Multimodal Retrieval in E-commerce
from Categories to Images, Text, and Back

Mariya Hendriksen[1]

Informatics Institute, University of Amsterdam m.hendriksen@uva.nl

Abstract. E-commerce provides rich multimodal data that is barely leveraged in practice. The majority of e-commerce search mechanisms are uni-modal, which are cumbersome and often fail to grasp the customer's needs. For the Ph.D. we conduct research aimed at combining information across multiple modalities to improve search and recommendations in e-commerce. The research plans are organized along the two principal lines. First, motivated by the mismatch between a textual and a visual representation of a given product category, we propose the task of category-to-image retrieval, i.e., the problem of retrieval of an image of a category expressed as a textual query. Besides, we propose a model for the task. The model leverages information from multiple modalities to create product representations. We explore how adding information from multiple modalities impacts the model's performance and compare our approach with state-of-the-art models. Second, we consider fine-grained text-image retrieval in e-commerce. We start off by considering the task in the context of reproducibility. Moreover, we address the problem of attribute granularity in e-commerce. We select two state-of-the-art (SOTA) models with distinct architectures, a CNN-RNN model and a Transformer-based model, and consider their performance on various e-commerce categories as well as on object-centric data from general domain. Next, based on the lessons learned from the reproducibility study, we propose the model for the fine-grained text-image retrieval.

1 Motivation

Multimodal retrieval is an important but understudied problem in e-commerce [50]. Even though e-commerce products are associated with rich multi-modal information, research currently focuses mainly on textual and behavioral signals to support product search and recommendation [1, 16, 44]. The majority of prior work in multimodal retrieval for e-commerce focuses on applications in the fashion domain, such as recommendation of fashion items [15, 36] and cross-modal fashion retrieval [13, 27]. In the more general e-commerce domain, multimodal retrieval has not been explored that well yet [19, 33]. Motivated by the knowledge gap, we lay out two directions for the research agenda: category-to-image retrieval, and fine-grained text-image retrieval.

Category-to-image retrieval. First, we focus on the category information in e-commerce. Product category trees are a key component of modern e-commerce as

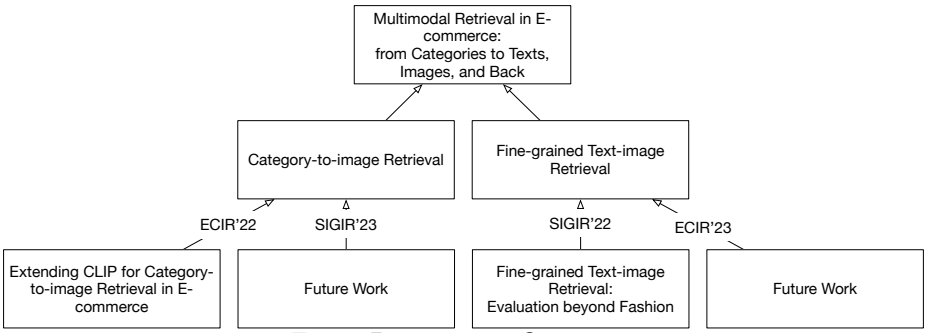

Fig. 1: Dissertation Overview.

they assist customers when navigating across large product catalogues [17, 26, 48, 52]. Yet, the ability to retrieve an image for a given product category remains a challenging task mainly due to noisy category and product data, and the size and dynamic character of product catalogues [30, 50]. Motivated by this challenge, we introduce the task of retrieving a ranked list of relevant images of products that belong to a given category, which we call the *category-to-image* (CtI) retrieval task. Unlike image classification tasks that operate on a predefined set of classes, in the CtI retrieval task we want to be able not only to understand which images belong to a given category but also to generalize towards unseen categories. Use cases that motivate the CtI retrieval task include (1) the need to showcase different categories in search and recommendation results [26, 48, 50]; (2) the task can be used to infer product categories in the cases when product categorical data is unavailable, noisy, or incomplete [54]; and (3) the design of cross-categorical promotions and product category landing pages [41].

Fine-grained text-image retrieval. Second, we address the problem of fine-grained text-image retrieval. Text-image retrieval is the task of finding similar items across textual and visual modalities. Successful performance on the task depends on the domain. In the general domain, where images typically depict complex scenes of objects in their natural contexts information across modalities is matched coarsely. Some examples of such datasets include MS COCO [35], and Flick30k [55]. By contrast, in the e-commerce domain, where there is typically one object per image, fine-grained matching is more important. Therefore, we focus on *fine-grained text-image retrieval*. We define the task as a combination of two subtasks: 1. *text-to-image retrieval*: given a noun phrase that describes an object, retrieve the image that depicts to the object; 2. *image-to-text retrieval*: given an image of an object, retrieve the noun phrase that describes an object.

We start off by examining the topic in the context of reproducibility. Reproducibility is one of the major pillars of the scientific method and is of utmost importance for Information Retrieval (IR) as a discipline rooted in experimentation [10]. One of the first works that touch upon reproducibility in IR is the study by Armstrong et al. [2] where the authors conducted a longitudinal analysis of papers published in proceedings of CIKM and SIGIR between 1998-2008 and discovered that the ad-hoc retrieval was not measurably improving. Later

on, Yang et al. [53] provided a meta-analysis of results reported on the TREC Robust04 and found out that some of the more recent neural models were outperformed by strong baselines. Similar discoveries were made in the domain of recommender systems research [5, 6]. Motivated by the findings, we explore the reproducibility of fine-grained text-image retrieval results. More specifically, we examine how SOTA models for fine-grained text-image fashion retrieval generalize towards other categories of e-commerce products. After analyzing SOTA models in the domain, we plan to improve upon them in a subsequent future work.

2 Related Work

Category-to-image retrieval. Early work in image retrieval grouped images into a restricted set of semantic categories and allowed users to retrieve images by using category labels as queries [46]. Later work allowed for a wider variety of queries ranging from natural language [22, 51], to attributes [39], to combinations of multiple modalities (e.g., title, description, and tags) [49]. Across these multimodal image retrieval approaches we find three common components: (1) an image encoder, (2) a query encoder, and (3) a similarity function to match the query to images [14, 42]. Depending on the focus of the work some components might be pre-trained, whereas the others are optimized for a specific task. In our work, we rely on pre-trained image and text encoders but learn a new multimodal composite of the query to perform CtI retrieval.

Fine-grained text-image retrieval. Early approaches to cross-modal mapping focused on correlation maximization through canonical correlation analysis [20, 21, 47]. Later approaches centered around convolutional and recurrent neural networks [11, 24, 25, 31]. They were further expanded by adding attention on top of encoders [31, 37, 40]. More recently, inspired by the success of transformers [8], a line of work centered around creating a universal vision-language encoder emerged [4, 32, 34, 38]. To address the problem of attribute granularity in the context of cross-modal retrieval, a line of work proposed to segment images into fragments [29], use attention mechanisms [28], combine image features across multiple levels [13], use pre-trained BERT as a backbone [12, 56]. Unlike prior work in this domain that focused on fashion, we focus on the general e-commerce domain.

3 Research Description & Methodology

The dissertation comprises two parts. Below, we describe every part of the thesis and elaborate on the methodology.

Category-to-image retrieval. Product categories are used in various contexts in e-commerce. However, in practice, during a user's session, there is often a mismatch between a textual and a visual representation of a given category. Motivated by the problem, we introduce the task of category-to-image retrieval

in e-commerce and propose a model for the task.

We use the XMarket dataset recently introduced by Bonab et al. [3] that contains textual, visual, and attribute information of e-commerce products as well as a category tree. Following [7, 23, 45] we use BM25, MPNet, CLIP as our baselines. To evaluate model performance, we use Precision@K where $K = \{1, 5, 10\}$, mAP@K where $K = \{5, 10\}$, and R-precision.

RQ1.1 *How do baseline models perform on the CtI retrieval task? Specifically, how do unimodal and bi-modal baseline models perform? How does the performance differ w.r.t. category granularity?*

To answer the question, we feed BM25 corpora that contain textual product information, i.e., product titles. We use an MPNet in a zero-shot manner. For all the products in the dataset, we pass the product title through the model. During the evaluation, we pass a category expressed as textual query through MPNet and retrieve top-k candidates ranked by cosine similarity w.r.t. the target category. We compare categories of the top-k retrieved candidates with the target category. Besides, we use pre-trained CLIP in a zero-shot manner with a text transformer and a vision transformer (ViT) [9] configuration. We pass the product image through the image encoder. For evaluation, we pass a category through the text encoder and retrieve top-k image candidates ranked by cosine similarity w.r.t. the target category. We compare categories of the top-k retrieved image candidates with the target category.

RQ1.2 *How does a model, named CLIP-I, that uses product image information for building product representations impact the performance on the CtI retrieval task?*

To answer the question, we build product representations by training on e-commerce data. We investigate how using product image data for building product representations impacts performance on the CtI retrieval task. To introduce visual information, we extend CLIP in two ways: (1) We use ViT from CLIP as an image encoder. We add a product projection head that takes as an input product visual information. (2) We use the text encoder from MPNet as category encoder; we add a category projection head on top of the category encoder. We name the resulting model CLIP-I. We train CLIP-I on category-product pairs from the training set. We only use visual information for building product representations.

RQ1.3 *How does CLIP-IA, which extends CLIP-I with product attribute information, perform on the CtI retrieval task?*

To answer the question, we extend CLIP-I by introducing attribute information to the product information encoding pipeline. We add an attribute encoder through which we obtain a representation of product attributes. We concatenate the resulting attribute representation with image representation and pass the resulting vector to the product projection head. Thus, the resulting product representation $\mathbf{p}$ is based on both visual and attribute product information. We name the resulting model CLIP-IA. We train CLIP-IA on category-product pairs and we use visual and attribute information for building product representation.

RQ1.4 *And finally, how does CLIP-ITA, which extends CLIP-IA with prod-*

uct text information, perform on the CtI task?

To answer the question, we investigate how extending the product information processing pipeline with the textual modality impacts performance on the CtI retrieval task. We add a title encoder to the product information processing pipeline and use it to obtain title representation. We concatenate the resulting representation with product image and attribute representations. We pass the resulting vector to the product projection head. The resulting model is CLIP-ITA. We train and test CLIP-ITA on category-product pairs. We use visual, attribute, and textual information for building product representations. The results are to be published in ECIR'22 [17]. The follow-up work is planned to be published at SIGIR 2023.

Fine-grained text-image retrieval. The ongoing work is focused on fine-grained text-image retrieval in the context of reproducibility. For the experiments, we select two SOTA models for fine-grained cross-modal fashion retrieval, each model with distinctive architecture. One of them is based on Transformer while another one is CNN-RNN-based. The Transformer-based model is Kaleido-BERT [56], that extends BERT [8]. Another model is a Multi-level Feature approach (MLF) [13]. Both models claim to deliver SOTA performance by being able to learn image representations that can better represent fine-grained attributes. They were evaluated on Fashion-Gen dataset [43] but, to the best of our knowledge, were not compared against each other.

In the work, we aim to answer the following research questions:

RQ2.1 *How well Kaleido-BERT and MLF perform on data from an e-commerce category that is different from Fashion?*

RQ2.2 *How well both models generalize beyond e-commerce domain? More specifically, how do they perform on object-centric data from the general domain?*

RQ2.3 *How Kaleido-BERT and MLF compare to each other w.r.t performance?*

The results are planned to be published as a paper at SIGIR 2022. The follow-up work is planned to be published at ECIR 2023 [18].

4 Appendix

Student's statement

As my work focuses on multimodal retrieval in e-commerce, I see ECIR as one of the most suitable venues to present my work. I submit my proposal for the doctoral consortium because I would like to discuss my research agenda with senior researches in the field. I hope to receive feedback on my ideas and get inspiration for my future work. Last but not least, I am very much looking forward to discussing my research with the other Ph.D. students attending the consortium.

Advisor's statement

I am happy to write this statement of endorsement for Mariya Hendriksen. Mariya has about two years left until the end of her PhD. So far, she has worked on different aspects of user behavior and multi-modal retrieval in the context of e-commerce, resulting in papers at the eCom workshops in 2020 and 2021, and in a full paper submission to ECIR 2022.

Despite being forced to work remotely for most of her PhD so far (because of the pandemic), Mariya has been a very constructive force in the team, both as a colleague who helps out to onboard new colleagues when they first arrive and in a more formal role, looking after reporting for industrial collaboration that funds her position. She has also broadened the perspective of the research team by bringing in expertise and challenges related to multimodal retrieval, and she has teamed up with a number of fellow PhD students in the lab to study the use of large transformer-based models in the context of cross-modal retrieval, combining and demonstrating her scholarly and leadership skills.

Mariya has written a four part PhD thesis proposal, partly based on research already performed, and partly outlining research that she still wants to perform. The doctoral consortium comes at exactly the right time for Mariya as it will allow her to get feedback on her proposed plans, with enough time to adjust based on feedback from the doctoral consortium. In particular, feedback on evaluation and the development of simulation-based training and testing environments would be particularly timely and helpful.

Maarten de Rijke, distinguished university professor, University of Amsterdam

Bibliography

[1] Ariannezhad M, Jullien S, Nauts P, Fang M, Schelter S, de Rijke M (2021) Understanding multi-channel customer behavior in retail. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp 2867–2871

[2] Armstrong TG, Moffat A, Webber W, Zobel J (2009) Improvements that don't add up: Ad-hoc retrieval results since 1998. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, Association for Computing Machinery, pp 601–610

[3] Bonab H, Aliannejadi M, Vardasbi A, Kanoulas E, Allan J (2021) Cross-market product recommendation. In: CIKM, ACM

[4] Chen YC, Li L, Yu L, Kholy AE, Ahmed F, Gan Z, Cheng Y, Liu J (2019) Uniter: Learning universal image-text representations. arXiv preprint arXiv:190911740

[5] Dacrema MF, Cremonesi P, Jannach D (2019) Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp 101–109

[6] Dacrema MF, Boglio S, Cremonesi P, Jannach D (2021) A troubling analysis of reproducibility and progress in recommender systems research. ACM Transactions on Information Systems (TOIS) 39(2):1–49

[7] Dai Z, Lai G, Yang Y, Le QV (2020) Funnel-transformer: Filtering out sequential redundancy for efficient language processing. arXiv preprint arXiv:200603236

[8] Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

[9] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929

[10] Ferro N, Fuhr N, Järvelin K, Kando N, Lippold M, Zobel J (2016) Increasing reproducibility in ir: Findings from the dagstuhl seminar on" reproducibility of data-oriented experiments in e-science". In: ACM SIGIR Forum, ACM New York, NY, USA, vol 50, pp 68–82

[11] Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato MA, Mikolov T (2013) Devise: A deep visual-semantic embedding model. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems 26, Curran Associates, Inc., pp 2121–2129

[12] Gao D, Jin L, Chen B, Qiu M, Li P, Wei Y, Hu Y, Wang H (2020) Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval.

In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 2251–2260
[13] Goei K, Hendriksen M, de Rijke M (2021) Tackling attribute fine-grainedness in cross-modal fashion search with multi-level features. In: SIGIR 2021 Workshop on eCommerce, ACM
[14] Gupta T, Vahdat A, Chechik G, Yang X, Kautz J, Hoiem D (2020) Contrastive learning for weakly supervised phrase grounding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, pp 752–768
[15] Hendriksen M, Overes V (2022) Unimodal vs. multimodal siamese networks for outfit completion. arXiv preprint arXiv:220710355
[16] Hendriksen M, Kuiper E, Nauts P, Schelter S, de Rijke M (2020) Analyzing and predicting purchase intent in e-commerce: Anonymous vs. identified customers. arXiv preprint arXiv:201208777
[17] Hendriksen M, Bleeker M, Vakulenko S, van Noord N, Kuiper E, de Rijke M (2022) Extending CLIP for category-to-image retrieval in e-commerce. In: ECIR 2022: 44th European Conference on Information Retrieval, Springer
[18] Hendriksen M, Vakulenko S, Kuiper E, de Rijke M (2023) Scene-centric vs. object-centric image-text cross-modal retrieval: A reproducibility study. arXiv preprint arXiv:230105174
[19] Hewawalpita S, Perera I (2019) Multimodal user interaction framework for e-commerce. In: 2019 International Research Conference on Smart Computing and Systems Engineering (SCSE), IEEE, pp 9–16
[20] Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research 47:853–899
[21] Hotelling H (1992) Relations between two sets of variates. In: Breakthroughs in statistics, Springer, pp 162–190
[22] Hu R, Xu H, Rohrbach M, Feng J, Saenko K, Darrell T (2016) Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4555–4564
[23] Jabeur LB, Soulier L, Tamine L, Mousset P (2016) A product feature-based user-centric ranking model for e-commerce search. In: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, pp 174–186
[24] Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137
[25] Kiros R, Salakhutdinov R, Zemel RS (2014) Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:14112539
[26] Kondylidis N, Zou J, Kanoulas E (2021) Category aware explainable conversational recommendation. arXiv preprint arXiv:210308733
[27] Laenen K, Moens MF (2019) Multimodal neural machine translation of fashion e-commerce descriptions. In: International Conference on Fashion communication: between tradition and future digital developments, Springer, pp 46–57

[28] Laenen K, Moens MF (2020) A comparative study of outfit recommendation methods with a focus on attention-based fusion. Information Processing & Management 57(6):102316
[29] Laenen K, Zoghbi S, Moens MF (2017) Cross-modal search for fashion attributes. In: Proceedings of the KDD 2017 Workshop on Machine Learning Meets Fashion, ACM, vol 2017, pp 1–10
[30] Laenen K, Zoghbi S, Moens MF (2018) Web search of fashion items with multimodal querying. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp 342–350
[31] Lee KH, Chen X, Hua G, Hu H, He X (2018) Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 201–216
[32] Li G, Duan N, Fang Y, Jiang D, Zhou M (2019) Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. arXiv preprint arXiv:190806066
[33] Li H, Yuan P, Xu S, Wu Y, He X, Zhou B (2020) Aspect-aware multimodal summarization for chinese e-commerce products. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 8188–8195
[34] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW (2019) Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:190803557
[35] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755
[36] Lin Y, Ren P, Chen Z, Ren Z, Ma J, de Rijke M (2019) Improving outfit recommendation with co-supervision of fashion generation. In: The World Wide Web Conference, pp 1095–1105
[37] Liu C, Mao Z, Liu AA, Zhang T, Wang B, Zhang Y (2019) Focus your attention: A bidirectional focal attention network for image-text matching. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 3–11
[38] Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems, pp 13–23
[39] Nagarajan T, Grauman K (2018) Attributes as operators: factorizing unseen attribute-object compositions. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 169–185
[40] Nam H, Ha JW, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 299–307
[41] Nielsen J, Molich R, Snyder C, Farrell S (2000) E-commerce user experience. Nielsen Norman Group
[42] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from natural language supervision. arXiv preprint arXiv:210300020
[43] Rostamzadeh N, Hosseini S, Boquet T, Stokowiec W, Zhang Y, Jauvin C,

Pal C (2018) Fashion-gen: The generative fashion dataset and challenge. arXiv preprint arXiv:180608317

[44] Rowley J (2000) Product search in e-shopping: a review and research propositions. Journal of consumer marketing

[45] Shen S, Li LH, Tan H, Bansal M, Rohrbach A, Chang KW, Yao Z, Keutzer K (2021) How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:210706383

[46] Smeulders A, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12):1349–1380

[47] Socher R, Fei-Fei L (2010) Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, pp 966–973

[48] Tagliabue J, Yu B, Beaulieu M (2020) How to grow a (product) tree: personalized category suggestions for ecommerce type-ahead. arXiv preprint arXiv:200512781

[49] Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ (2016) Yfcc100m: The new data in multimedia research. Communications of the ACM 59(2):64–73

[50] Tsagkias M, King TH, Kallumadi S, Murdock V, de Rijke M (2020) Challenges and research opportunities in ecommerce search and recommendations. SIGIR Forum 54(1)

[51] Vo N, Jiang L, Sun C, Murphy K, Li LJ, Fei-Fei L, Hays J (2019) Composing text and image for image retrieval-an empirical odyssey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6439–6448

[52] Wirojwatanakul P, Wangperawong A (2019) Multi-label product categorization using multi-modal fusion models. arXiv preprint arXiv:190700420

[53] Yang W, Lu K, Yang P, Lin J (2019) Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, pp 1129–1132

[54] Yashima T, Okazaki N, Inui K, Yamaguchi K, Okatani T (2016) Learning to describe e-commerce images from noisy online data. In: Asian Conference on Computer Vision, Springer, pp 85–100

[55] Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics 2:67–78

[56] Zhuge M, Gao D, Fan DP, Jin L, Chen B, Zhou H, Qiu M, Shao L (2021) Kaleido-bert: Vision-language pre-training on fashion domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12647–12657