Efficient Search with an Interactive Video Retrieval System for Novice Users in IVR4B

Thao-Nhu Nguyen[*]
Bunyarit Puangthamawathanakun[*]
thaonhu.nguyen24@mail.dcu.ie
bunyarit.puangthamawathanakun2@mail.dcu.ie
Dublin City University
Dublin, Ireland

Chonlameth Arpnikanondt
chonlameth@sit.kmutt.ac.th
King Mongkut's University of Technology Thonburi
Bangkok, Thailand

Cathal Gurrin
Graham Healy
Annalina Caputo
cathal.gurrin@dcu.ie
graham.healy@dcu.ie
annalina.caputo@adaptcentre.ie
Dublin City University
Dublin, Ireland

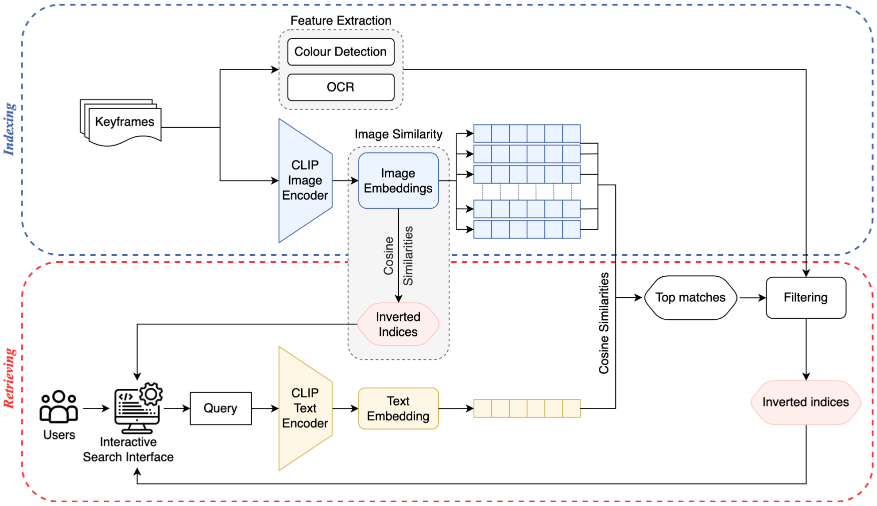

Figure 1: The system workflow, inherited from VideoFall in VBS'22, consists of two stages of indexing and retrieving.

ABSTRACT

In this paper, we present the second release of VideoCLIP, an interactive CLIP-based video retrieval system that participated in the Video Browser Showdown 2023. While we continue to use the underlying architecture to map the content between image and text, we concentrate on improving the user experience for novice users. Specifically, we have implemented three different query modalities and redesigned the user interface in order to adapt to the context of the Interactive Video Retrieval for Beginners (IVR4B) workshop. These modifications ultimately aim to provide newcomers with a simple and efficient user experience to locate the desired videos.

CCS CONCEPTS

• **Information systems → Search interfaces**; • **Human-centered computing → User interface programming**.

KEYWORDS

video retrieval, multimodal data, user interface, retrieval system

ACM Reference Format:
Thao-Nhu Nguyen, Bunyarit Puangthamawathanakun, Chonlameth Arpnikanondt, Cathal Gurrin, Graham Healy, and Annalina Caputo. 2023. Efficient Search with an Interactive Video Retrieval System for Novice Users in IVR4B. In *20th International Conference on Content-based Multimedia Indexing (CBMI 2023), September 20–22, 2023, Orléans, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3617233.3617273

1 INTRODUCTION

Interactive Video Retrieval (IVR) systems allow users to search for and retrieve video content from large video archives. Over the recent years, there have been many IVR systems developed, each of which takes a unique approach to tackling the IVR challenge. To assist in the comparative evaluation of various approaches to interactive video and multimodal data retrieval, a number of benchmarking challenges have been established, such as Video Browser Showdown (VBS) [7, 10] or Lifelog Search Challenge (LSC) [5] to

[*] Both authors contributed equally to this research.

CBMI 2023, September 20–22, 2023, Orléans, France
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0912-8/23/09.
https://doi.org/10.1145/3617233.3617273

provide large-scale datasets and evaluation protocols that measure progress and support impactful comparisons between different systems. In IVR challenges, participants are required to complete multiple tasks of finding relevant video clips given an information need (either a video clip or a textual description) within a given period of time. The system's performance will be evaluated based on the score, which measures both the accuracy of retrieval and the search time elapsed when relevant content was found. In such benchmarking challenges, these systems are usually operated by experienced system developers (called experts), rather than novice users, who would be typical of the end-users of IVR systems. To address this issue, the Interactive Video Retrieval for Beginners (IVR4B) workshop is organised to evaluate the effectiveness of video retrieval systems specifically for novice users who have little knowledge of how IVR systems work and VideoCLIP in particular. Similarly to the VBS2022 competition, the IVR4B's organisers provide a collection of 17,235 videos from two set of the Vimeo Creative Commons Collection: V3C1 [3] and V3C2 [15].

Based on a system [12] that participated in VBS2023, VideoCLIP, is a re-designed IVR system for the IVR4B workshop that is designed for novice users. In particular, the system, which exploits state-of-the-art joint image-text embedding models, enables users to search for videos using various search modalities such as free-text search, search by example, and temporal search. Meanwhile, the user interface (UI) is also being revised to offer users an intuitive and efficient search experience.

2 RELATED RESEARCH

There are many high-scoring systems that have participated in the VBS in recent years. Vibro, the winner of both VBS2022 and VBS2023, integrates state-of-the-art search features with a novel visualisation interface [8, 16]. Both systems support text querying, query-by-example, and nearest neighbour search, with year-on-year advancements in all system components. Another high performing system, VISIONE [1, 2] supports a related feature set, such as spatial colours and object search, visual similarity, semantic similarity, and free text search.

Reviewing top-performing VBS systems in recent years suggests that free text search and temporal search are required components, along with a strong embedding model supporting retrieval. The popular embedding model, CLIP [14], is a powerful joint-embedding model for both text and images and has been integrated into many state-of-the-art systems, taking top places in VBS. Our previous retrieval systems, VideoFall [13] and VideoCLIP [12], are both CLIP-integrated retrieval systems. For VideoFall, we attempted to create a hierarchical multi-user search taking advantage of diversity in terms of prompts, but the final performance is lower than we expected with regard to network instability between users and overhead on data transmission. In the next version, the multi-user feature has been removed in VideoCLIP aiming for better performance expectations with a new feature, meta-search. In this paper, we describe VideoCLIP, which is a refined version of VideoCLIP aimed at novice users.

3 SYSTEM OVERVIEW

The system workflow is inherited from our previous version of VideoFall [13] as illustrated in Figure 1. While the latest embedding models are implemented to exploit the semantic meanings between image and text alongside the metadata extracted from the given dataset, three different search modalities are integrated to support users, especially non-expert users, in looking for the desired moments in different scenarios.

3.1 Embedding Models

The CLIP model, which stands for Contrastive Language-Image Pre-Training, has gained popularity in various research fields ranging from image classification, and image similarity to image captioning due to its ability to map the semantic meanings between visual concepts and natural language. This model has proven to be very effective, and it is considered the state-of-the-art embedding model for many teams in the VBS challenge, including our team VideoCLIP. To further enhance the performance of the embedding model, we create an ensemble model from the two latest large CLIP models [14], ResNet-50 [6] model (ResNet50x64), and a Vision Transformer [4] pre-trained at 336-pixel resolution (ViT-L/14@336p). The model is responsible for converting both visual and textual information into numerical vectors in the same feature space, and then measuring the similarity between them. By implementing this model, we can support complex queries that go beyond the traditional keyword-matching queries.

3.2 Search Modalities

Free-text search. The free-text search is widely considered the most popular query modality among video retrieval systems, which allows users to type in any kind of text describing the visual content they are looking for in the target keyframes. With advances in natural language processing, input queries, natural text without requiring any specialized knowledge of the underlying retrieval algorithms, will be converted into high-dimensional feature vectors in the same space as the preprocessed visual features. By mapping pairs of textual and visual feature vectors, the system enables users to measure the similarity score between them and return the most relevant videos for the given descriptions.

Query-by-example search. The query-by-example search function provides an alternative search option when the input query is difficult to describe. This works using the FAISS library (Facebook AI Similarity Search) [9], providing efficient similarity search and clustering of dense vectors in high-dimensional feature vectors. This feature is handy when searching for specific visual elements that might be difficult to describe textually, or when seeking visually similar content to a single keyframe.

Temporal search. In addition to the two above search options, we also support temporal search, allowing users to look for a sequence of content, rather than a single item. To enable temporal search, we rank the consecutive sequence of frames from each video by measuring the possibility that they match the sequences of timestamped inputs. Initially, users will input a hint from the description as a normal text-based search. Then, each time a new hint is displayed, users will add a new sub-query based on their temporal order. Consequently, the new score will be calculated by the combination of the previous score for each video between the tabs with the same weights. For example, suppose a user has a query such as "*Hands of a kid applying glue to an egg carton and then a view of a sculpture made of those cartons. In the second shot, the camera pans up along green and turquoise egg cartons. In the first shot, we see a jar with white glue, the bottom of an egg carton, and the kid holding a brush.*". In this case, users can input the initial hint as a text-based search and then add other two subsequent sub-queries with the order in which events occurred to refine their search results. The final ranked list result will consist of half of the first query and half of the second, taking into account both the textual information and the temporal order of events specified by the users.

4 USER INTERFACE AND USER INTERACTION

4.1 User Interface

To support novice users, we prioritise improving the user interface (UI) and user experience (UX) to ease each use case while browsing. The UI of the system consists of four main areas: (**A**) the search section, (**B**) the main result section, (**C**) temporal result visualisation, and (**D**) the miscellaneous section as illustrated in Figure 2.

(A) Search Section. In this section, users may use the query input for free-text search along with OCR filtering for visible text in the content. The search section also includes three temporal tabs that are ordered chronologically, with the first tab representing the earliest point in the video, followed by the other tabs. While searching in a particular tab, users can switch between tabs freely, allowing them to quickly and easily find the desired moment with temporal information. The bottom bar is the historical action result list, which was inspired by the LifeSeeker system [11]. This bar will support users in their search by providing a reference point to return to any previous stage when necessary. For instance, the current stage in Figure 2 is depicted as "Search", and users can navigate back to any previous search or filter step without experiencing any waiting time.

(B) Main Result Section. The result from part **A** is displayed in descending order regarding similarity scores in this area. Moreover, when users switch temporal tabs in part **A**, the result in this main section will be updated accordingly, as well as navigated back and forth in the historical action result list. In some cases, the number of keyframes in the result section may exceed the height of the panel, making it necessary for users to scroll down to locate the target keyframes.

(C) Temporal Result Visualisation. Keyframes listed in this section are the result of part **A** according to the re-ranked score. Each item is based on the main result in part **B** with the score adjusted as mentioned on temporal search in Section 3.2. Temporal search will be enabled when a search section in more than one temporal tab is filled. Since each item is based on the main result, results in this section are changed when users switch temporal tabs.

(D) Miscellaneous Section. The two tabs in this area take responsibility for visualising the result list of similar keyframes and the bookmarked relevant keyframes. Unlike the main result section and temporal result section, which are dependent on the temporal tab, this section is independent, making it easier for novice users to navigate.

4.2 User Interaction

The user interaction in such a retrieval system involves a number of steps that enable users to input their queries and refine their search criteria until they find the desired image or video segment.

At the initial stage, the system presents an empty visualisation. Users can type the natural language description in the query box, which shows "*Type anything*" to begin the process. Users may also add words to the OCR filter box to narrow down the search space. Afterwards, the result panel will display the score-based ranked list result of images matching the given input, enabling users to navigate through it to find the most relevant one. Each displayed keyframe will have four buttons corresponding to submit, bookmark, query-by-example, and view all keyframes, providing users with further navigation options.

As per task requirements, some queries are based on chronological order, and users may utilise temporal search by using the free-text search on each temporal tab. When users input all the queries needed in timely order, they may switch temporal tabs to adjust the temporal result list to find the target moment.

Novice users may have difficulty expressing the right prompts, and free-text search may not always be possible to find the target keyframes. On the other hand, they may have a desire to look for images related to one of the previous results. Query-by-example is designed for this use case in which users are free to choose any keyframes to search for it visually similar ones. This function is also helpful when it comes to the ad-hoc search task that requires locating as many relevant videos to the given query as possible.

The last two features are bookmarking and viewing all keyframes. While users are searching for each task, they may bookmark relevant keyframes for future references when they are uncertain which keyframe is the most relevant. To view all keyframes, users can choose a keyframe to inspect other keyframes from the same video in a timely order.

Once the users find the most relevant keyframes representing the videos they are looking for, they can submit them at any time during the allocated time for that task. Lastly, when a single task is finished, users have the option to reset all the prompts and result lists via the clear button to be ready for the incoming task.

5 DISCUSSION

One of the major challenges our system, VideoCLIP, faced in VBS2023 was related to the network connection and storage. The VideoCLIP system is currently implemented and hosted in our remote server machine, which requires a stable network connection to access and runs the system in real time. We encountered unpredictably slow response and display times due to the enormous size of the dataset and the high demand for the network response. This unfortunately had a negative impact on VideoCLIP's performance. To overcome this problem, we propose to move the architecture to our local machine and use the network for submission only which will significantly reduce the burden on the network and improve the

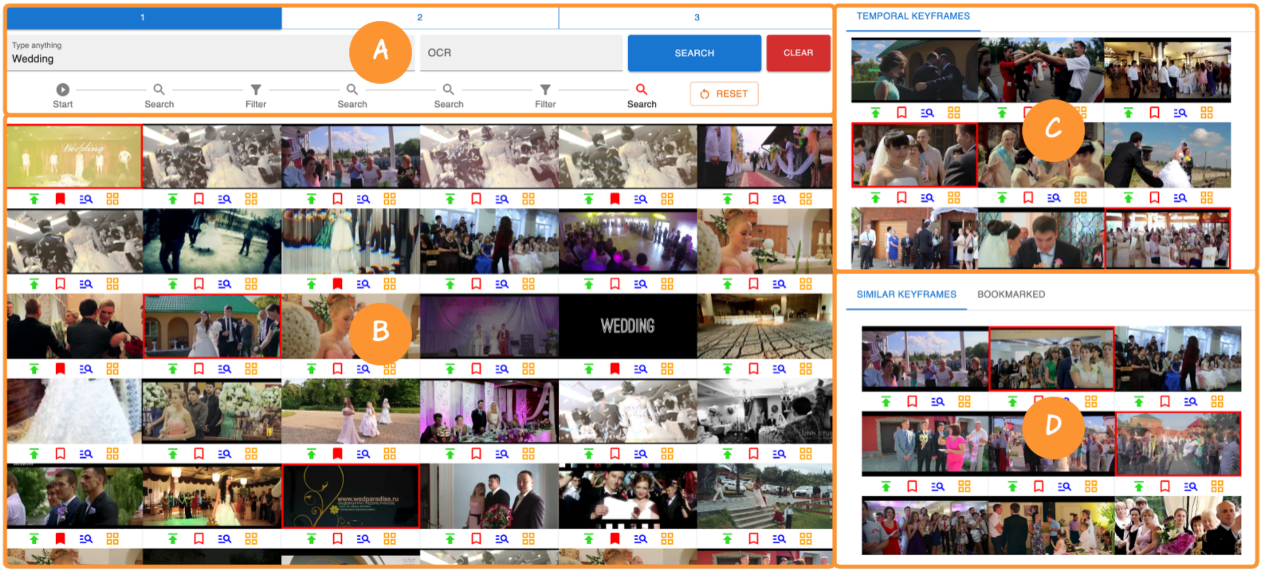

Figure 2: User Interface Prototype of VideoCLIP in IVR4B

overall performance of the system. Besides, an external hard disk will be used to store all datasets and the processed data for faster image visualisation.

6 CONCLUSION

In conclusion, the improvements made to our retrieval system, VideoCLIP, for the video retrieval benchmark for beginners in IVR4B workshop are significant. By incorporating the state-of-the-art embedding model, we have enhanced the video representations as keyframes, which will enable more efficient retrieval of relevant videos given the descriptions. In addition, we have offered three search options including free-text, search with example, and temporal search allowing users to vary the way to look for the targets. Furthermore, the UI has also been improved with better visualisation and additional functionalities in an effort to make the user experience not only more convenient but also more efficient in the process of video retrieval. We believe that these modifications will enable our system to achieve highly competitive results in this year's competition.

ACKNOWLEDGMENTS

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 18/CRT/6223, and 13/RC/2106_P2 at the ADAPT SFI Research Centre at DCU. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

REFERENCES

[1] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2022. VISIONE at Video Browser Showdown 2022. In *MultiMedia Modeling (Lecture Notes in Computer Science)*, Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Binh Huynh Thi Thanh, and Benoit Huet (Eds.). Springer International Publishing, Cham, 543–548. https://doi.org/10.1007/978-3-030-98355-0_52

[2] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2023. VISIONE at Video Browser Showdown 2023. In *MultiMedia Modeling (Lecture Notes in Computer Science)*, Duc-Tien Dang-Nguyen, Cathal Gurrin, Martha Larson, Alan F. Smeaton, Stevan Rudinac, Minh-Son Dao, Christoph Trattner, and Phoebe Chen (Eds.). Springer International Publishing, Cham, 615–621. https://doi.org/10.1007/978-3-031-27077-2_48

[3] Fabian Berns, Luca Rossetto, Klaus Schoeffmann, Christian Beecks, and George Awad. 2019. V3C1 Dataset: An Evaluation of Content Characteristics. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (Ottawa ON, Canada) *(ICMR '19)*. Association for Computing Machinery, New York, NY, USA, 334–338. https://doi.org/10.1145/3323873.3325051

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/ARXIV.2010.11929

[5] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *Proceedings of the 2021 International Conference on Multimedia Retrieval* (Taipei, Taiwan) *(ICMR '21)*. Association for Computing Machinery, New York, NY, USA, 690–691. https://doi.org/10.1145/3460426.3470945

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

[7] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoč, Andreas Leibetseder, František Mejzlík, Ladislav Peška, Luca Rossetto, Konstantin Schall, Klaus Schoeffmann, Heiko Schuldt, Florian Spiess, Ly-Duyen Tran, Lucia Vadicamo, Patrik Veselý, Stefanos Vrochidis, and Jiaxin Wu. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *International Journal of Multimedia Information Retrieval* 11 (2022), 1 – 18.

[8] Nico Hezel, Konstantin Schall, Klaus Jung, and Kai Uwe Barthel. 2022. Efficient Search and Browsing of Large-Scale Video Collections with Vibro. In *MultiMedia Modeling (Lecture Notes in Computer Science)*, Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Binh Huynh

Thi Thanh, and Benoit Huet (Eds.). Springer International Publishing, Cham, 487–492. https://doi.org/10.1007/978-3-030-98355-0_43

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[10] Jakub Lokoč, Patrik Veselý, František Mejzlík, Gregor Kovalčík, Tomáš Souček, Luca Rossetto, Klaus Schoeffmann, Werner Bailer, Cathal Gurrin, Loris Sauter, Jaeyub Song, Stefanos Vrochidis, Jiaxin Wu, and Björn þóR Jónsson. 2021. Is the Reign of Interactive Search Eternal? Findings from the Video Browser Showdown 2020. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 3, Article 91 (jul 2021), 26 pages. https://doi.org/10.1145/3445031

[11] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Sinéad Smyth, Annalina Caputo, and Cathal Gurrin. 2023. E-LifeSeeker: An Interactive Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual on Lifelog Search Challenge* (Thessaloniki, Greece) *(LSC'23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3592573.3593098

[12] Thao-Nhu Nguyen, Bunyarit Puangthamawathanakun, Annalina Caputo, Graham Healy, Binh T. Nguyen, Chonlameth Arpnikanondt, and Cathal Gurrin. 2023. VideoCLIP: An Interactive CLIP-Based Video Retrieval System At&nbsp;VBS2023. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I* (Bergen, Norway). Springer-Verlag, Berlin, Heidelberg, 671–677. https://doi.org/10.1007/978-3-031-27077-2_57

[13] Thao-Nhu Nguyen, Bunyarit Puangthamawathanakun, Graham Healy, Binh T. Nguyen, Cathal Gurrin, and Annalina Caputo. 2022. Videofall - A Hierarchical Search Engine for VBS2022. In *MultiMedia Modeling*, Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Binh Huynh Thi Thanh, and Benoit Huet (Eds.). Springer International Publishing, Cham, 518–523.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[15] Luca Rossetto, Klaus Schoeffmann, and Abraham Bernstein. 2021. Insights on the V3C2 Dataset. *CoRR* abs/2105.01475 (2021). arXiv:2105.01475 https://arxiv.org/abs/2105.01475

[16] Konstantin Schall, Nico Hezel, Klaus Jung, and Kai Uwe Barthel. 2023. Vibro: Video Browsing with Semantic and Visual Image Embeddings. In *MultiMedia Modeling*, Duc-Tien Dang-Nguyen, Cathal Gurrin, Martha Larson, Alan F. Smeaton, Stevan Rudinac, Minh-Son Dao, Christoph Trattner, and Phoebe Chen (Eds.). Springer International Publishing, Cham, 665–670.