saliency-Based Diversity and Fairness Metric Evaluation

To evaluate the proposed Saliency-Based Diversity and Fairness Metric, we conducted experiments on six datasets commonly used for assessing data diversity in images. These datasets include the Diverse Dataset [32], FFHQ [39], WIKI [99], IMDB [99], LFW [100], and UTK [101]. The DiverseDatasetconsists of only 81 images, while we randomly selected 100 images from each of the remaining five datasets. The images were manually classified into male and female groups. To handle the inherent class imbalance, we created balanced versions of the datasets through an undersampling data augmentation technique, as illustrated in Figure 6.14.

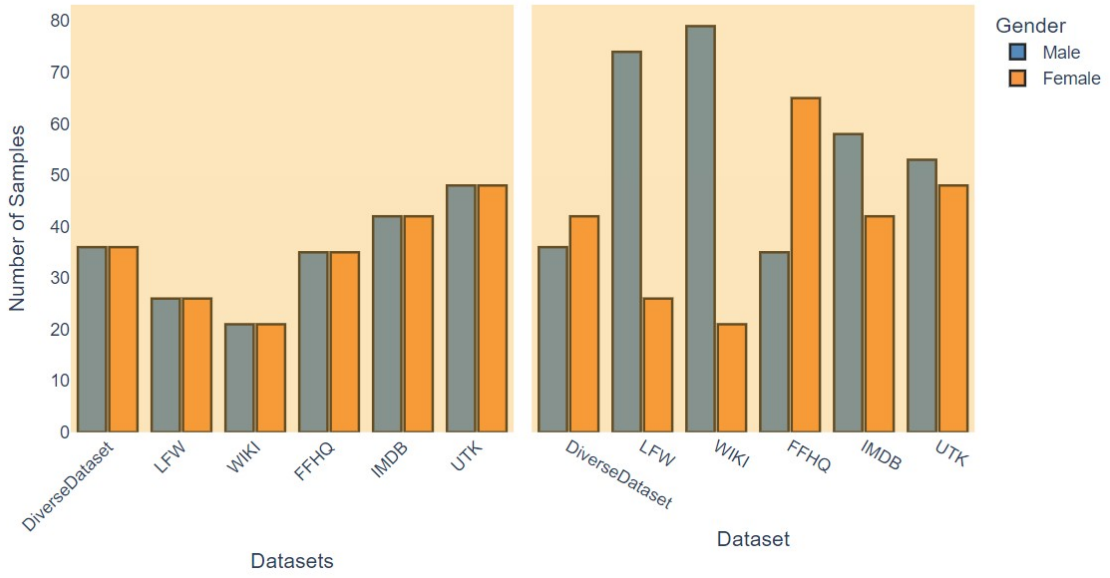


Figure 6.14: Number of the samples in Balance and Imbalance datasets

In addition to gender-based diversity, we extended our evaluation to measure diversity and fairness across nine geographically distinct language-location datasets [97,

32], which include: West Asia & North Africa, North America, Western Europe, South Asia, South East Asia, East Asia, Eastern Europe, Latin America, and Sub-Saharan Africa. For each of these regions, we manually organised the images into two gender groups: male and female.

To further generalise the applicability of the metric beyond gender, we assessed saliency-based diversity and fairness across five professions (CEO, Engineer, Nurse, Politician, and School Teacher) in the same nine language-location combinations as discussed in [97, 32]. Each profession was evaluated across these nine sub-categories to ensure a comprehensive understanding of diversity and fairness in various professional and cultural contexts.

Table 6.10: Diversity and Fairness Metrics: Comparison Between Balanced and Imbalanced Datasets.

Dataset	**Balanced**			**Imbalanced**		
	D_{within}	D_{inter}	$M_{\text{fairness-diversity}}$	D_{within}	D_{inter}	$M_{\text{fairness-diversity}}$
Baseline						
DiverseDataset	0.83±0.01	0.35±0.00	0.59±0.00	0.82±0.02	0.35±0.00	0.59±0.01
FFHQ	0.83±0.02	0.35±0.00	0.59±0.01	0.82±0.01	0.32±0.00	0.57±0.01
IMDB	0.80±0.01	0.35±0.00	0.57±0.00	0.80±0.01	0.34±0.00	0.57±0.00
LFW	0.84±0.02	0.36±0.00	0.60±0.01	0.81±0.01	0.27±0.00	0.54±0.01
UTK	0.80±0.01	0.35±0.00	0.57±0.00	0.80±0.01	0.35±0.00	0.57±0.00
WIKI	0.84±0.01	0.36±0.00	0.60±0.00	0.80±0.01	0.23±0.00	0.52±0.00
With FaceKeepOriginalAugment [102]						
DiverseDataset	0.83±0.00	0.36±0.00	0.59±0.00	0.82±0.01	0.35±0.00	0.59±0.00
FFHQ	0.83±0.02	0.36±0.00	0.59±0.01	0.83±0.01	0.32±0.00	0.57±0.00
IMDB	0.81±0.01	0.35±0.00	0.58±0.00	0.81±0.01	0.34±0.00	0.57±0.00
LFW	0.83±0.01	0.36±0.00	0.59±0.01	0.81±0.01	0.27±0.00	0.54±0.00
UTK	0.81±0.02	0.35±0.00	0.58±0.01	0.80±0.01	0.35±0.00	0.58±0.00
WIKI	0.84±0.01	0.36±0.00	0.60±0.00	0.82±0.00	0.24±0.00	0.53±0.00

In Table 6.10, the comparison of balanced and imbalanced datasets across the three key metrics within group diversity (D_{within}), between group diversity (D_{inter}), and the proposed fairness-diversity metric ($M_{\text{fairness-diversity}}$) highlights the effect of data augmentation strategies. Overall, FaceKeepOriginalAugment demonstrates a slight improvement in D_{within} and D_{inter} across most datasets when compared to the baseline. For instance, in the IMDB dataset, FaceKeepOriginalAugment increases D_{within} from 0.80 to 0.81 and maintains D_{inter} at 0.35 for balanced datasets. Similarly, in the WIKI dataset, FaceKeepOriginalAugment slightly improves $M_{\text{fairness-diversity}}$

from 0.52 to 0.53 in imbalanced conditions, suggesting improved fairness when dealing with underrepresented groups. The results for FFHQ and LFW datasets remain stable with minimal fluctuations, showcasing that the augmentation method preserves diversity and fairness without significant deviations from the baseline metrics. The notable consistency in D_{within} across diverse datasets implies that KeepOriginalAugment maintains intra-group diversity effectively, a strength of the approach. Meanwhile, $M_{\text{fairness-diversity}}$ is relatively stable across both balanced and imbalanced datasets, indicating robustness to bias. FaceKeepOriginalAugment marginally improves fairness while preserving diversity, offering a balanced solution that enhances representation, especially in datasets where class imbalance poses a challenge.

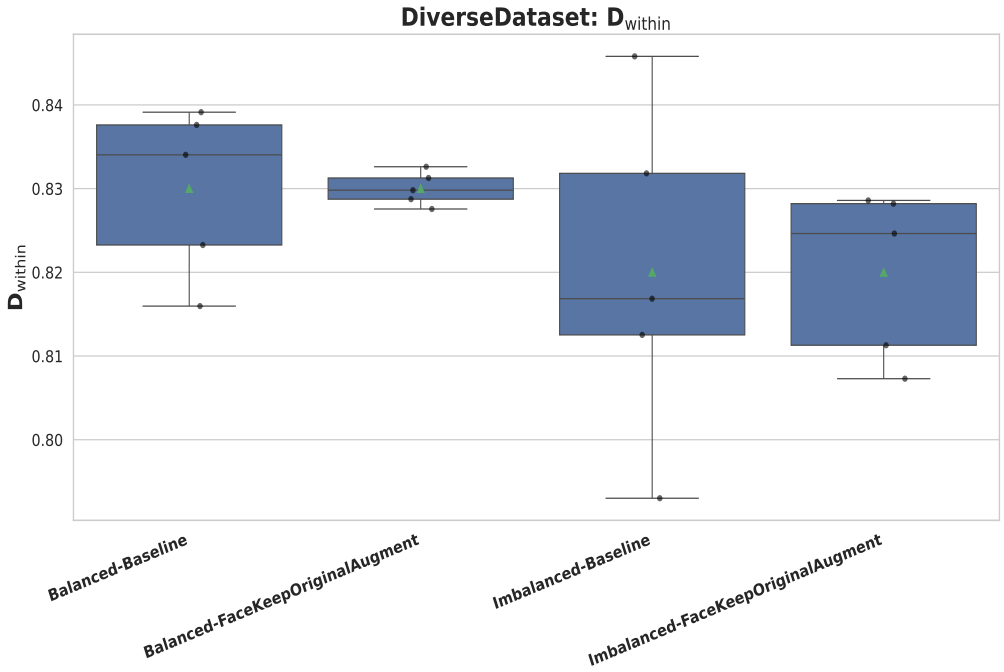
Distribution of $\mathbf{D}_{within}$ and $\mathbf{D}_{inter}$ for balance and imbalance datasets

Figure 6.15 shows the distributions of D_{within} and D_{inter} for DiverseDataset. All four conditions yield very similar D_{inter} values (around 0.35–0.36) with almost no variation, while D_{within} remains high (approximately 0.82–0.84) across runs. The balanced setting with FaceKeepOriginalAugment gives a slight increase in within-group diversity, indicating that the dataset is already diverse and the augmentation mainly helps to preserve this property.

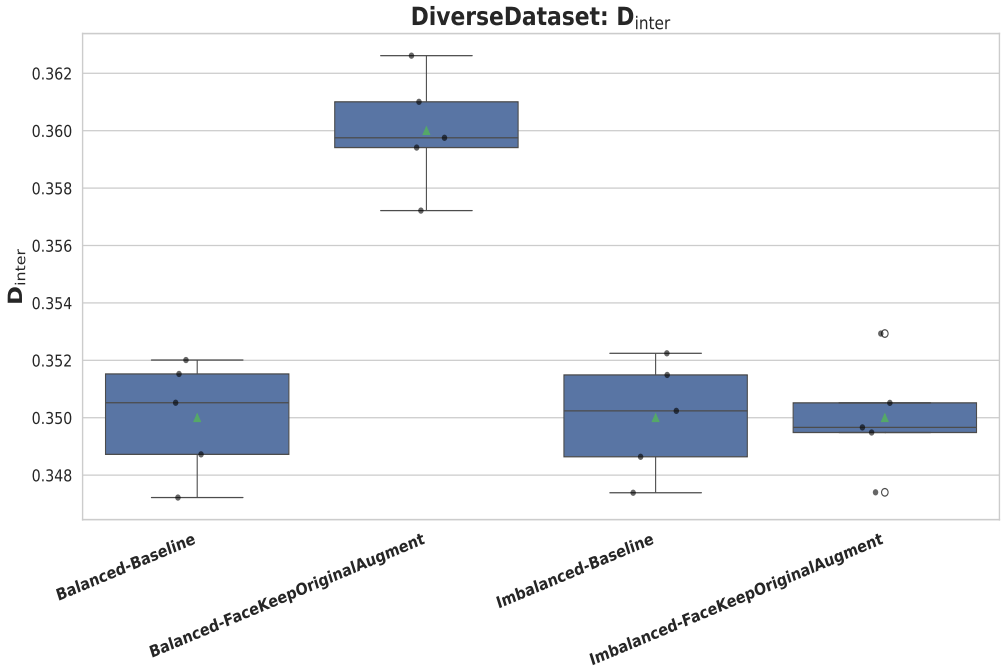
Figure 6.16 presents D_{within} and D_{inter} for FFHQ. In this case, the imbalance obviously lowers the value of D_{inter} (from 0.35 to 0.32) in the balanced case, but FaceKeepOriginalAugment in the balanced case moderately increases the value of Dinter and maintains the value of Dwithin high (approximately 0.82-0.84). It means that imbalance has a negative impact on between-group diversity, but augmentation has a positive effect on preserving a strong within-group diversity, and the effect on between-group diversity is negative and insignificant in the event of balanced data.

Figure 6.17 shows the behaviour of both metrics on IMDB. The difference in D_{inter} between balanced and imbalanced baselines is small (roughly 0.35 vs. 0.34), and both augmented configurations largely preserve these levels with very little run-to-run variability. For D_{within}, the balanced baseline is around 0.80, and Face-

KeepOriginalAugment increases it slightly (to about 0.81) in both balanced and imbalanced settings. The plots in general imply that IMDB has high within group diversity and a relatively small decrease in between group diversity in the face of imbalance.

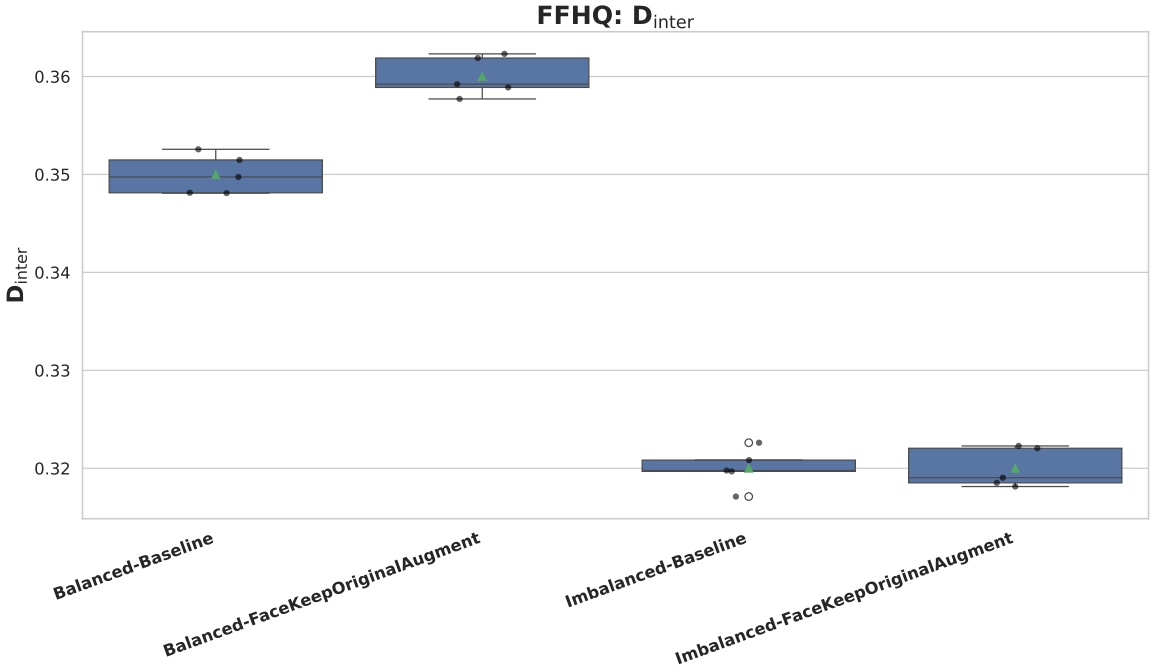


(a) D_{within}

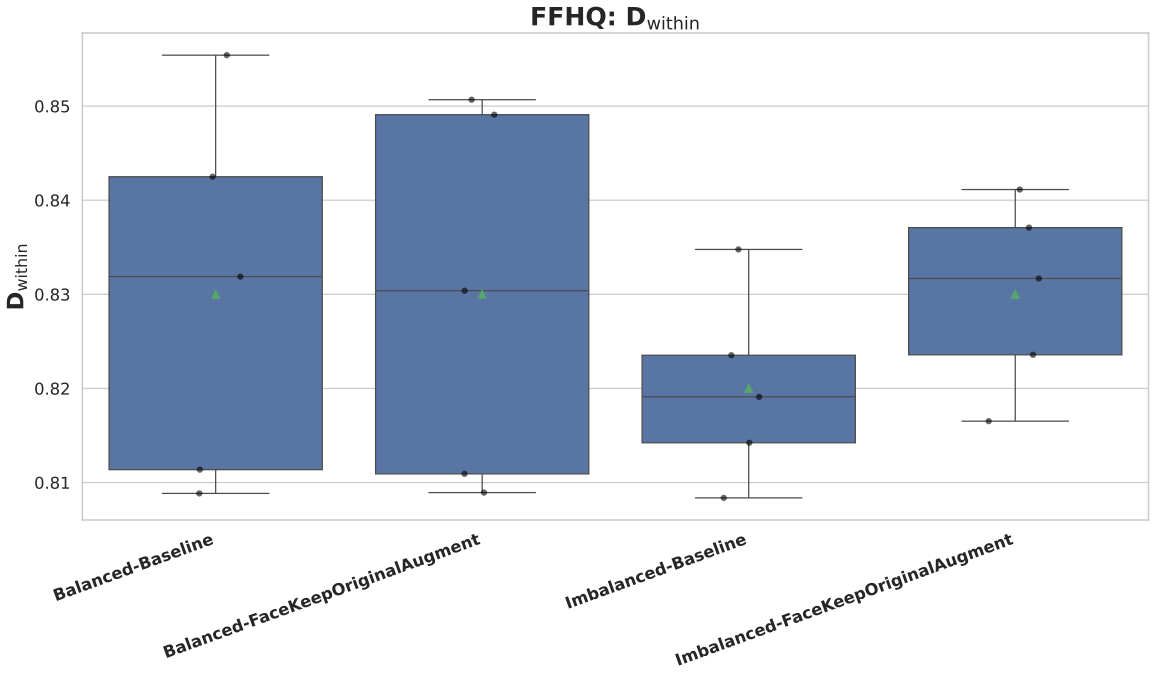


(b) D_{inter}

Figure 6.15: Distribution of D_{within} and D_{inter} metrics on **DiverseDataset** across the four conditions (Balanced/Imbalanced × Baseline/FaceKeepOriginalAugment).

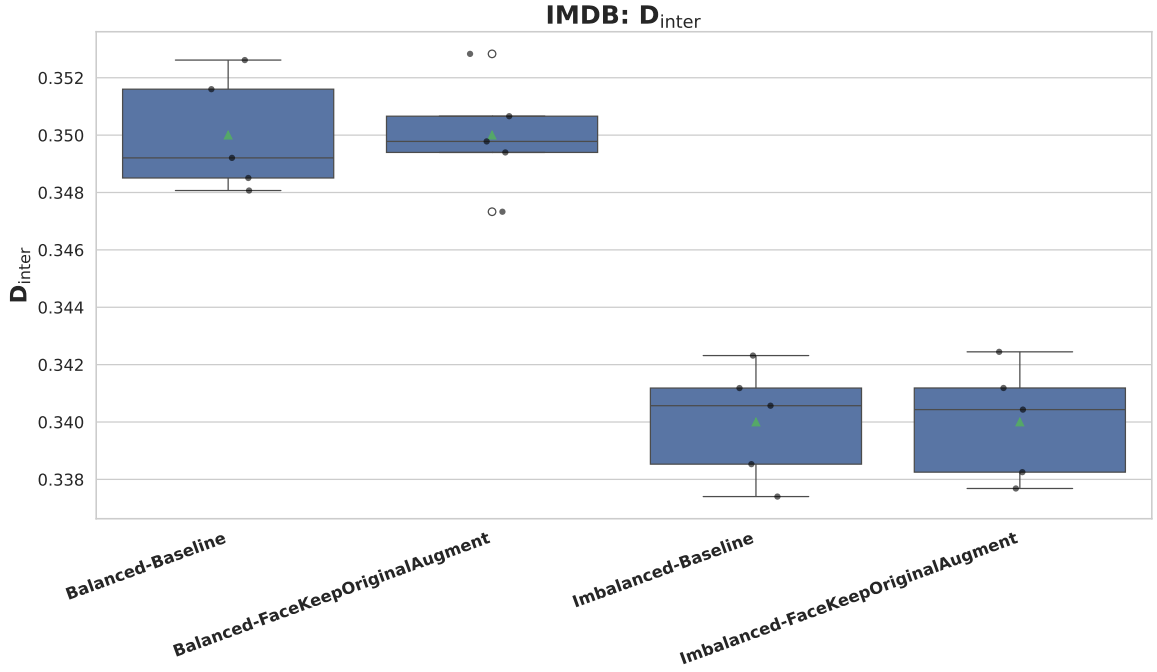


(a) D_{inter}

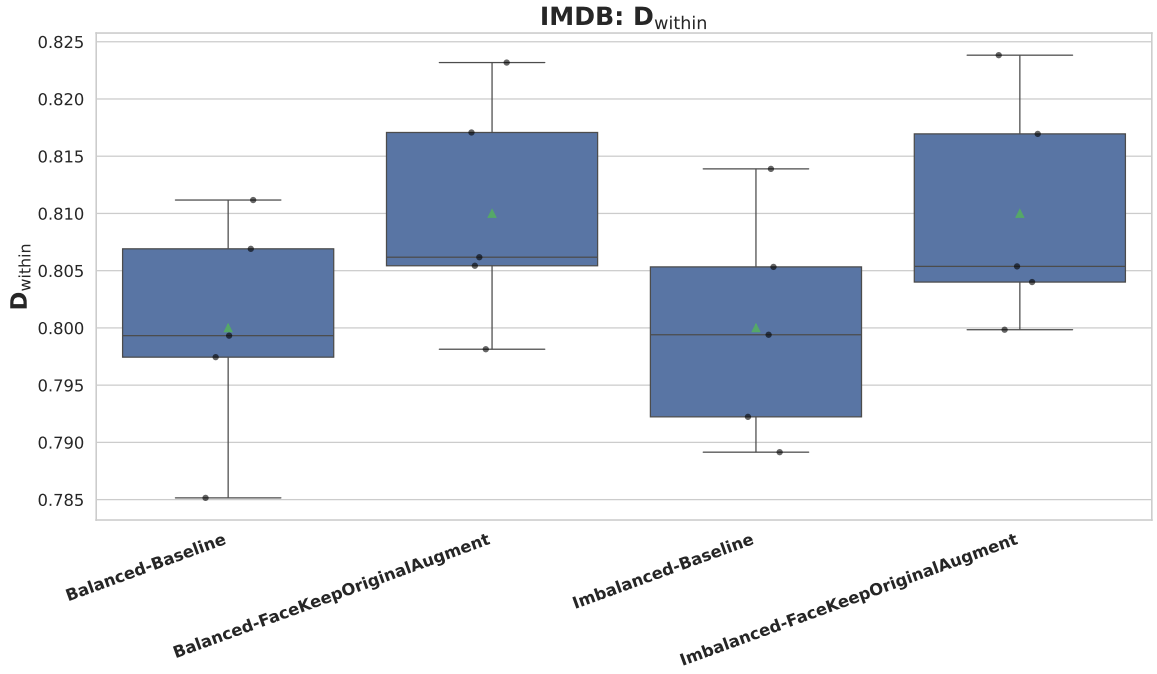


(b) D_{within}

Figure 6.16: Distribution of D_{within} and D_{inter} metrics on **FFHQ** across the four conditions (Balanced/Imbalanced × Baseline/FaceKeepOriginalAugment).

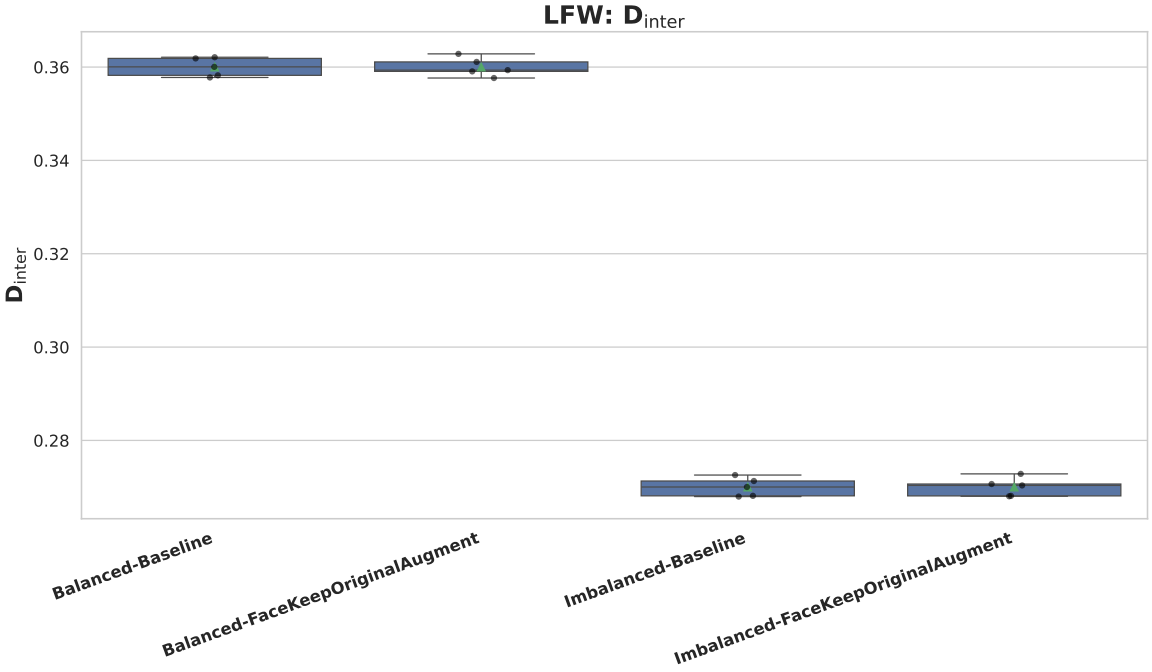


(a) D_{inter}

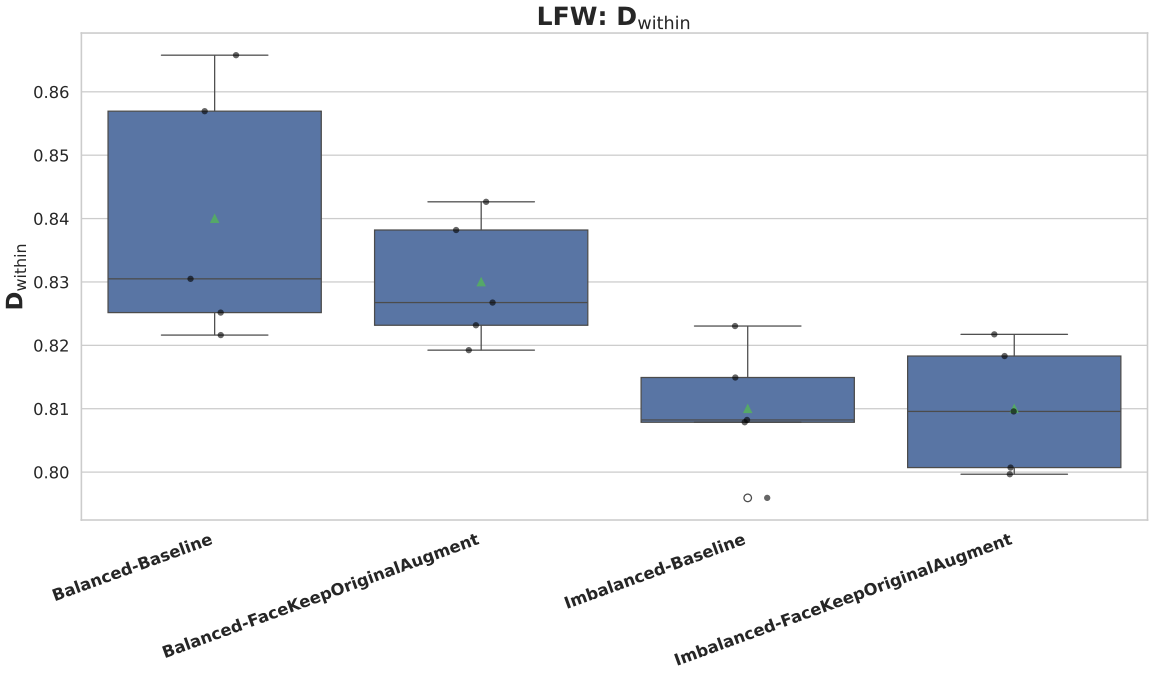


(b) D_{within}

Figure 6.17: Distribution of D_{within} and D_{inter} metrics on **IMDB** across the four conditions (Balanced/Imbalanced × Baseline/FaceKeepOriginalAugment).

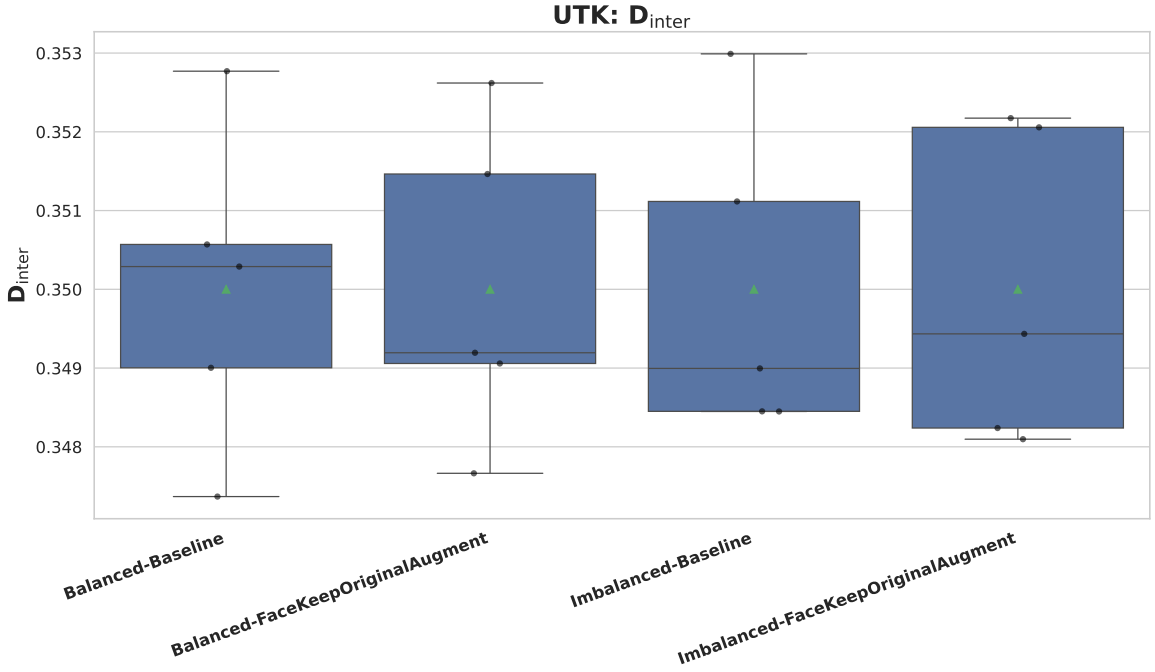


(a) D_{inter}

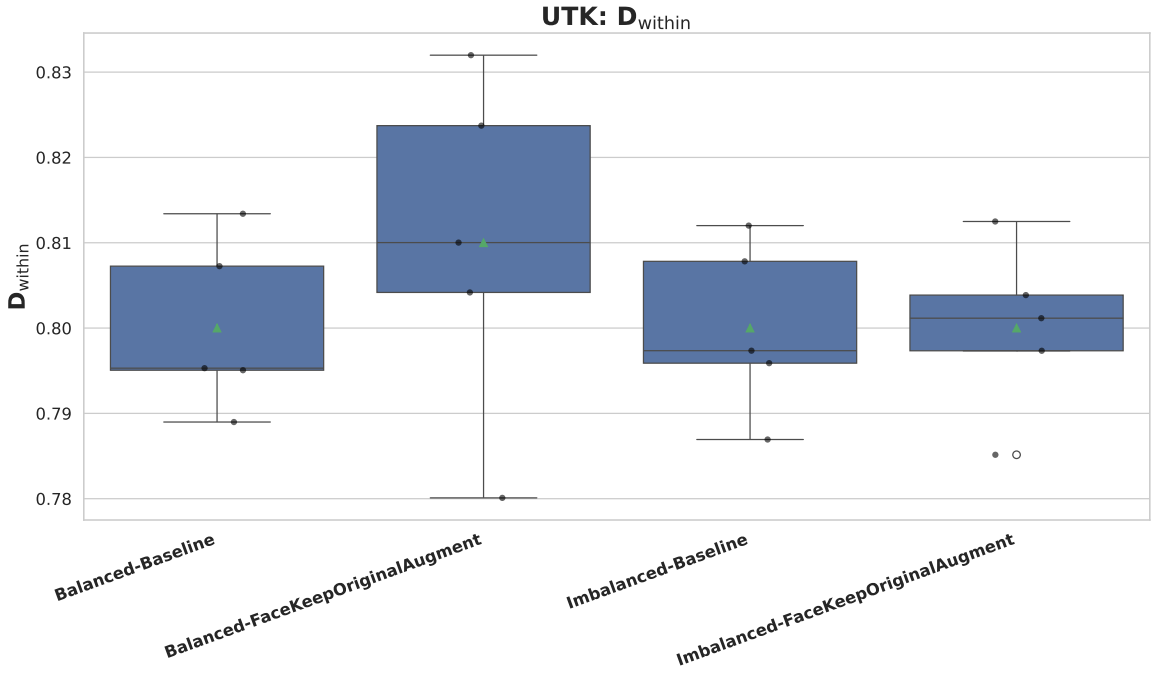


(b) D_{within}

Figure 6.18: Distribution of D_{within} and D_{inter} metrics on **LFW** across the four conditions (Balanced/Imbalanced × Baseline/FaceKeepOriginalAugment).

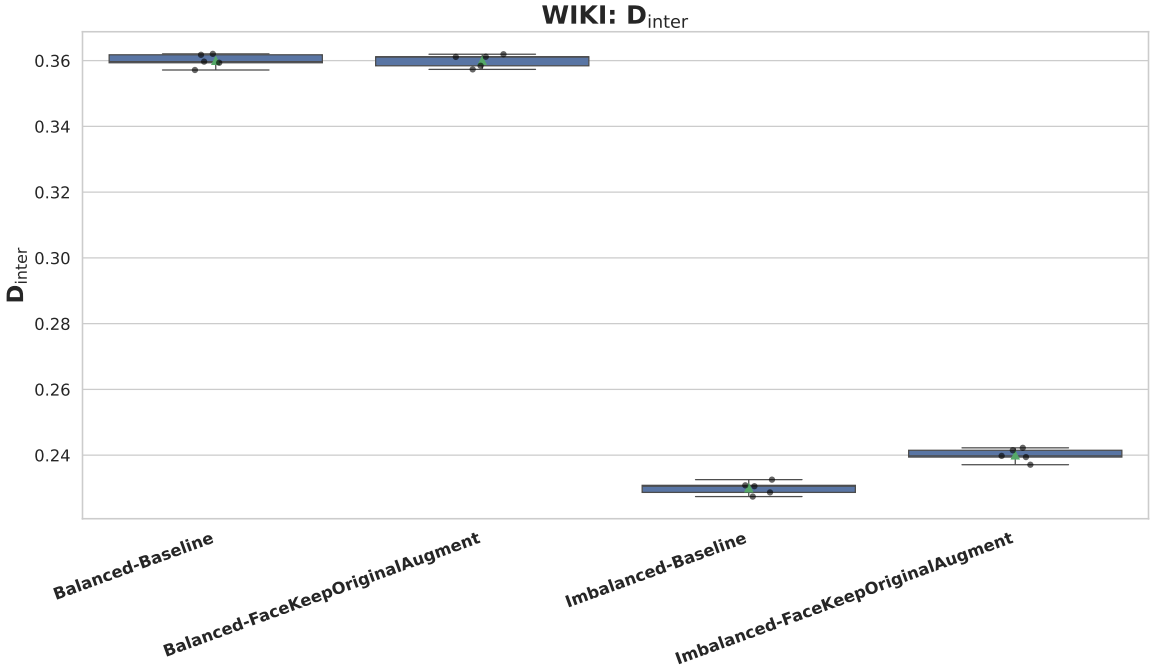


(a) D_{inter}

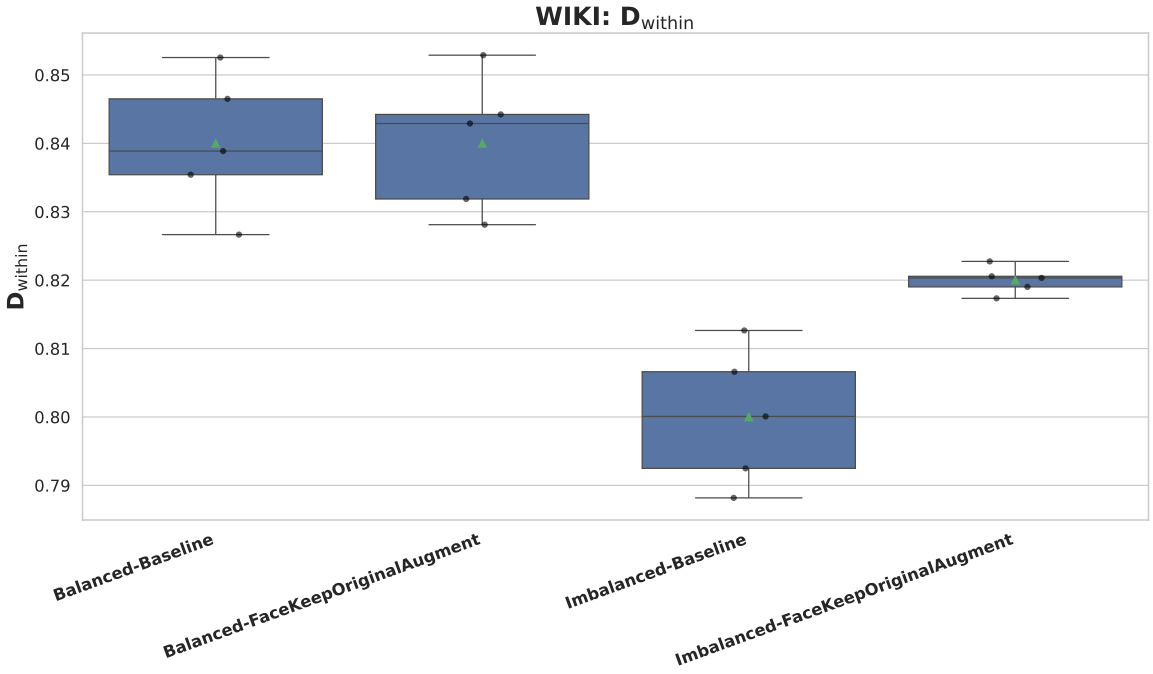


(b) D_{within}

Figure 6.19: Distribution of D_{within} and D_{inter} metrics on **UTK** across the four conditions (Balanced/Imbalanced × Baseline/FaceKeepOriginalAugment).



(a) D_{inter}



(b) D_{within}

Figure 6.20: Distribution of D_{within} and D_{inter} metrics on **WIKI** across the four conditions (Balanced/Imbalanced × Baseline/FaceKeepOriginalAugment).

Figure 6.18 shows a stronger effect of imbalance on **LFW**. As the balanced baseline, the imbalanced one falls to approximately 0.27, which means that there is definitely a loss of between-group diversity. Using FaceKeepOriginalAugment in the balanced case maintains the larger D_{inter}, whereas in the imbalanced case maintains the smaller value of D_{inter} near the smaller value. In all scenarios, within group diversity is strong (approximately 0.81- 0.84), indicating that in-group diversity is strong even when inter-group diversity becomes bad in the case of imbalance.

Figure 6.19 reports the distributions for UTK. The values of the D_{inter} diversity are very similar (around 0.35) with very few variations across all four conditions, which suggests the consistency of between-group diversity and minimal influence of balance and augmentation. The values of the D_{within} are in the range of about 0.80 to 0.82 and the balanced setting is FaceKeepOriginalAugment where a minor increase is being encountered against the balanced baseline. In general, the UTK findings indicate high diversity properties, and FaceKeepOriginalAugment has a slight reinforcing effect on within-group diversity but does not change between-group diversity.

Figure 6.20 shows the distributions of D_{within} and D_{inter} for WIKI across the four conditions. In the balanced setting, both the baseline and FaceKeepOriginalAugment configurations yield relatively high D_{within} (around 0.84) and higher D_{inter} (around 0.36), whereas the imbalanced baseline clearly reduces D_{inter} (down to roughly 0.23–0.24) and slightly lowers D_{within} (to about 0.80). With FaceKeepOriginalAugment applied to the imbalanced data, both metrics are partially recovered (with a small increase in D_{within} and a slight improvement in D_{inter}), indicating that augmentation can mitigate, but not completely eliminate, the loss of diversity induced by imbalance on this dataset. In Table 6.11, we present the diversity and fairness metrics, within group diversity (D_{within}), between-group diversity (D_{inter}), and the fairness-diversity metric ($M_{\text{fairness-diversity}}$) for various language-location pair queries both in their baseline state and with the FaceKeepOriginalAugment. The baseline metrics exhibit consistent values across different pairs, with D_{within} ranging

Table 6.11: Diversity and Fairness Metrics for Language-Location Pairs Across Gender

Language-Location Pairs	**Baseline**			**With FaceKeepOriginalAugment**		
	D_{within}	D_{inter}	$M_{\text{fairness-diversity}}$	D_{within}	D_{inter}	$M_{\text{fairness-diversity}}$
Arabic-West Asia & North Africa	0.59±0.27	0.49±0.16	0.54±0.06	0.79±0.01	0.35±0.00	0.57±0.01
English-North America	0.59±0.27	0.49±0.16	0.54±0.06	0.81±0.01	0.35±0.00	0.58±0.01
English-West Europe	0.59±0.28	0.50±0.17	0.55±0.06	0.80±0.02	0.35±0.00	0.57±0.01
Hindi-South Asia	0.59±0.27	0.48±0.15	0.53±0.06	0.80±0.01	0.35±0.00	0.57±0.00
Indonesian-Southeast Asia	0.59±0.27	0.49±0.16	0.54±0.06	0.79±0.01	0.35±0.00	0.57±0.01
Mandarin-East Asia	0.60±0.27	0.51±0.18	0.55±0.04	0.79±0.01	0.35±0.00	0.57±0.00
Russian-East Europe	0.59±0.27	0.50±0.17	0.54±0.05	0.80±0.01	0.35±0.00	0.57±0.01
Spanish-Latin America	0.60±0.27	0.49±0.16	0.54±0.06	0.79±0.00	0.35±0.00	0.57±0.00
Swahili-Sub-Saharan Africa	0.59±0.27	0.50±0.17	0.55±0.05	0.80±0.01	0.35±0.00	0.57±0.00

from 0.59 to 0.60 and $M_{\text{fairness-diversity}}$ values hovering around 0.53 to 0.55, raising fairness concerns in these datasets. With FaceKeepOriginalAugment, there is a noticeable enhancement in the metrics. For instance, the D_{within} values improve to 0.79 for most pairs, showcasing a substantial increase in intra-group diversity, while $M_{\text{fairness-diversity}}$ also sees an uplift to around 0.57 across the board. Notably, pairs such as Arabic-West Asia & North Africa and English - North America show an increase in D_{within} from 0.59 to 0.79 and $M_{\text{fairness-diversity}}$ from 0.54 to 0.57, respectively. This improvement signifies that the augmentation strategy enhances both diversity and fairness in these pairs, addressing potential biases that might be present in the baseline datasets. The results illustrate that FaceKeepOriginalAugment effectively enriches the diversity while maintaining fairness across gender in various language-location pairs. This highlights the method's strength in fostering a more equitable representation in datasets that may initially show limited diversity. The insights drawn from these metrics suggest that employing such augmentation strategies is crucial in developing models that prioritise fairness and diversity, particularly in linguistically diverse contexts.

In Table 6.12, we present the diversity and fairness metrics within group diversity for various professions both in their baseline state and with FaceKeepOriginalAugment. The baseline results indicate a high level of within-group diversity for each profession, with D_{within} values consistently around 0.82 to 0.83. This suggests that the profession datasets are already fairly diverse in terms of representation. The $M_{\text{fairness-diversity}}$ values for the baseline remain stable at approximately 0.72, indi-

Table 6.12: Diversity and Fairness Metrics measurement of different Profession datasets across Language Location pairs

Profession	**Baseline**			**With FaceKeepOriginalAugmentaiton**		
	D_{within}	D_{inter}	$M_{\text{fairness-diversity}}$	D_{within}	D_{inter}	$M_{\text{fairness-diversity}}$
CEO	0.83± 0.01	0.61± 0.00	0.72± 0.00	0.86± 0.00	0.63± 0.00	0.74± 0.00
Engineer	0.83± 0.01	0.62± 0.00	0.73± 0.00	0.85± 0.00	0.63± 0.00	0.74± 0.00
Nurse	0.82± 0.01	0.61± 0.00	0.72± 0.01	0.86± 0.00	0.63± 0.00	0.74± 0.00
Politician	0.82± 0.00	0.61± 0.00	0.72± 0.00	0.86± 0.00	0.62± 0.00	0.74± 0.00
School Teacher	0.83± 0.00	0.62± 0.00	0.72± 0.00	0.86± 0.01	0.62± 0.00	0.74± 0.00

cating a reasonable balance between fairness and diversity across these professions. with FaceKeepOriginalAugment, there is an observable enhancement in the diversity metrics. The D_{within} values increase to between 0.85 and 0.86, reflecting an improvement in intra-group diversity. Similarly, the $M_{\text{fairness-diversity}}$ remains consistent at 0.74 for all professions, demonstrating that the augmentation method maintains fairness while enhancing diversity. Notably, the D_{inter} values show a slight increase in some cases, such as the CEO and Nurse professions, where the values are maintained around 0.61 to 0.63 after augmentation. This suggests that the method does not negatively impact the between-group diversity while improving within-group diversity. Overall, the results indicate that FaceKeepOriginalAugment effectively increases the diversity of profession datasets while preserving fairness across language location pairs. The enhancements in the metrics highlight the importance of using augmentation strategies to ensure a more equitable representation in datasets used for model training, particularly in contexts where professions may be underrepresented or biased.

Distribution of $\mathbf{D}_{within}$ and $\mathbf{D}_{inter}$ for profession datasets Figure 6.21 compares the CEO profession in the Baseline and FaceKeepOriginalAugment conditions, and it can be seen that FaceKeepOriginalAugment modestly enhances both between-group diversity, denoted by D_{inter}, and within-group diversity, denoted by D_{within}. Numerically, D_{within} increases from about 0.83 to 0.86 and D_{inter} from about 0.61 to 0.63, and the variability across repetitions is minimal, indicating that with FaceKeepOriginalAugment the CEO images are more diverse and balanced without introducing instability.

This same behaviour can be observed in Figure 6.22 for Engineers: FaceKeepOriginalAugment has higher D_{inter} and D_{within} than the Baseline (increasing D_{within} from roughly 0.83 to 0.85 and D_{inter} from about 0.62 to 0.63), and the slim boxes and strips indicate that results do not vary across runs, showing that the augmentation enhances both cross-group and in-group diversity in this occupation.

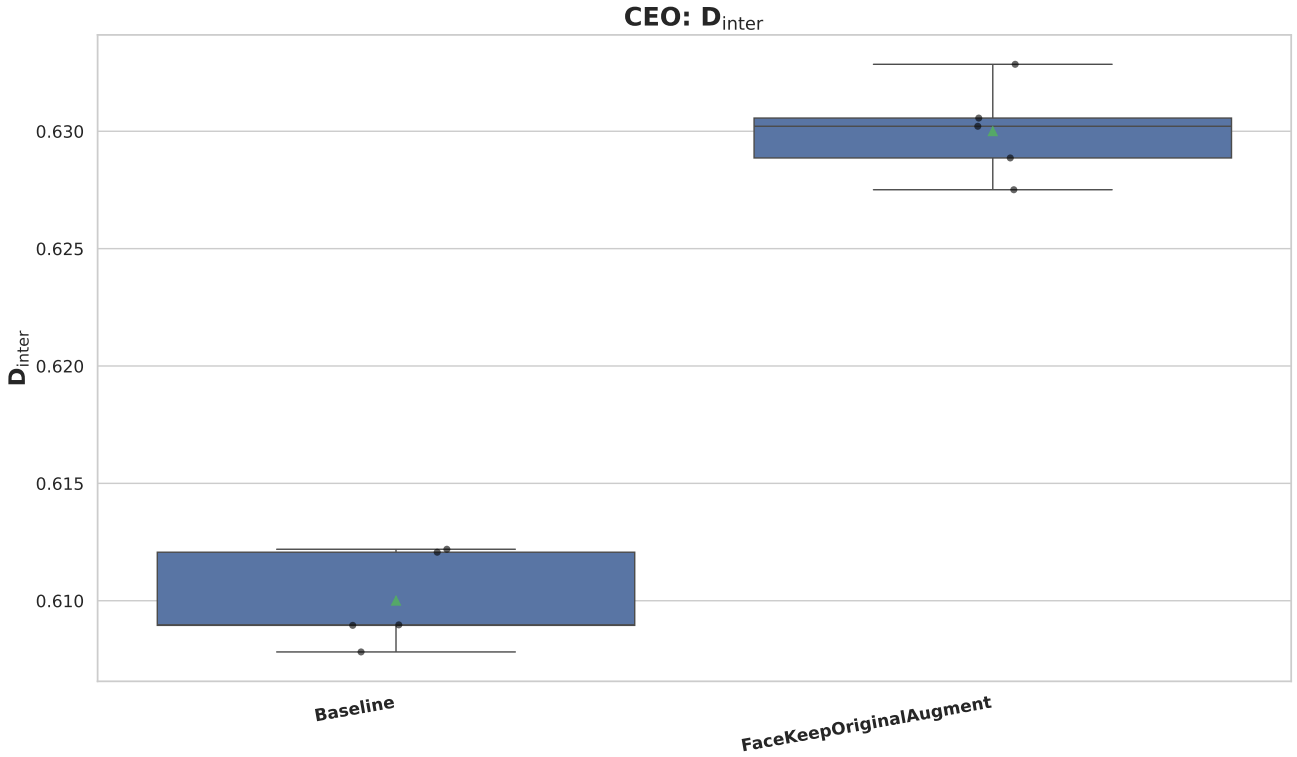
As displayed in Figure 6.23, FaceKeepOriginalAugment enhances both D_{inter} and D_{within} in comparison to the Baseline when Nurses are considered, with D_{within} rising from around 0.82 to 0.86 and D_{inter} from about 0.61 to 0.63. The distributions are shifted upwards and remain compact rather than spread in a disorderly manner, which means that the augmented environment gives nurses better representation without instability.

In Figure 6.24, the Politician boxplots reveal that, when compared to the Baseline, the values of D_{inter} and D_{within} are slightly larger with FaceKeepOriginalAugment (from roughly 0.82 to 0.86 for D_{within} and from about 0.61 to 0.62 for D_{inter}), and the fact that the points are not far apart again supports the idea that the augmentation assists in covering more subgroups related to this occupation while remaining stable across runs.

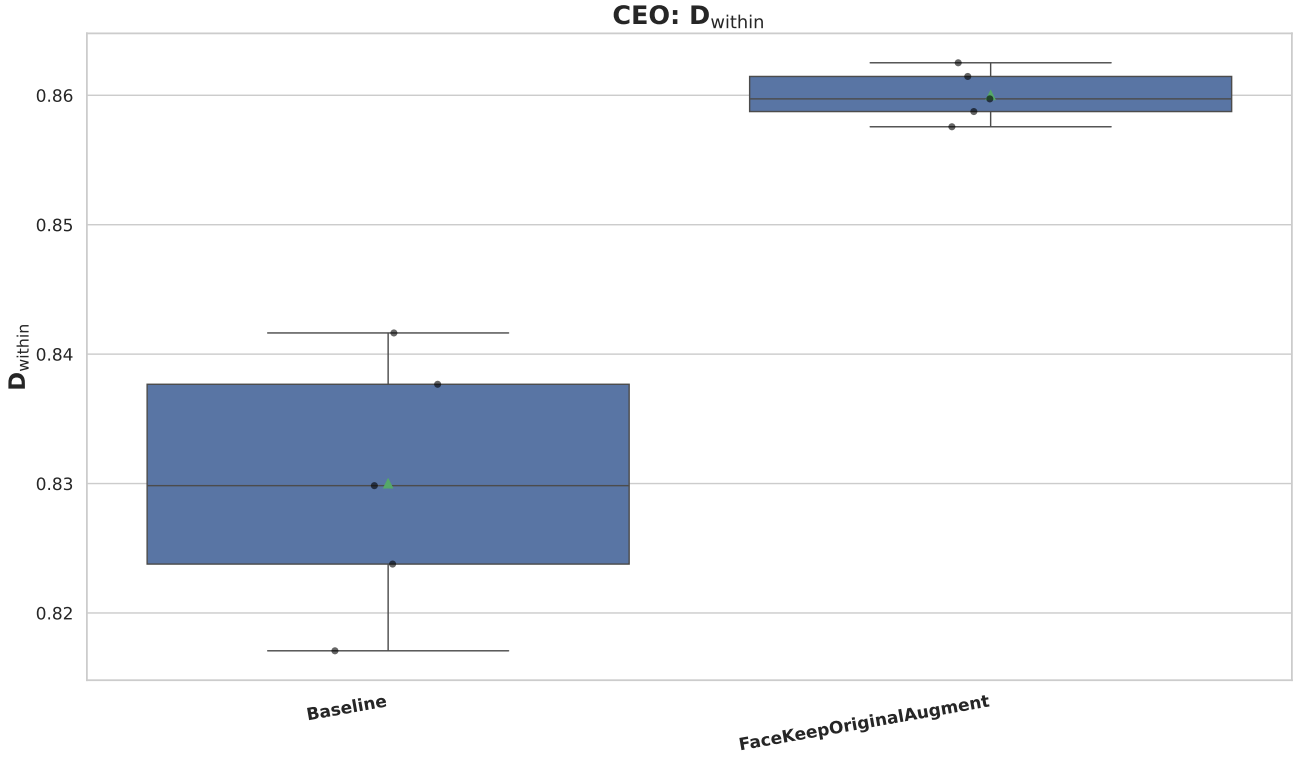
Figure 6.25 provides the measures for School Teachers, where FaceKeepOriginalAugment demonstrates better values of D_{within} (increasing from about 0.83 to 0.86) and slightly improved values of D_{inter} (both around 0.62) compared to the Baseline, with very compact boxplots and point clouds. This shows that the augmentation systematically increases the diversity of this profession without altering the behaviour across repetitions.

6.3.6 Additional work evaluation -Partial Mix (PM) and Noise Addition(NA)

We have explored two data augmentation techniques evaluations.

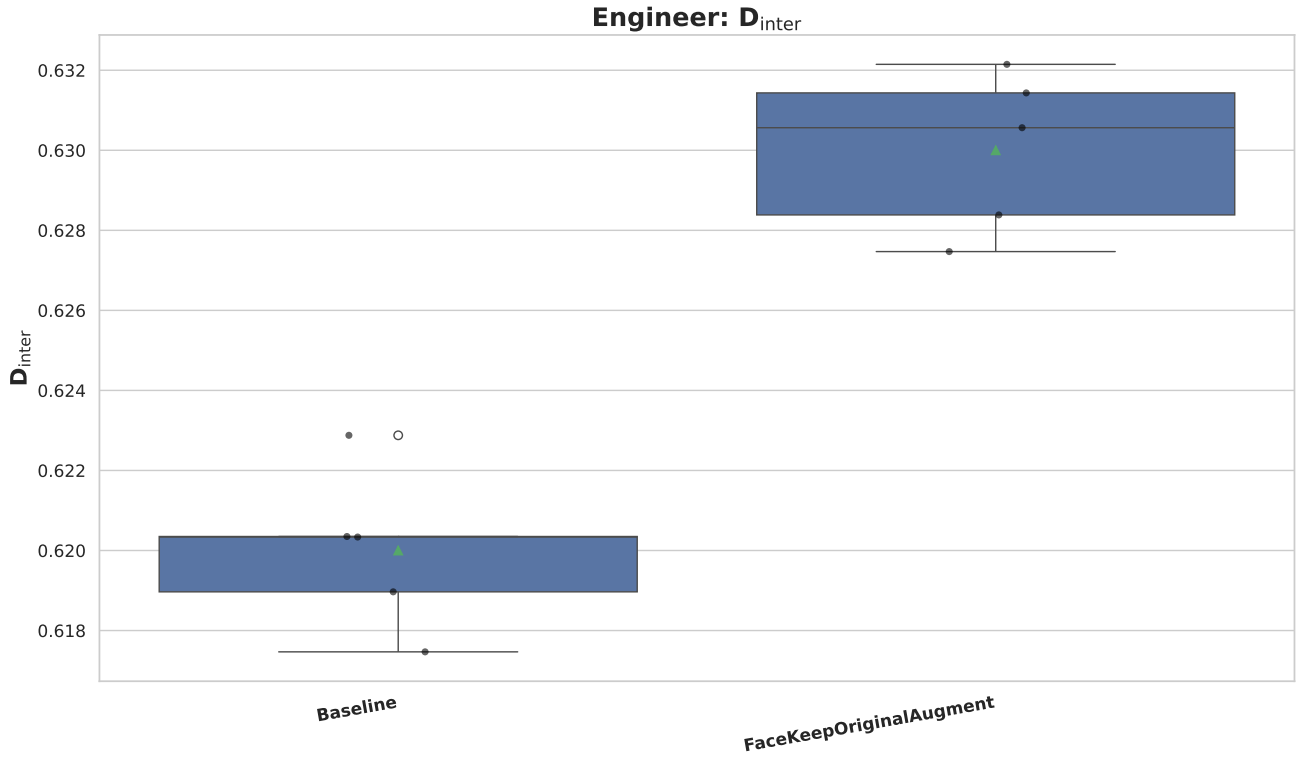


(a) D_{inter}

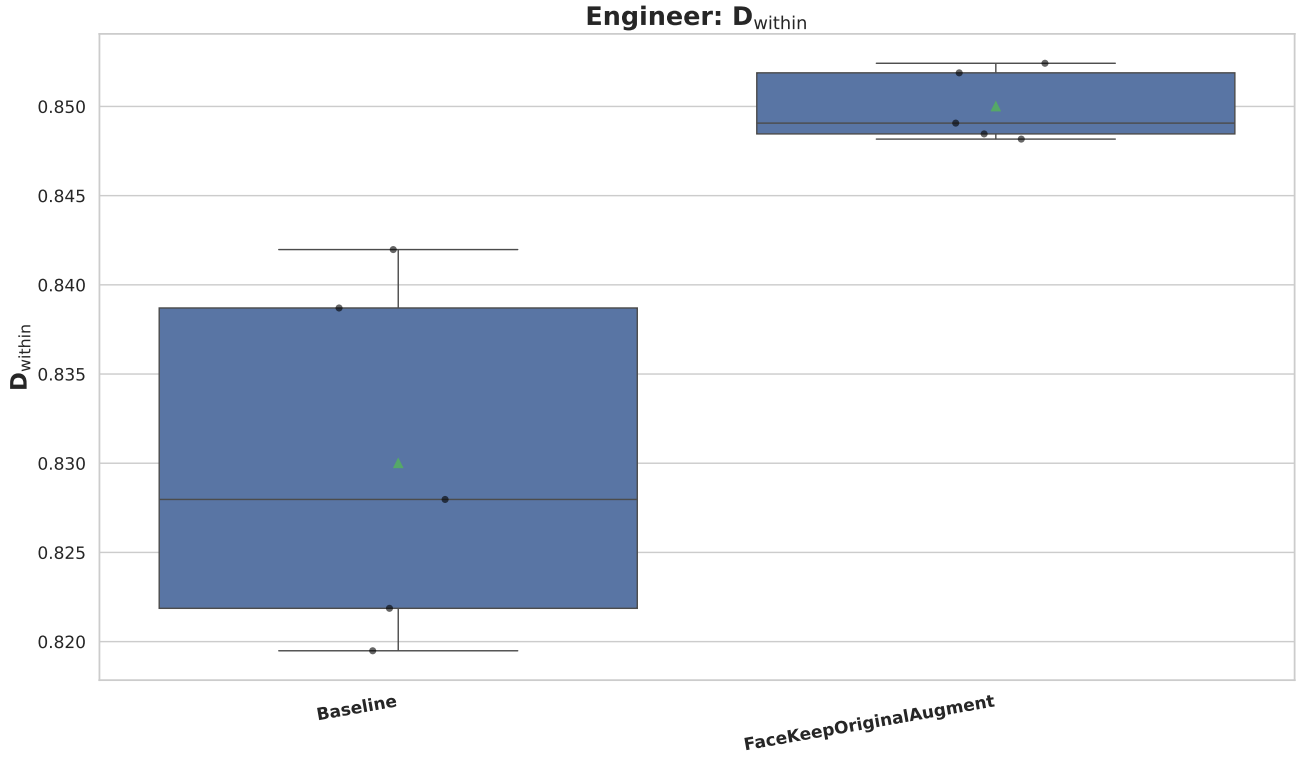


(b) D_{within}

Figure 6.21: Distribution of D_{within} and D_{inter} metrics for the **CEO** profession Baseline and FaceKeepOriginalAugment comparison.

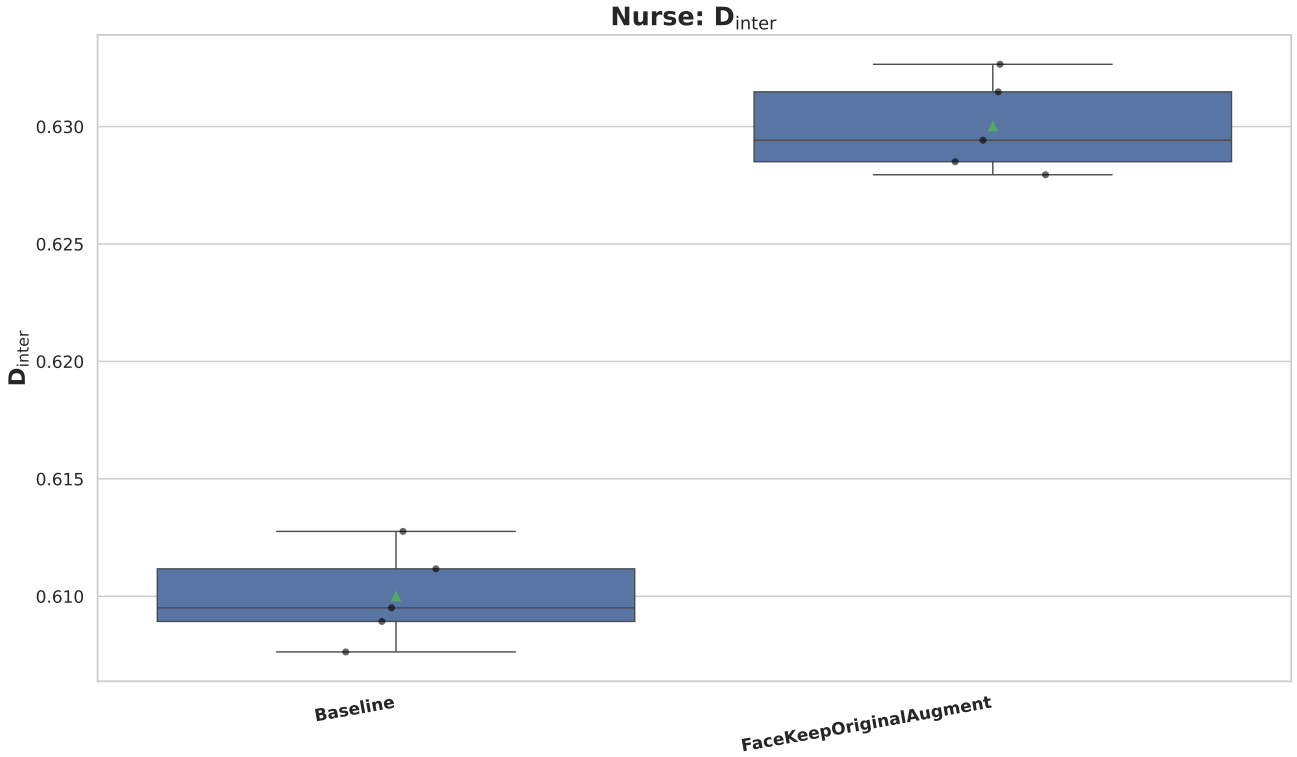


(a) D_{inter}

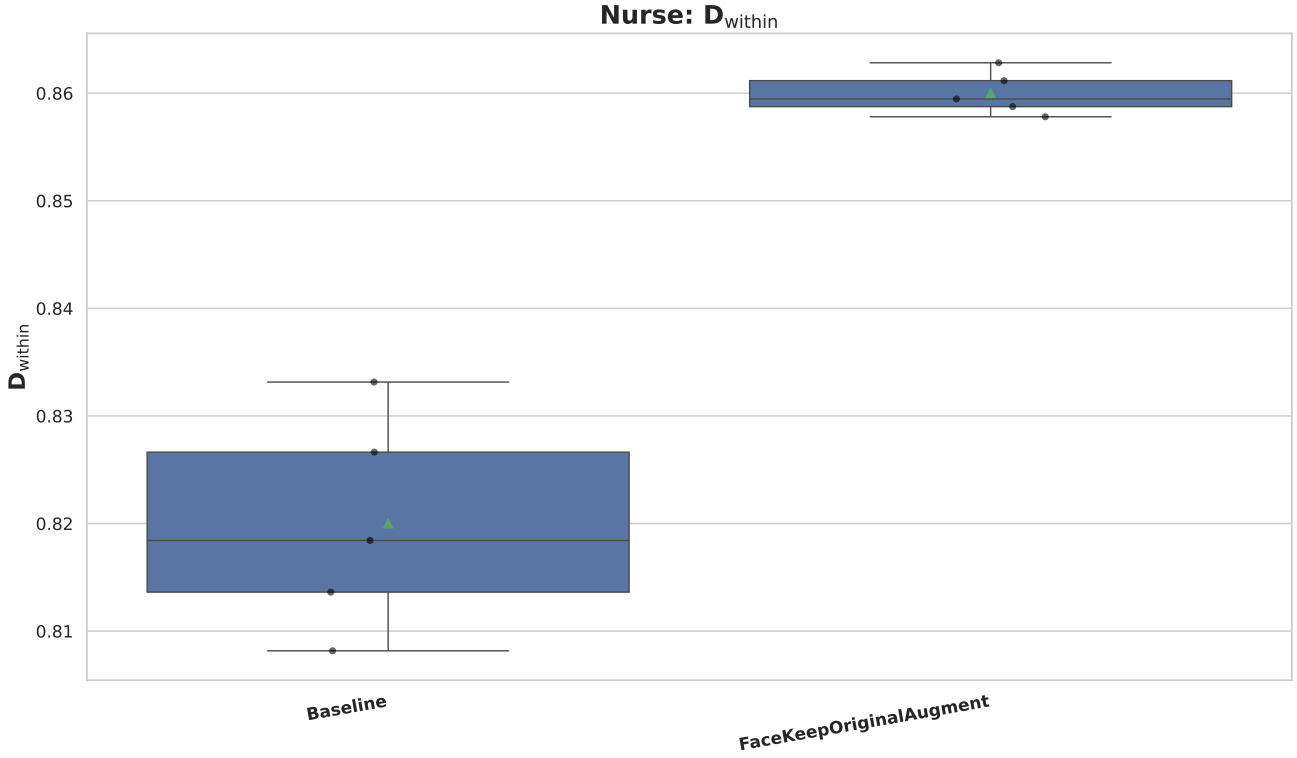


(b) D_{within}

Figure 6.22: Distribution of D_{within} and D_{inter} metrics for the **Engineer** profession Baseline and FaceKeepOriginalAugment comparison.

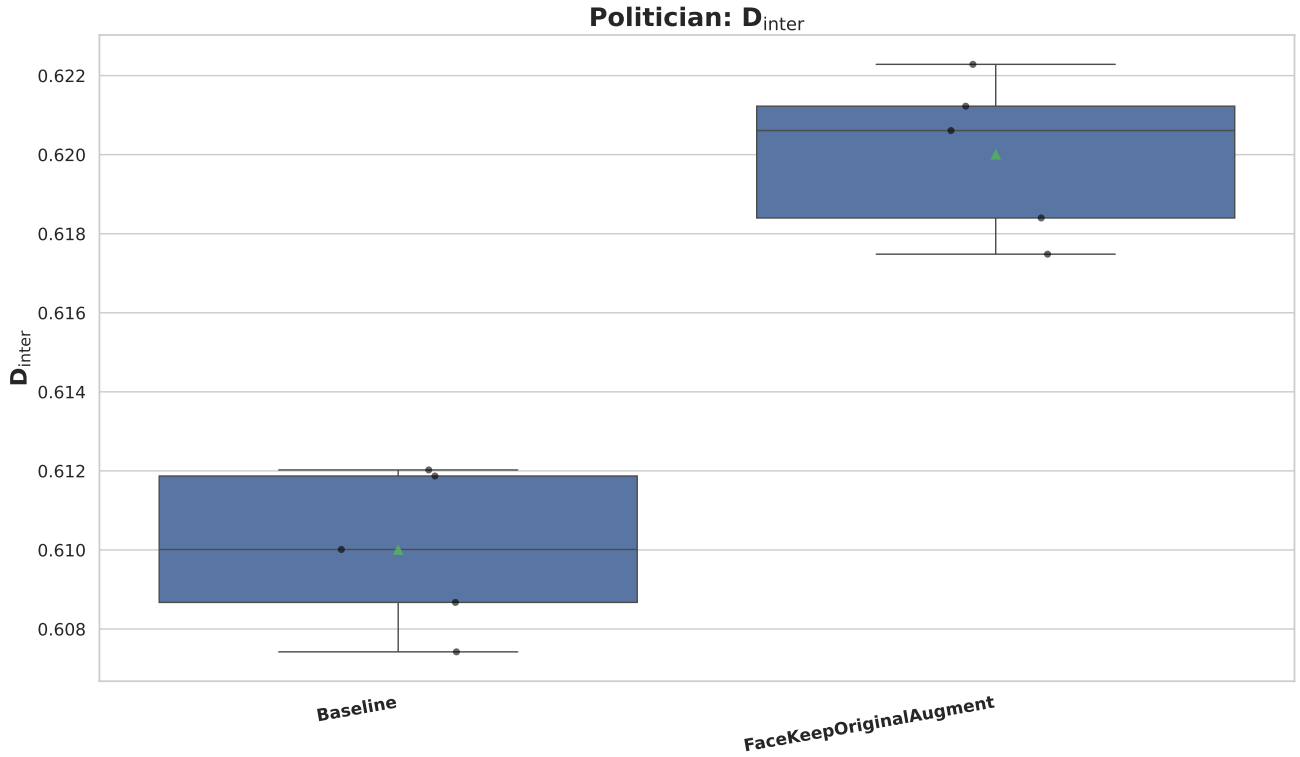


(a) D_{inter}

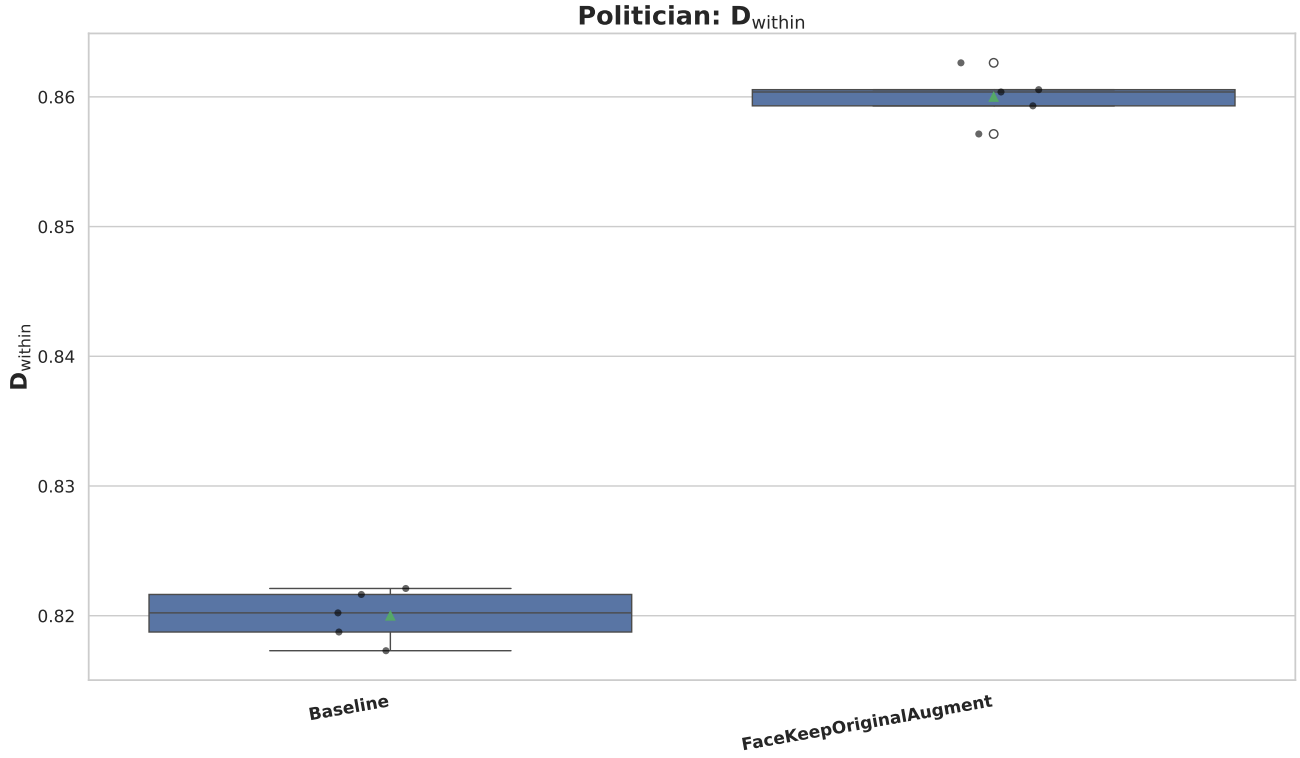


(b) D_{within}

Figure 6.23: Distribution of D_{within} and D_{inter} metrics for the **Nurse** profession Baseline and FaceKeepOriginalAugment comparison.

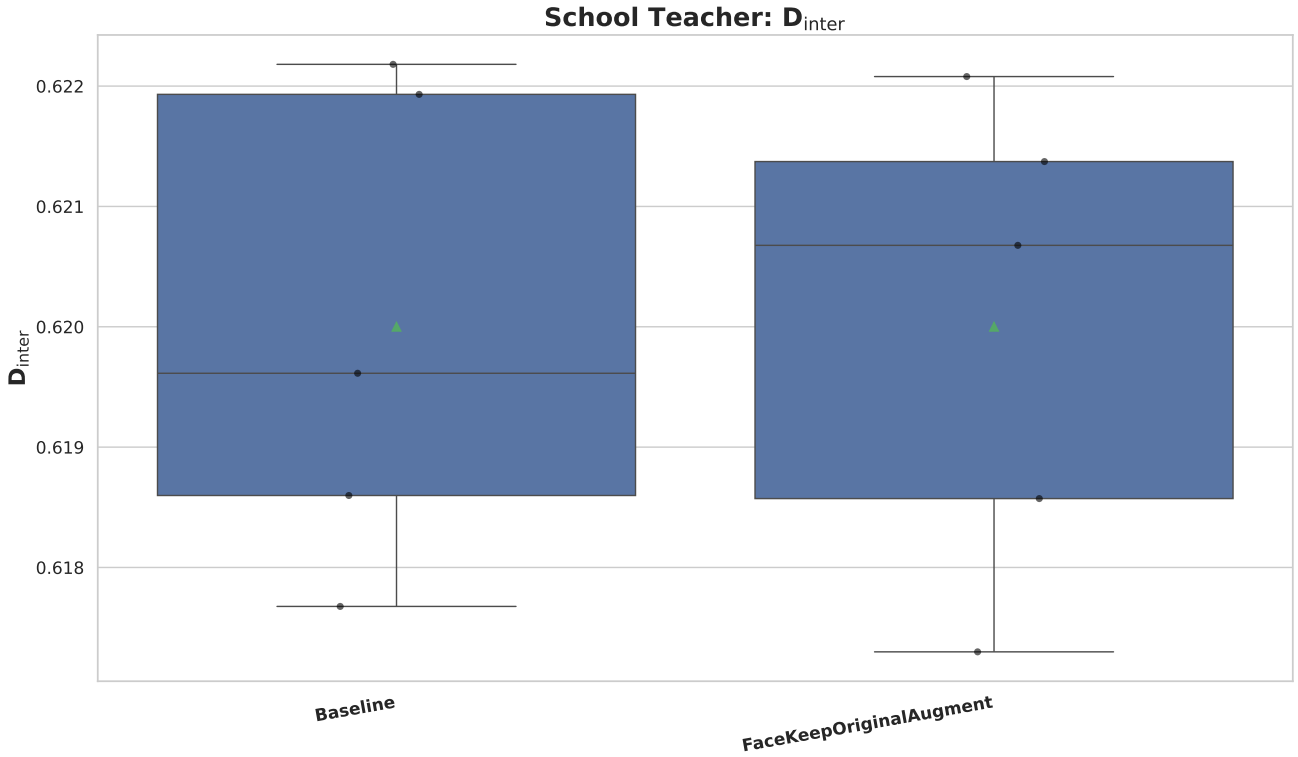


(a) D_{inter}

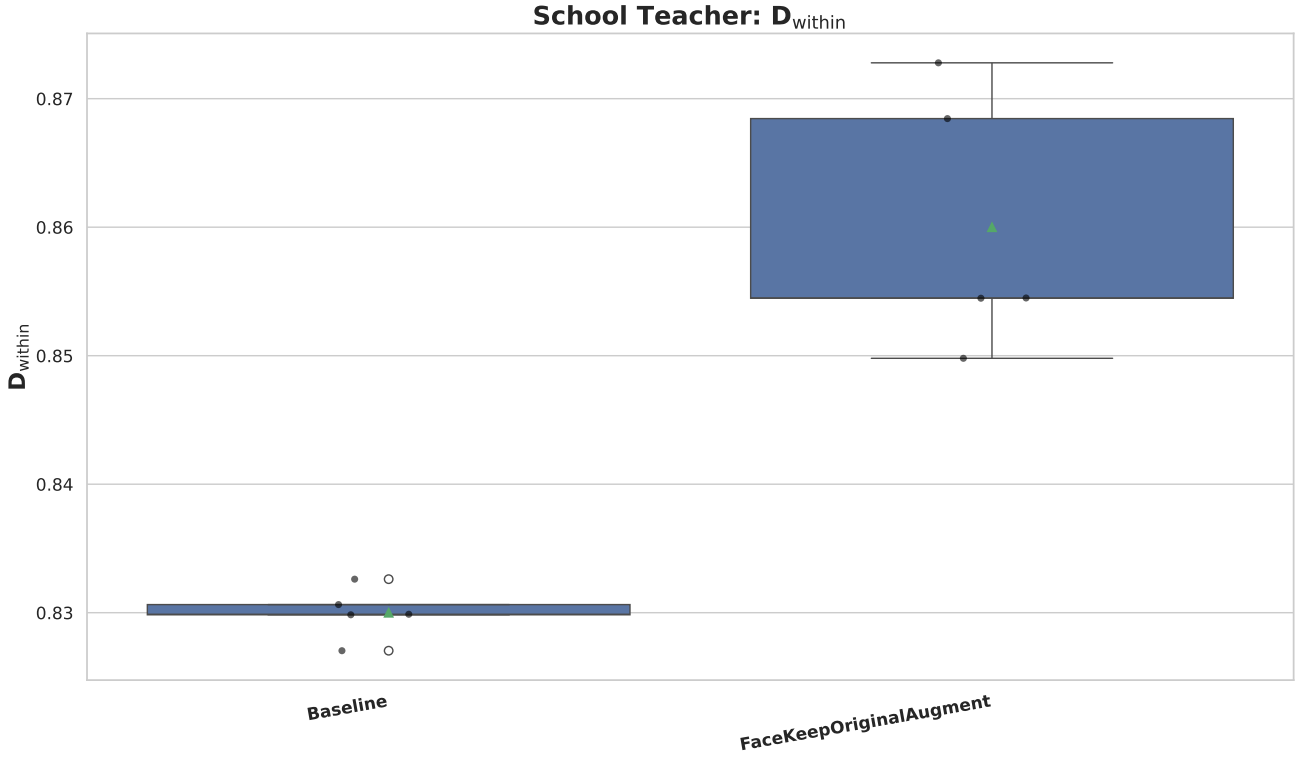


(b) D_{within}

Figure 6.24: Distribution of D_{within} and D_{inter} metrics for the **Politician** profession Baseline and FaceKeepOriginalAugment comparison.



(a) D_{inter}



(b) D_{within}

Figure 6.25: Distribution of D_{within} and D_{inter} metrics for the **School Teacher** profession Baseline and FaceKeepOriginalAugment comparison.

Intra-dataset Image Similarity (ISS_{intra}) Evaluation:

The results in Table 6.13 demonstrate the effectiveness of both of our methods, "with PM" and "with NA", in improving the Intra-dataset Image Similarity Score (ISS$intra$) across multiple datasets. Our PM approach consistently outperforms the baseline for all datasets, achieving the highest ISS$intra$ values representing greater diversity in the results. Specifically, the PM method achieves notable improvements on the FFHQ, Diverse Dataset, WIKI, IMDB, LFW, and UTK datasets, with the highest ISS_{intra} observed for the IMDB dataset at 1.21. In contrast, the NA approach, while still improving upon the baseline, yields slightly lower scores compared to the PM method but consistently surpasses the baseline. This demonstrates that both methods contribute to improved dataset diversity, with the PM approach being more effective overall.

Table 6.13: ISS_{intra} of datasets for baseline results are from [32].

Dataset	**Baseline**	**with PM**	**with NA**
FFHQ [39]	0.9940	**1.04**	1.01
Diverse Dataset [32]	0.9895	**1.07**	1.02
WIKI [99]	0.9786	**1.11**	1.01
IMDB [99]	0.9661	**1.21**	1.00
LFW [100]	0.9536	**1.11**	1.03
UTK [101]	0.9418	**1.11**	1.02

Query-based ISS_{intra} Analysis for Various Language-Location Pairs:

Table 6.14 provides a detailed analysis of the ISS_{intra} values across different queries and language-location pairs, further comparing our PM and NA methods to the baseline. Across nearly all queries, both approaches show improvements over the baseline, with the "with NA" method often slightly outperforming "with PM" in specific regions and queries.

For the CEO query, the NA method demonstrates the highest improvements, particularly in the Arabic-West Asia & North Africa region, achieving an ISS$intra$ of 0.9923, significantly surpassing the baseline (0.8990). A similar trend is observed

Table 6.14: Image Similarity Score across all possible queries. Baseline results are from [32].

Query	**Language Location Pair**	**ISS**$_{Intra}$		
		Baseline	**with PM**	**with NA**
CEO	Arabic-West Asia & North Africa	0.8990	0.9901	**0.9923**
	English-North America	0.9690	**0.9724**	0.9711
	English-West Europe	0.9295	0.9558	**0.9571**
	Hindi-South Asia	0.9978	0.9993	**0.9998**
	Indonesian-SE Asia	0.9837	0.9926	**0.9931**
	Mandarin-East Asia	0.9895	0.9974	**0.9986**
	Russian-East Europe	0.9597	0.9962	**0.9977**
	Spanish-Latin America	0.9747	0.9904	**0.9933**
	Swahili-Sub Saharan Africa	0.9771	0.9917	**0.9939**
Engineer	Arabic-West Asia & North Africa	0.9864	0.9946	**0.9967**
	English-North America	0.9883	**1.0021**	1.0014
	English-West Europe	1.0009	**1.0017**	1.0009
	Hindi-South Asia	1.0031	1.0025	**1.0039**
	Indonesian-SE Asia	0.9872	0.9885	**0.9899**
	Mandarin-East Asia	0.9911	0.9921	**0.9935**
	Russian-East Europe	1.0072	**1.0082**	1.0076
	Spanish-Latin America	0.9850	0.9967	**0.9981**
	Swahili-Sub Saharan Africa	0.9837	0.9952	**0.9961**
Nurse	Arabic-West Asia & North Africa	1.0026	1.0029	**1.0032**
	English-North America	0.9716	0.9921	**0.9936**
	English-West Europe	0.9956	0.9974	**0.9985**
	Hindi-South Asia	0.9845	0.9959	**0.9973**
	Indonesian-SE Asia	0.9759	0.9925	**0.9941**
	Mandarin-East Asia	0.9890	0.9972	**0.9985**
	Russian-East Europe	0.9980	0.9972	**0.9983**
	Spanish-Latin America	1.0006	**1.0011**	1.0007
	Swahili-Sub Saharan Africa	0.9585	0.9937	**0.9955**
Politician	Arabic-West Asia & North Africa	0.9773	0.9942	**0.9955**
	English-North America	0.9959	**0.9984**	0.9976
	English-West Europe	0.9794	0.9954	**0.9967**
	Hindi-South Asia	0.9799	0.9929	**0.9941**
	Indonesian-SE Asia	0.9723	0.9916	**0.9924**
	Mandarin-East Asia	0.9763	0.9948	**0.9961**
	Russian-East Europe	0.9384	**0.9987**	0.9982
	Spanish-Latin America	0.9885	**0.9941**	0.9935
	Swahili-Sub Saharan Africa	0.9436	**0.9978**	0.9971
School Teacher	Arabic-West Asia & North Africa	1.0143	1.0149	**1.0155**
	English-North America	0.9977	0.9976	**0.9981**
	English-West Europe	0.9401	0.9987	**0.9998**
	Hindi-South Asia	1.0000	**1.0006**	1.0003
	Indonesian-SE Asia	0.9860	**1.0015**	1.0012
	Mandarin-East Asia	1.0086	**1.0092**	1.0087
	Russian-East Europe	0.9762	0.9948	**0.9955**
	Spanish-Latin America	0.9659	0.9975	**0.9982**
	Swahili-Sub Saharan Africa	0.9859	**1.0030**	1.0024

for the Engineer query, where the NA method outperforms PM in most regions, especially for Russian-East Europe and Spanish-Latin America, where NA achieves ISS*intra* values of 1.0076 and 0.9981, respectively. For the Nurse query, the NA method again consistently outperforms both the baseline and PM, with a remarkable improvement in the Swahili-Sub Saharan Africa region, where the ISS_{intra} increases from 0.9585 (baseline) to 0.9955. The Politician query also shows substantial gains with both approaches, particularly in Russian-East Europe, where the NA method reaches an ISS_{intra} of 0.9982, an increase over the baseline of 0.9383. Finally, for the School Teacher query, both methods show increased performance in nearly all regions, with the NA method showing slightly higher ISS_{intra} scores, particularly in Arabic-West Asia & North Africa (1.0155) and Spanish-Latin America (0.9982).

Table 6.15: Overall Image Similarity Score for Professions. Baseline results are from [32].

	ISS_{intra}			ISS_{cross}		
Query	Baseline	with PM	with NA	Baseline	with PM	with NA
CEO	0.9644	**0.9873**	0.9862	0.9846	0.9956	**0.9960**
Engineer	0.9925	0.9980	**0.999**	0.9939	0.9972	**0.9980**
Nurse	0.9862	**0.9967**	0.9931	0.9900	0.9961	**0.9965**
Politician	0.9724	**0.9953**	0.9930	0.9836	**0.9964**	0.9952
School Teacher	0.9860	**1.0020**	1.0010	0.9904	**0.9977**	0.9931
Mean Value	0.9803	**0.9958**	0.9944	0.9885	**0.9966**	0.9957

Overall ISS Performance: Intra-dataset and Cross-dataset Comparison:

As presented in Table 6.15, both approaches, PM and NA, consistently outperform the baseline in both ISS_{intra} and ISS_{cross} evaluations. For the CEO query, the cross-dataset score (ISS_{cross}) for the NA method is slightly higher (0.9960) compared to PM (0.9956), showing a marginal improvement over the baseline. A similar pattern is observed for the Engineer and Politician queries, where the NA method again shows higher cross-dataset performance. For Nurse and School Teacher, the PM method performs slightly better in ISS_{intra}, but the NA method maintains higher cross-dataset scores. This is particularly evident for the School Teacher query, where

the NA method scores 1.001 in ISS_{intra} and 0.9931 in ISS_{cross}, outperforming both the baseline and PM.

When averaged across all queries, the PM method achieves the highest mean ISS_{intra} score (0.9958), while the NA method follows closely with 0.9944, both outperforming the baseline (0.9803). Similarly, for ISS_{cross}, PM leads with 0.9966, followed closely by NA (0.9957), both again surpassing the baseline value of 0.9885. These results emphasise the effectiveness of our methods, particularly the PM approach, in increasing dataset diversity both within and across datasets.

Bias reduction in CNNs and ViTs: As shown in Table 6.16, CNN models demonstrated improvements in accuracy with the application of the Uniform Noise Blur technique, and in two cases (Inception V3 and Xception) with Partial Mix, though none surpassed the performance of manually debiased data. Inception V3 and Xception showed the most consistent gains, while VGG16 saw only minor improvement. In contrast, Vision Transformers (ViTs) showed no improvements with either augmentation method, indicating that these techniques were ineffective in reducing bias for ViTs. This highlights the need for more tailored approaches for bias mitigation in ViTs, which likely depend on cues beyond facial features.

Table 6.16: Accuracy of all models on the gender-balanced test dataset. Accuracies higher than the biased dataset are in bold.

Model Type	Model	Biased Accuracy	Partial Mix	Uniform Noise Blur	Unbiased Accuracy
CNN	Inception V3	0.72	**0.73**	**0.74**	0.79
	ResNet 152	0.76	0.76	**0.77**	0.85
	VGG 16	0.57	0.56	**0.58**	0.66
	Xception	0.74	**0.75**	**0.75**	0.79
ViT	ViT B/16	0.55	0.52	0.55	0.57
	ViT B/32	0.50	0.50	0.49	0.57
	ViT L/16	0.39	0.37	0.39	0.40
	ViT L/32	0.56	0.54	0.54	0.60

Table 6.17: Summary of the proposed augmentation methods, their empirical behaviour and practical advantages and limitations. FSA, FKOA, PM and NA represent FaceSaliencyAug, FaceKeepOriginalAugment, Partial Mix and Noise addition, respectively.

Method	Behaviour on diversity metrics	Behaviour on gender-bias metrics (IIAS)	Advantages and limitations
FSA	Moderate but consistent improvements in ISS_{intra} and ISS_{cross} over the baseline and RSMDA (Tables 6.1 and 6.3). Diversity gains are positive but smaller than those of FaceKeepOriginalAugment.	Strongest bias mitigation overall. Achieves large reductions in IIAS for both CNNs and ViTs (Table 6.4), including up to ∼12–34× reduction for CNNs in masked setting and up to ∼37–53× for ViTs in unmasked settings. Particularly effective on unmasked, gender-balanced data.	**Advantages:** Very effective at reducing gender bias while leaving most of the image intact; simple to integrate into existing pipelines; does not require complex hyperparameter tuning. **Limitations:** Operates only on the facial region and therefore does not directly address non-facial cues (e.g. clothing, background, objects); diversity improvements are weaker than with FaceKeepOriginalAugment.
FKOA	Largest gains in diversity. Consistently improves ISS_{intra} and ISS_{cross} across face datasets and profession datasets (Tables 6.6, 6.7, and 6.8). Also improves the proposed fairness–diversity metric $M_{\text{fairness-diversity}}$ on multiple datasets and language–location pairs (Tables 6.10–6.12).	Substantial bias reduction for both CNNs and ViTs, though typically less extreme than FSA on unmasked ViTs. Across masked / unmasked and biased / unbiased settings, FKOA reduces total absolute IIAS by roughly 4–21× (Table 6.9).	**Advantages:** Best overall method for simultaneously increasing diversity and improving fairness; works across many datasets (faces, language–location pairs, professions); preserves original content while enriching it with augmented variants. **Limitations:** More complex than FSA (multiple placement and augmentation strategies, plus hyperparameters); bias reduction on some ViT settings is weaker than for FSA; slightly higher implementation and computational overhead.
PM	Very strong increases in ISS_{intra} across all face datasets (Table 6.13), often the highest diversity scores, and improved ISS_{cross} for most professions (Table 6.15).	PM leads to small accuracy gains for some CNNs (Table 6.16), but does not consistently reduce bias for ViTs.	**Advantages:** Simple to implement; highly effective at increasing visual diversity; can slightly improve CNN accuracy on biased data. **Limitations:** Changes facial identity and semantics more aggressively; effect on fairness is not as systematically analysed as for FSA / FKOA; does not improve ViT bias and may not be suitable when identity preservation is important.
NA	Improves ISS_{intra} and ISS_{cross} over the baseline on most datasets and professions (Tables 6.13 and 6.15), but typically less than Partial Mix.	Similar to PM, shows small accuracy gains for some CNNs (Table 6.16), but no clear improvement for ViTs. Bias reduction effects are weaker and less consistent than for FSA / FKOA.	**Advantages:** Very easy to implement; inexpensive; can be combined with other augmentations; modest gains in diversity and CNN performance. **Limitations:** Limited impact on bias, especially for ViTs; may introduce noise that does not correspond to realistic variations; overall less effective than FSA and FKOA for fairness.

6.3.7 Summary of augmentation methods

In this chapter we investigated several saliency-based augmentation strategies from the perspective of diversity and bias. Our two main methods are FaceSaliencyAug (FSA), which performs saliency-guided erasing within the facial region, and FaceKeepOriginalAugment (FKOA), which mixes original and augmented content by pasting (augmented) salient regions into non-salient areas. In addition, we explored two adversarial augmentations, Partial Mix (PM) and Noise Addition (NA), mainly as complementary strategies. Table 6.17 summarises the behaviour of these methods across the main evaluation criteria: diversity (ISS_{intra}, ISS_{cross} and the proposed fairness–diversity metric), gender bias reduction (IIAS), and impact on accuracy.

In terms bias mitigation, FSA provides the strongest overall reduction in IIAS, especially for Vision Transformers on unmasked, gender-balanced data, where reductions of up to approximately 53-fold are observed relative to the baseline (Table 6.4). FKOA also consistently reduces bias (typically by 4–21× across CNNs and ViTs; Table 6.9), but its main strength lies in different prospective.

From a diversity perspective, FKOA achieves the largest gains. It systematically increases ISS_{intra} and ISS_{cross} across facial datasets and profession datasets (Tables 6.6 and 6.8), and also improves the proposed fairness–diversity metric $M_{\text{fairness-diversity}}$ for both generic face datasets and language–location pairs (Tables 6.10–6.12). FSA also improves ISS scores, but the effect is more modest compared to FKOA.

The adversarial augmentations, PM and NA, mainly act as supporting methods. Partial Mix produces the highest ISS_{intra} values across several datasets (Table 6.13) and often improves ISS_{cross} (Table 6.15); it also yields small accuracy gains for some CNNs (Table 6.16). Noise Addition generally has a weaker effect than Partial Mix but still improves diversity and accuracy for some CNNs. However, neither PM nor NA substantially reduce bias in ViTs, which appear to rely on broader contextual cues beyond the face. Practically, if the primary goal is maximising bias reduction with minimal change to the overall image, FSA is the preferred choice. If the goal is to maximise dataset diversity and fairness across gender, geography and profession,

FKOA is more suitable. PM and NA can be used as additional augmentations for CNNs when further diversity and small accuracy gains are desired, but they should be seen as complementary to FSA and FKOA rather than as replacements.

6.4 Conclusion

This chapter examined how data augmentation can be used to improve fairness in computer vision models by mitigating geographical, gender, and stereotypical biases.. To do this, two data augmentation methods were proposed. The first, FaceSaliencyAug, leverages salient region detection to mask and restore important facial regions, thereby improving data diversity and mitigating bias. The second, FaceKeepOriginalAugment, combines salient and non-salient areas, striking a balance between preserving image features and enhancing diversity. Both methods were tested across a variety of datasets, including FFHQ, WIKI, IMDB, LFW, UTK Faces, and a custom Diverse Dataset. Results demonstrated improvements in diversity metrics (ISS_{intra} and ISS_{inter}), as well as a significant reduction in gender bias across four professions, as measured by the IIAS metric for both CNNs and ViTs.

Building on these findings, we explored the use of adversarial data augmentation techniques—Partial Mix (PM) and Noise Addition (NA)—which were shown to further enhance dataset diversity and reduce gender bias, particularly in CNN models. These techniques yielded significant improvements in both intra- and cross-dataset image similarity scores (ISS), indicating better representation of diverse gender features.

Furthermore, we introduced a saliency-based diversity and fairness metric that offers a more nuanced evaluation of gender, location, and profession-related biases. This metric provides a deeper insight into how well the models perform in mitigating these biases. Together, these findings underscore the importance of addressing biases in computer vision models, while highlighting that ViTs require more comprehensive strategies that go beyond facial-region augmentation. Future work should focus on developing such strategies to ensure fair and equitable representation across all

model architectures.

Chapter 7

Conclusion

7.1 Introduction

This chapter concludes the thesis by revisiting the four key research questions. First, it highlights the central problem addressed in the thesis, followed by a summary of each question and its impact on data augmentation. Lastly, we discuss potential future directions for research in these areas.

7.2 Introduction

This thesis introduced novel image data augmentation methods designed to enhance the generalisation, robustness, and fairness of deep learning models in computer vision. The proposed techniques—Random Slices Mixing Data Augmentation (RSMDA), RandSaliencyAug, and KeepOriginalAugment—aim to improve model accuracy while also addressing the ethical concerns associated with biases in AI systems. In this conclusion, we summarise the key findings of the thesis, outline the contributions made in each chapter, and discuss the challenges and future directions for further research.

7.3 Random Slices Mixing Data Augmentation (RSMDA)

Addressing Hypothesis 1 (H1) and Research Question 1 (RQ1), Chapter 3, we introduced RSMDA, a new data augmentation technique that combines image slices to reduce feature loss and improve model generalisation. We proposed three strategies for mixing image slices: row-wise slicing (RSMDA-R), column-wise slicing (RSMDA-C), and combined row-column slicing (RSMDA-RC). Our extensive experiments on datasets such as CIFAR-10, CIFAR-100, and FashionMNIST demonstrated that RSMDA significantly outperforms traditional methods like random erasing in terms of model generalisation and accuracy. The method also showed improved robustness in the face of adversarial attacks, confirming its practical utility in real-world applications. Taken together, these results confirm H1 and provide a positive answer to RQ1: mixing row, column, and combined slices from different images is an effective way to reduce feature loss and enhance feature diversity, leading to improved model generalisation.

7.4 RandSaliencyAug

In line with Hypothesis 2 (H2) and Research Question 2 (RQ2), Chapter 4, we proposed RandSaliencyAug, a saliency-guided data augmentation technique that detects salient region, applies on the six proposed masking on that salient region to prevent overfitting while preserving critical contextual information. This method was shown to outperform existing techniques, such as Cutout and Random Erasing, by offering a balanced trade-off between feature diversity and contextual integrity. Through experiments on datasets like CIFAR-10 and CIFAR-100, we demonstrated that RandSaliencyAug can improve model robustness and accuracy without compromising the retention of important features. This technique was particularly effective in reducing overfitting on smaller datasets, making it a valuable tool for enhancing model performance across a wide range of tasks. Overall, these findings support H2 and address RQ2, showing that saliency-aware erasing can be designed to balance

feature loss and contextual information loss, while RQ2.1 and RQ2.2 are specifically answered by the way RandSaliencyAug handles occlusion and overfitting through targeted masking of salient regions.

7.5 KeepOriginalAugment

Chapter 5 tackled Hypothesis 3 (H3) and Research Question 3 (RQ3) by presenting KeepOriginalAugment, a data augmentation strategy that integrates salient regions with non-salient areas in images to optimise both data diversity and information retention. This method was shown to outperform other selective augmentation techniques, such as KeepAugment, by reducing domain shifts between salient and non-salient regions, which could potentially hinder model learning. Through evaluations on CIFAR-10, TinyImageNet, and other datasets, we observed that KeepOriginalAugment significantly improved model performance and robustness, particularly in large-scale image classification tasks. Additionally, it demonstrated superior efficiency compared to traditional methods, making it suitable for a range of deep learning models. These results validate H3 and provide concrete answers to RQ3.1 and RQ3.2: carefully choosing both the placement strategy and the type of augmented region (salient, non-salient, or both) allows us to increase diversity while preserving contextual information, thereby improving model generalisation.

7.6 FaceSaliencyAug and FaceKeepOriginalAugment

Focusing on Hypothesis 4 (H4) and Research Question 4 (RQ4), Chapter 6 addressed the ethical implications of data augmentation in AI, particularly the issue of bias We extended our proposed augmentation methods to reduce gender and geographical biases in computer vision models, evaluating their effectiveness using masked and unmasked datasets. These datasets simulate real-world scenarios where data may be occluded or incomplete, providing a robust testing ground for assessing fairness. In addition, we used professional datasets (e.g., CEO, Engineer, Nurse, and Teacher) to

test how well our methods mitigate bias across different occupations. We introduced two additional data augmentation techniques, Partial Mix and Noise Addition, to enhance diversity and reduce bias.

In summary, the results in Chapter 6 provide strong evidence for **H4** and address RQ4.1 and RQ4.2: data augmentation that respects salient and non-salient facial regions, as well as their placement, can mitigate gender, geographical, and occupational stereotypical biases while maintaining competitive performance.

7.7 Saliency-Based Diversity Fairness Metric

To support a more principled analysis of fairness in the context of RQ4, we introduced the Saliency-Based Diversity Fairness Metric (SBD-FM), a novel approach to quantitatively measure fairness and assess how well our techniques reduce bias in model predictions. This directly relates to RQ4.3, which asks how to capture and measure fairness and diversity regardless of whether the data is balanced or imbalanced.

Our results showed significant improvements in fairness, with the proposed techniques successfully reducing bias in both facial recognition and occupational classification tasks, while SBD-FM provided a saliency-aware lens on the trade-offs between diversity and fairness. Thus, SBD-FM answers RQ4.3 by offering a metric that remains informative under both balanced and imbalanced data distributions.

7.8 Limitations

While the proposed methods show promising results, there are several challenges and limitations, which also clarify the boundaries of our answers to the research questions. The computational complexity of these techniques, particularly when applied to larger datasets such as ImageNet, can be a limiting factor in practice, especially for large-scale deployment. Additionally, while the methods introduced in this thesis significantly reduce bias, they do not completely eliminate all forms of

bias present in the training data. There remain residual and potentially emergent biases that require further exploration, especially in more diverse and real-world settings and for additional demographic attributes beyond those considered here.

7.9 Future Directions

Future work could explore ways to improve the computational efficiency of the proposed augmentation strategies, for example through more efficient architectures, model compression techniques, or hardware accelerations, thereby extending the practical impact of the answers to RQ1–RQ3. Additionally, integrating SBD-FM with other fairness evaluation frameworks could offer a more comprehensive understanding of the ethical implications involved in data augmentation and further refine the insights related to RQ4.

Further investigation into how these methods can be applied to multimodal datasets (e.g., combining images with text or audio) could also open up new areas of research. Moreover, hybrid methods that combine RSMDA, RandSaliencyAug, and KeepOriginalAugment—and their fairness-oriented variants—could be explored to further enhance the robustness and fairness of models, building on the positive findings for H1–H4. Finally, extending these ideas to real-time or continual-learning settings, and to domains beyond standard image classification, represents an important direction for future work and a natural continuation of the research agenda set out in Chapter 1.

Appendix A

Publications

A.1 Publication

Publications work from the thesis:

A.2 Journal Publications

- Kumar, Teerath, et al. *Rsmda: Random slices mixing data augmentation.* Applied Sciences 13.3 (2023): 1711.

- Kumar, Teerath, et al. *Image data augmentation approaches: A comprehensive survey and future directions.* IEEE Access (2024).

- Kumar, Teerath, Alessandra Mileo, and Malika Bendechache. *Facesaliencyaug: mitigating geographic, gender and stereotypical biases via saliency-based data augmentation.* Signal, Image and Video Processing 19.1 (2025): 1-11.

- Kumar, Teerath, Alessandra Mileo, and Malika Bendechache. *Saliency-based metric and FaceKeepOriginalAugment: a novel approach for enhancing fairness and Diversity.* Multimedia Systems 31.2 (2025): 1-14.

A.3 Conference Publication

- Kumar, Teerath, Rob Brennan, and Malika Bendechache. *Stride random erasing augmentation.* CS & IT Conference Proceedings. Vol. 12. No. 2. CS & IT Conference Proceedings, 2022.

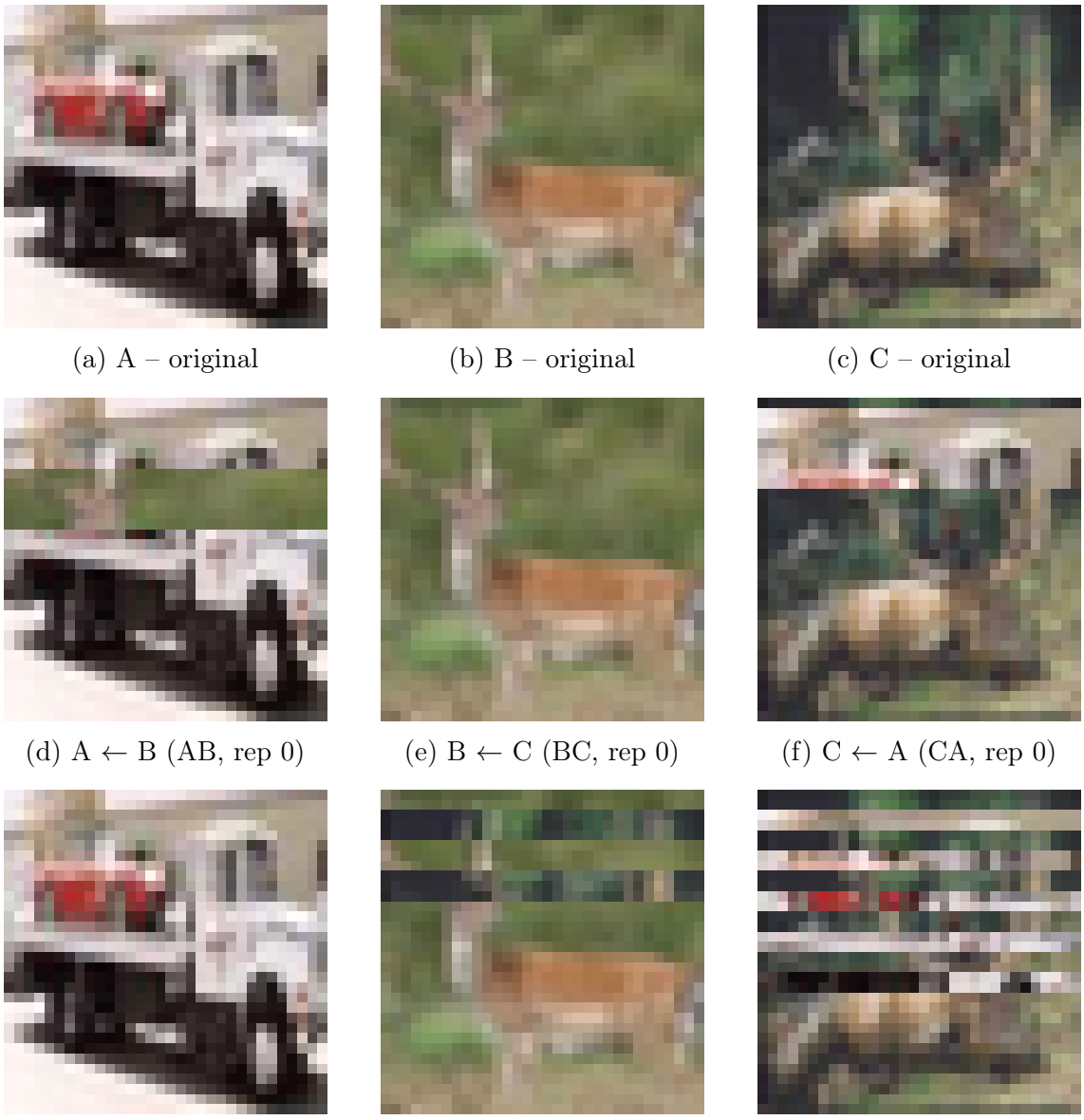
- Kumar, Teerath, Alessandra Mileo, and Malika Bendechache. *Keeporiginalaugment: Single image-based better information-preserving data augmentation approach.* IFIP International Conference on Artificial Intelligence Applications and Innovations. Cham: Springer Nature Switzerland, 2024.

- Kumar, T., Mileo, A. and Bendechache, M. (2025). *RandSaliencyAug: Balancing Saliency-Based Data Augmentation for Enhanced Generalization.* In Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP

- Kumar, Teerath, et al. *Debiasing Computer Vision Models using Data Augmentation based Adversarial Techniques.* (2024).

- Kumar, Teerath, Alessandra Mileo, and Malika Bendechache. *RandSaliencyAug++: Saliency-Based Data Augmentation for Improved Generalization Across Diverse Tasks* IMVIP 2025 (Published)

A.4 Supplimentary material

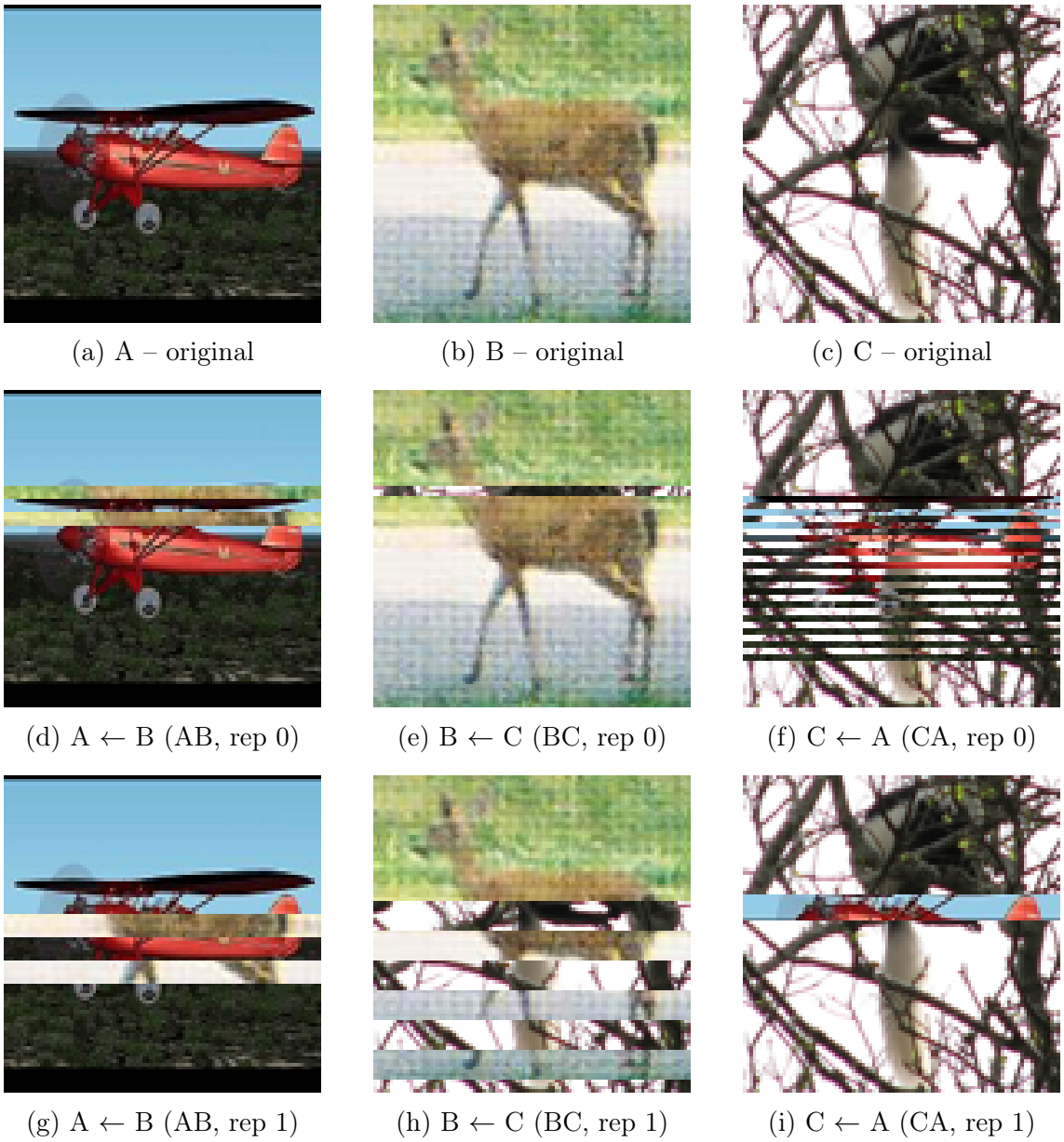


(a) A – original (b) B – original (c) C – original

(d) A ← B (AB, rep 0) (e) B ← C (BC, rep 0) (f) C ← A (CA, rep 0)

(g) A ← B (AB, rep 1) (h) B ← C (BC, rep 1) (i) C ← A (CA, rep 1)

Figure A.1: RSMDA visualisation on low-resolution CIFAR-10 images (32×32). Top row: original images A, B, C. Middle row: RSMDA samples for pairs A←B, B←C, and C←A (first random draw, rep 0). Bottom row: second independent draw (rep 1), illustrating the stochastic nature of the augmentation at the same resolution.

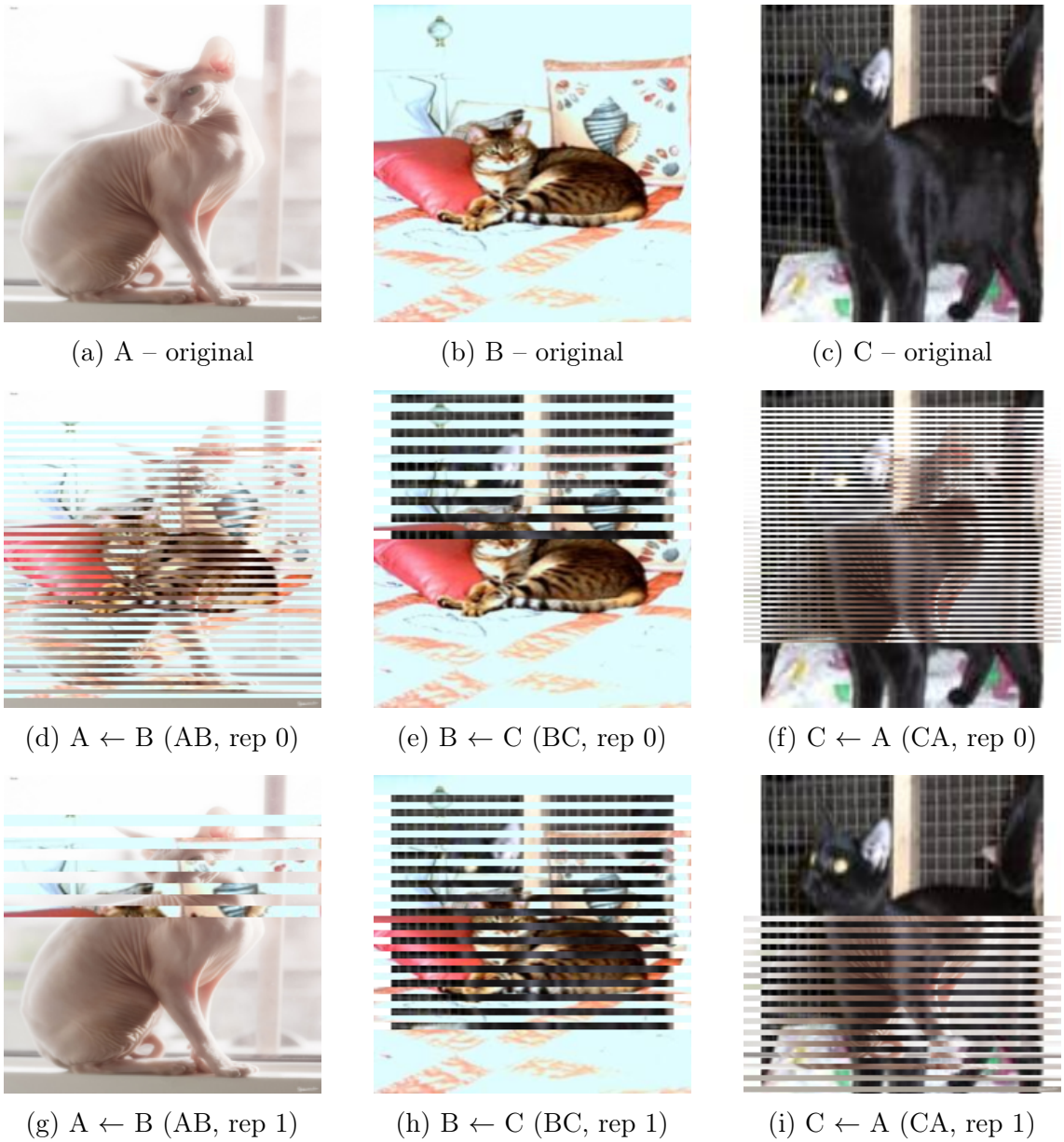


(a) A – original (b) B – original (c) C – original

(d) A ← B (AB, rep 0) (e) B ← C (BC, rep 0) (f) C ← A (CA, rep 0)

(g) A ← B (AB, rep 1) (h) B ← C (BC, rep 1) (i) C ← A (CA, rep 1)

Figure A.2: RSMDA visualisation on medium-resolution STL-10 images (96×96). Top row: originals. Middle and bottom rows: two independent RSMDA realisations for the same pairs A←B, B←C, and C←A.



(a) A – original (b) B – original (c) C – original

(d) A ← B (AB, rep 0) (e) B ← C (BC, rep 0) (f) C ← A (CA, rep 0)

(g) A ← B (AB, rep 1) (h) B ← C (BC, rep 1) (i) C ← A (CA, rep 1)

Figure A.3: RSMDA visualisation on high-resolution Oxford-IIIT Pet images (224×224). Top row: originals. Middle and bottom rows: two independent samples of the same RSMDA pairings A←B, B←C, and C←A, showing that at higher resolution the augmentation remains stochastic but less semantically destructive than at 32×32.

Bibliography

[1] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[3] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[4] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[5] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[6] Xiaolong Liu, Zhidong Deng, and Yuhan Yang. "Recent progress in semantic image segmentation". In: *Artificial Intelligence Review* 52.2 (2019), pp. 1089–1106.

[7] Victor Lempitsky et al. "Image segmentation with a bounding box prior". In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 277–284.

[8] Jiss Kuruvilla et al. "A review on image processing and image segmentation". In: *2016 international conference on data mining and advanced computing (SAPIENCE)*. IEEE. 2016, pp. 198–203.

[9] JunHao Liew et al. "Regional interactive image segmentation networks". In: *2017 IEEE international conference on computer vision (ICCV)*. IEEE Computer Society. 2017, pp. 2746–2754.

[10] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1 (2019), pp. 1–48.

[11] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[12] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[13] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[14] Sangdoo Yun et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6023–6032.

[15] Jan Kukačka, Vladimir Golkov, and Daniel Cremers. "Regularization for deep learning: A taxonomy". In: *arXiv preprint arXiv:1710.10686* (2017).

[16] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[17] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[18] Z. Zhong et al. "Random erasing data augmentation". In: *Proceedings Of The AAAI Conference On Artificial Intelligence*. Vol. 34. 2020, pp. 13001–13008.

[19] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. "Data augmentation using random image cropping and patching for deep CNNs". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.9 (2019), pp. 2917–2931.

[20] Agnieszka Mikołajczyk and Michał Grochowski. "Data augmentation for improving deep learning in image classification problem". In: *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE. 2018, pp. 117–122.

[21] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. "A group-theoretic framework for data augmentation". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 9885–9955.

[22] Jason Wei and Kai Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks". In: *arXiv preprint arXiv:1901.11196* (2019).

[23] Álvaro Acción, Francisco Argüello, and Dora B Heras. "Dual-window superpixel data augmentation for hyperspectral image classification". In: *Applied Sciences* 10.24 (2020), p. 8833.

[24] Krishna Kumar Singh and Yong Jae Lee. "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3524–3533.

[25] Pengguang Chen et al. "Gridmask data augmentation". In: *arXiv preprint arXiv:2001.04086* (2020).

[26] Terrance DeVries and Graham W Taylor. "Improved regularization of convolutional neural networks with cutout". In: *arXiv preprint arXiv:1708.04552* (2017).

[27] Hongyi Zhang et al. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).

[28] Teerath Kumar et al. “Image data augmentation approaches: A comprehensive survey and future directions”. In: *IEEE Access* (2024).

[29] AFM Uddin et al. “Saliencymix: A saliency guided data augmentation strategy for better regularization”. In: *arXiv preprint arXiv:2006.01791* (2020).

[30] J. Choi et al. “SalfMix: a novel single image-based data augmentation technique using a saliency map”. In: *Sensors* 21.8444 (2021).

[31] Chengyue Gong et al. “Keepaugment: A simple information-preserving data augmentation approach”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 1055–1064.

[32] Abhishek Mandal, Susan Leavy, and Suzanne Little. “Dataset diversity: measuring and mitigating geographical bias in image search and retrieval”. In: *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing*. 2021, pp. 19–25.

[33] Abhishek Mandal, Suzanne Little, and Susan Leavy. “Gender bias in multimodal models: A transnational feminist approach considering geographical region and culture”. In: *arXiv preprint arXiv:2309.04997* (2023).

[34] Joy Buolamwini and Timnit Gebru. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[35] Kimmo Karkkainen and Jungseock Joo. “Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 1548–1558.

[36] A. Birhane, V. Prabhu, and E. Kahembwe. “Multimodal datasets: misogyny, pornography, and malignant stereotypes”. In: *ArXiv Preprint ArXiv:2110.01963* (2021).

[37] Peter Smith and Karl Ricanek. "Mitigating Algorithmic Bias: Evolving an Augmentation Policy That Is Non-Biasing". In: *Proceedings Of The IEEE/CVF Winter Conference On Applications Of Computer Vision Workshops*. 2020, pp. 90–97.

[38] N. Norori et al. "Addressing bias in big data and AI for health care: A call for open science". In: *Patterns* 2 (2021).

[39] T. Karras, S. Laine, and T. Aila. *NVlabs/ffhq-dataset*. 2019. URL: `https://github.com/NVlabs/ffhq-dataset`.

[40] Yi Zhang and Jitao Sang. "Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 4346–4354.

[41] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. "Biaswap: Removing dataset bias with bias-tailored swapping augmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14992–15001.

[42] Jungsoo Lee et al. "Learning debiased representation via disentangled feature augmentation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25123–25133.

[43] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[44] Aleksander Madry et al. "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083* (2017).

[45] Geoffrey E Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (2012).

[46] Jimmy Ba and Brendan Frey. "Adaptive dropout for training deep neural networks". In: *Advances in neural information processing systems* 26 (2013).

[47] Li Wan et al. "Regularization of neural networks using dropconnect". In: *International conference on machine learning*. PMLR. 2013, pp. 1058–1066.

[48] Matthew D Zeiler and Rob Fergus. "Stochastic pooling for regularization of deep convolutional neural networks". In: *arXiv preprint arXiv:1301.3557* (2013).

[49] Jonathan Tompson et al. "Efficient object localization using convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 648–656.

[50] Cecilia Summers and Michael J Dinneen. "Improved mixed-example data augmentation". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 1262–1270.

[51] Shiori Sagawa et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization". In: *arXiv preprint arXiv:1911.08731* (2019).

[52] E. Cubuk et al. "Randaugment: Practical automated data augmentation with a reduced search space". In: *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition Workshops*. 2020, pp. 702–703.

[53] Kimmo Kärkkäinen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age". In: *ArXiv Preprint ArXiv:1908.04913* (2019).

[54] Alan Wang et al. "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets". In: *International Journal Of Computer Vision* 130 (2022), pp. 1790–1810.

[55] Abhijit Mandal, Sarah Leavy, and Suzanne Little. "Measuring Bias in Multimodal Models: Multimodal Composite Association Score". In: *International Workshop On Algorithmic Bias In Search And Recommendation*. 2023, pp. 17–30.

[56] Faisal Kamiran and Toon Calders. "Data preprocessing techniques for classification without discrimination". In: *Knowledge and Information Systems*. Vol. 33. 1. 2012, pp. 1–33. DOI: 10.1007/s10115-011-0463-8.

[57] Michael Feldman et al. "Certifying and removing disparate impact". In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 259–268. DOI: 10.1145/2783258.2783311.

[58] Alekh Agarwal et al. "A reductions approach to fair classification". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. ICML. PMLR, 2018, pp. 60–69.

[59] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016, pp. 3315–3323.

[60] Ling Li et al. "Bias Oriented Unbiased Data Augmentation for Cross-Bias Representation Learning". In: *Multimedia Systems* 29 (2023), pp. 725–738.

[61] Dongyoon Han, Jiwhan Kim, and Junmo Kim. "Deep pyramidal residual networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5927–5935.

[62] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[63] Adam Coates, Andrew Ng, and Honglak Lee. "An analysis of single-layer networks in unsupervised feature learning". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 215–223.

[64] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747* (2017).

[65] Gao Huang et al. "Deep networks with stochastic depth". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer. 2016, pp. 646–661.

[66] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[67] Vikas Verma et al. "Manifold mixup: Better representations by interpolating hidden states". In: *International conference on machine learning*. PMLR. 2019, pp. 6438–6447.

[68] Yoshihiro Yamada et al. "Shakedrop regularization for deep residual learning". In: *IEEE Access* 7 (2019), pp. 186126–186136.

[69] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[70] Akshay Agarwal, Richa Singh, and Mayank Vatsa. "The Role of'Sign'and'Direction'of Gradient on the Performance of CNN". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 646–647.

[71] Bolei Zhou et al. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[72] Peng-Tao Jiang et al. "Layercam: Exploring hierarchical class activation maps for localization". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 5875–5888.

[73] Hyungsik Jung and Youngrock Oh. "Towards better explanations of class activation mapping". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1336–1344.

[74] S. Montabone and A. Soto. "Human detection using a mobile platform and novel features derived from a visual saliency mechanism". In: *Image And Vision Computing* 28 (2010), pp. 391–402.

[75] Devesh Walawalkar et al. "Attentive Cutmix: An Enhanced Data Augmentation Approach for Deep Learning Based Image Classification". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*. 2020.

[76] Sergey Zagoruyko and Nikos Komodakis. "Wide Residual Networks". In: *British Machine Vision Conference 2016*. British Machine Vision Association. 2016.

[77] Xavier Gastaldi. "Shake-shake regularization". In: *arXiv preprint arXiv:1705.07485* (2017).

[78] Yann Le and Xuan Yang. "Tiny imagenet visual recognition challenge". In: *CS 231N* 7.7 (2015), p. 3.

[79] Teerath Kumar, Alessandra Mileo, and Malika Bendechache. "KeepOriginalAugment: Single Image-based Better Information-Preserving Data Augmentation Approach". In: *20th International Conference on Artificial Intelligence Applications and Innovations*. 2024.

[80] M. Everingham et al. *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/inde

[81] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. "A-fast-rcnn: Hard positive generation via adversary for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2606–2615.

[82] Teerath Kumar et al. "RSMDA: Random Slices Mixing Data Augmentation". In: *Applied Sciences* 13.3 (2023), p. 1711.

[83] Minghui Liu et al. "Focuseddropout for convolutional neural network". In: *Applied Sciences* 12.15 (2022), p. 7682.

[84] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. "Dropblock: A regularization method for convolutional networks". In: *Advances in neural information processing systems* 31 (2018).

[85] Ekin D Cubuk et al. "Autoaugment: Learning augmentation strategies from data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 113–123.

[86] Dan Hendrycks et al. "Augmix: A simple data processing method to improve robustness and uncertainty". In: *arXiv preprint arXiv:1912.02781* (2019).

[87] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. "Fractalnet: Ultra-deep neural networks without residuals". In: *arXiv preprint arXiv:1605.07648* (2016).

[88] Ethan Harris et al. "Fmix: Enhancing mixed sample data augmentation". In: *arXiv preprint arXiv:2002.12047* (2020).

[89] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. "Puzzle mix: Exploiting saliency and local statistics for optimal mixup". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5275–5285.

[90] Jianchao Zhu et al. "Automix: Mixup networks for sample interpolation via cooperative barycenter learning". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer. 2020, pp. 633–649.

[91] Jie Qin et al. "Resizemix: Mixing data with preserved object information and true labels". In: *arXiv preprint arXiv:2012.11101* (2020).

[92] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[93] Jin-Woo Seo, Hong-Gyu Jung, and Seong-Whan Lee. "Self-augmentation: Generalizing deep networks to unseen classes for few-shot learning". In: *Neural Networks* 138 (2021), pp. 140–149.

[94] K Simonyan, A Vedaldi, and A Zisserman. "Deep inside convolutional networks: visualising image classification models and saliency maps". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR. 2014.

[95] Cheng-Yang Fu et al. "DSSD: Deconvolutional Single Shot Detector". In: *ArXiv Preprint ArXiv:1701.06659* (2017).

[96] Abhijit Mandal, Sarah Leavy, and Suzanne Little. "Biased Attention: Do Vision Transformers Amplify Gender Bias More Than Convolutional Neural Networks?" In: *ArXiv Preprint ArXiv:2309.08760* (2023).

[97] A. Mandal, S. Leavy, and S. Little. "Multimodal composite association score: Measuring gender bias in generative multimodal models". In: *ArXiv Preprint ArXiv:2304.13855* (2023).

[98] Teerath Kumar, Alessandra Mileo, and Malika Bendechache. "FaceSaliencyAug: mitigating geographic, gender and stereotypical biases via saliency-based data augmentation". In: *Signal, Image and Video Processing* 19.1 (2025), pp. 1–11.

[99] Rasmus Rothe, Radu Timofte, and Luc Van Gool. "Dex: Deep expectation of apparent age from a single image". In: *Proceedings of the IEEE international conference on computer vision workshops*. 2015, pp. 10–15.

[100] Erik Learned-Miller et al. "Labeled faces in the wild: A survey". In: *Advances in face detection and facial image analysis* (2016), pp. 189–248.

[101] Zhifei Zhang, Yang Song, and Hairong Qi. "Age progression/regression by conditional adversarial autoencoder". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5810–5818.

[102] Teerath Kumar, Alessandra Mileo, and Malika Bendechache. "Saliency-based metric and FaceKeepOriginalAugment: a novel approach for enhancing fairness and Diversity". In: *Multimedia Systems* 31.2 (2025), pp. 1–14.