ProVLA: Compositional Image Search with Progressive Vision-Language Alignment and Multimodal Fusion

Zhizhang Hu*
University of California, Merced
zhu42@ucmerced.edu

Xinliang Zhu
Amazon Visual Search & AR
xlzhu@amazon.com

Son Tran
Amazon M5
sontran@amazon.com

René Vidal
Amazon Visual Search & AR
rvidal@amazon.com

Arnab Dhua
Amazon Visual Search & AR
aduha@amazon.com

Abstract

Traditional image-to-image and text-to-image search struggle with comprehending complex user intentions, particularly in fashion e-commerce, where users search for similar products with text modifications to a reference image. This paper introduces Progressive Vision-Language Alignment and Multimodal Fusion (ProVLA), a novel approach which utilizes a transformer-based vision and language model to generate multimodal embeddings. Our method involves a two-step learning process and a cross-attention-based fusion encoder to facilitate robust information fusion, and a momentum queue-based hard negative mining mechanism to handle noisy training data. Extensive evaluations on the Fashion 200k and Shoes benchmark datasets demonstrate that our model outperforms state-of-the-art methods.

1. Introduction

Image search, a cornerstone of computer vision, encodes the semantics of an input query to retrieve the most relevant images [9, 24, 31, 12, 10]. However, these traditional image-to-image and text-to-image modalities often struggle to accurately understand complex user intentions. Particularly in the fashion e-commerce domain, where users intend to search for a similar product given a reference picture, but they also have certain changes over the reference, such as changing the color and style. In such cases, the traditional image-to-image and text-to-image methods are limited in fulfilling the users' search requirements.

Compositional image search [29, 1, 3] aims to address these challenges by using both both image and text inputs. Figure 1 depicts the process of compositional image search. The image serves as a reference, and the text is used as a modification guidance to describe how the reference image should be changed to approximate users' intentions and obtain the target image. This approach enables users to express their ideas more precisely by leveraging both visual and textual information. However, one intrinsic challenge is the difficulty of compositing image-text representations that transform only the image features relevant to the text modification while preserving other features. Also, given the labor-intensive data collection and labeling processes, the training data is prone to be noisy, i.e., wrongly labeled color, feature, products, etc. Previous works [29, 1, 3] have made strides in this direction but still face several shortcomings.

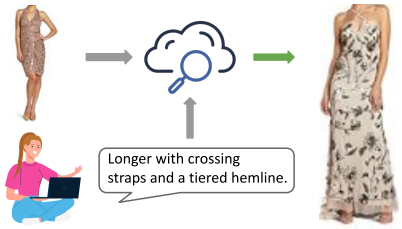

Figure 1. Compositional image search: A new product is retrieved, resembling the reference image while incorporating changes specified in the text.

This paper proposes ProVLA, a new compositional image search model that addresses these challenges. It employs transformer-based vision [21] and language [7] backbone models to generate embedding tokens for each modality. We introduce a two-step learning procedure and a cross-attention-based fusion encoder module to solve the first challenge, and a momentum queue-based hard negative mining mechanism to tackle the second challenge. We evaluate ProVLA on the Fashion 200k [15] and Shoes [14] benchmarking datasets. Quantitative results indicate our ProVLA outperforms existing state-of-the-art (SOTA) methods, and qualitative visualization demonstrates effective interaction and exchange of information between modalities.

*This work is done during the internship at Amazon.

In summary, our main contributions are the following:

- The introduction of ProVLA, a multimodal compositional learning model for the combined representation of image and text queries.
- A two-step learning framework to progressively condition the model in the complex knowledge for downstream tasks.
- The development of a cross-attention-based multimodal fusion encoder and a momentum queue-based hard negative mining mechanism for robust and efficient information fusion.
- The validation of ProVLA on multiple image-text retrieval benchmarking datasets, demonstrating superior performance over recent state-of-the-art methods.

2. Related Works

Image Retrieval. Image retrieval, a longstanding problem in computer vision, relies on learning an embedding space that maps query and target retrieval candidates closely or far apart depending on relevance [5]. Traditional unimodal-based queries are effective for simple searches but struggle with more complex retrieval targets [12, 24, 22, 20, 32, 36, 11, 34, 35]. In response, Vo et al. proposed a compositional image retrieval model that combines image and text queries [29]. Subsequent models, like TIRG [29], ARTEMIS [6], VAL [3], DATIR [13], CoSMo [18], and MAAF [8], have further developed this approach. However, they tend to neglect contextualizing backbone models pre-trained on the open-domain data, which may limit the model to align domain-specific text terms with image contents.

Vision-language Pre-training. Inspired by the success of transformer-based pre-training in natural language processing [7], vision-language pre-training leverages large-scale image-text pairs to learn a joint vision-language embedding space. Models like CLIP [26] and ALIGN [16] have shown promising results, yet they often ignore the interaction between modalities. Other studies such as UNITER [4], ViLT [17], ALBEF [19], and TCL [33] have proposed ways to learn joint embeddings of image content and natural language during pre-training, thereby increasing the interaction between modalities. This work builds upon these findings, proposing a vision-language training approach that tailors the backbone models to the joint embedding space with the semantics of the target retrieval task.

3. Proposed Method

The proposed framework, referred to as ProVLA, aims to create a fused representation of the reference image and modification text that is well aligned with the target image. It comprises a two-step progressive learning: Pre-fine-tuning and Image-Text Composition, and two main components: multimodal fusion encoder and momentum queue-based negative mining (Figure 2). The pre-fine-tuning step conditions the model into the domain-specific semantic spaces [37, 3]. During this phase, the model learns to align domain-specific text terms with image contents. Subsequently, the Image-Text Composition process executes multimodal fusion for the image search and retrieval task. The Pre-fine-tuning and Image-text composition networks utilize a vision encoder, a text encoder, and a multimodal fusion encoder, all based on transformer models [28]. The Pre-fine-tuning network inputs are the image and its caption, while the Image-text Composition network takes the reference image, target image, and corresponding modification text as inputs. In both steps, a Swin Transformer [21] is employed to extract visual representation from images. The compositional image search task requires both fine-grained visual information for the features mentioned in the modification text and coarse-grained visual information for the remaining features that should be preserved. As the Swin Transformer learns visual concepts at different levels of granularity in a hierarchical order, we concatenate output image tokens from the penultimate (Stage 3) layers and the final (Stage 4) as the image embedding. To encode the semantics of the modification text, we adopt a 6-layer transformer as the language backbone.

We view the image-text composition task as an image-to-image retrieval task, conditioned by the explicit description in the modification text. This process mainly relies on the cross-attention mechanism in the multimodal fusion encoder. The reference image embedding serves as the query in the cross-attention operation, acting as the initial search context for visual information. The modification text embedding, functioning as both keys and values, is the potential transformation repository, from which relevant modifications are identified and applied to the image, effectively bridging the image and text domains.

The primary objective of image-text composition training is to bring the embeddings of matched composed image-target image pairs closer while distancing the unmatched pairs. Therefore we employ a momentum queue-based negative mining strategy. For this, we leverage contrastive learning and a mechanism of momentum-based distillation, maintaining queues to store the most recent composed image, reference image, and target image embeddings [30, 19]. This results in optimized contrastive learning tasks and the calculation of softmax-normalized similarity.

More formally, let I^r denote the reference image, T the modifying text, C the composed image, and I^t the target image. Let $Y^{c2I^t}(C)$ and $P^{c2I^t}(C)$ denote, respectively, the one hot and soft-max normalized similarities between the composed image and the set of target images, as defined in [19]. Likewise, let $Y^{I^tc2}(I^t)$ and $P^{I^tc2}(I^t)$ denote the one-hot and soft-max normalized similarities between the target image and the set of composed images. Here, (I^t, C)

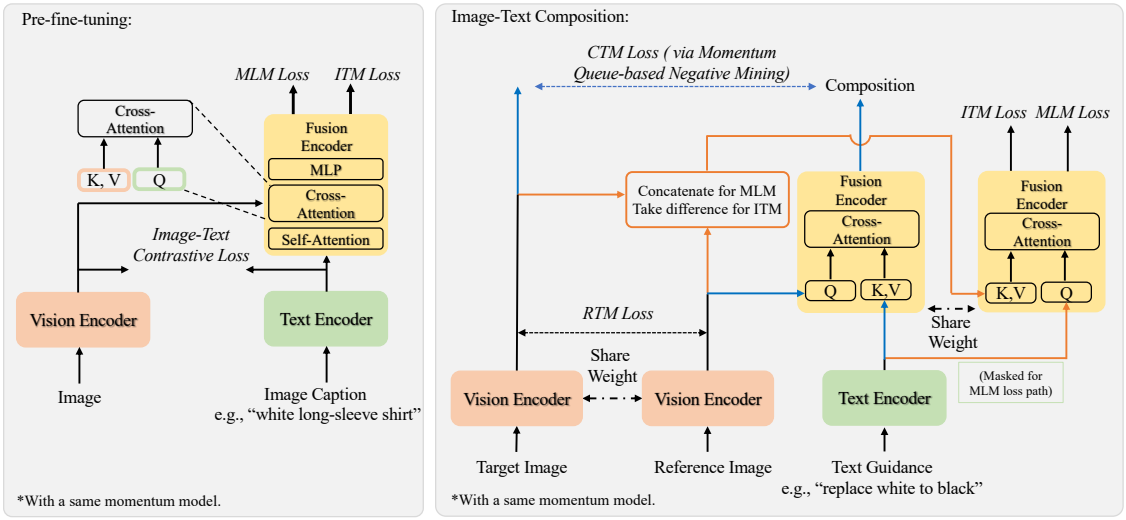

Figure 2. Overview of the proposed ProVLA architecture. It consists of two steps: Pre-fine-tuning and Image-text Composition. In the Image-text Composition step, two main forward paths are highlighted with blue and orange colors. The blue-colored path is the multimodal fusion path and the orange colored is the auxiliary loss path. The centered multimodal fusion path is also the inference path. Please be notified the Fusion Encoder plot in the Image-text Composition step is a simplified plot.

is a positive pair if the tuple (I^r, T) used to compute C is paired with I^t, in which case the one-hot similarity has a probability of 1. Otherwise, the pair is negative and the one-hot similarity has a probability of 0. The Composed-Target matching (CTM) loss is then mathematically defined as the cross-entropy $\mathbb{H}$ between softmax-normalized similarity P and the one-hot similarity label Y [19], i.e.:

$$\mathcal{L}_{CTM} = \frac{1}{2}\mathbb{E}_{(C,I^t)}[\mathbb{H}(Y^{c2I^t}(C), P^{c2I^t}(C)) + \mathbb{H}(Y^{I^t2c}(I^t), P^{I^t2c}(I^t))]. \quad (1)$$

In order to encourage the vision encoder to preserve visual attributes not modified by the text, we adopt the Reference-Target matching (RTM) loss. This loss employs the same design as the CTM loss, but the pair of inputs is the embedding of reference and target images. Also, to facilitate the training of the fusion encoder, we adopt the image-text matching (ITM) and masked language modeling (MLM) losses, which are widely used in previous vision-language fusion research [19, 33]. Given the triplet of the reference image, target image, and text, the task of image-text matching is to predict whether the text matches the difference between the two images. The masked language modeling task aims to predict the masked word in the given text by looking at the remaining text and two images. We adopt the same fusion encoder for the image-text composition, but the text embedding is the query to the cross-attention. For the ITM, the difference between target and reference image embedding is the input to the fusion encoder, and the concatenation of two image embeddings is used as the input for the MLM task.

In summary, the overall training objective of the Image-Text Composition is:

$$\mathcal{L} = \mathcal{L}_{CTM} + \mathcal{L}_{RTM} + \mathcal{L}_{ITM} + \mathcal{L}_{MLM}. \quad (2)$$

Please note that we omit the loss from the momentum model [19] for a clearer presentation.

4. Experiment

4.1. Experimental Setup

Datasets. We test our method on two widely-used benchmarking datasets: Fashion 200k [15] and Shoes [14].

The Fashion 200k dataset, sourced from multiple websites, comprises roughly 200,000 fashion images with product information and product images. We adopt the approach introduced in TIRG [29], using only images and corresponding text captions, where the text style is an attribute list. We filter image pairs with a single word difference in the caption and generate modification text in a "replace A with B" format. This results in triplets of a reference image, target image, and modification text. The dataset comprises around 170,000 training and 30,000 testing triplets. The retrieval is performed by matching target attributes, not images.

The Shoes dataset was initially designed for attribute discovery [2], featuring image-text pairs of products and titles. It is then augmented with relative captions in natural language for dialog-based interactive image retrieval [14]. It consists of 10,000 training and 4,658 testing triplets. The modification text in this dataset is human-annotated, reflecting real user feedback in e-commerce.

Baseline Models. We compare ProVLA with a wide range of baselines including early works and recent SOTA models on this task, including TIRG [29], VAL [3], MAAF [8], CoSMo [18], ARTEMIS [6] and DATIR [13].

Evaluation Metric. For both datasets, we adopt the standard evaluation metric in retrieval, i.e., Recall@K, denoted as R@K for short.

Implementation. We use the PyTorch [25] deep learning framework to conduct all our experiments. The Swin Transformer is pre-trained on ImageNet-21k [27] at resolution 224x224. The text encoder and fusion encoder are initialized with a $\text{BERT}_{\text{base}}$ [7] model. We pre-fine-tune the model for 50 and 100 epochs for the Fashion 200K and Shoes datasets, respectively, using a batch size of 192 on 8 NVIDIA V100 GPUs. The optimizer is AdamW [23] with a weight decay rate of 0.02. We adopt a cosine-scheduled learning rate with a warmup set to $1e^{-5}$ in the first 30 epochs, and decayed to $1e^{-6}$. The momentum parameter is set to 0.995, and the size of the hard-negative mining queue is set to 38,400. The distillation weight α is linearly ramped up from 0 to 0.6 within the 1st epoch.

4.2. Compositional Search Performance

Table 1 and Table 2 illustrate the retrieval performance on the Fashion 200K and Shoes datasets. From those tables, we observe that our model achieves compelling results compared to other methods. On the Fashion 200k dataset, our model surpasses the SOTA method by a significant margin, with a 2.6 percent points increase in R@10 and a 1.9 percent points increase in R@50. However, we note a minor setback, as our model underperforms by 1.6 percent points on the R@1 metric. Similarly, in the Shoes dataset, our model exhibits a competitive edge, outperforming the SOTA by 0.5 percent points on R@1 and by a notable 3.1 percent points on R@50. A limitation surfaces on the R@50 metric, where our model falls behind the SOTA.

The improvement in both datasets can be attributed to our model's efficacy in leveraging relevant image and text data, generating meaningful modifications for improved retrieval performance. However, the slight deficits in Recall@1 on the Fashion 200k dataset and Recall@50 on the Shoes dataset underscore room for further refinement. We conjecture that the shortcomings may stem from the model's current handling of edge cases where minor attribute variations significantly impact the retrieval process.

4.3. Ablation Study

To better understand the impact of the two key components of our model, namely the two-step progressive learning and the queue-based hard negative mining, we conducted an ablation study. This analysis aims to isolate the contributions of each component and provide insights into their functionality and importance within the model. Table 3 depicts the result of the ablation study.

Initially, a simplified model without these components showed R@10 scores of 18.2 and 23.6 on the Fashion 200k and Shoes datasets, respectively. Adding queue-based hard negative mining significantly increased the scores to 30.3 and 53.7, demonstrating its vital role in improving retrieval precision by handling diverse negative samples. Incorporating both components further improved R@10 scores to 53.7 and 56.2 on the respective datasets, emphasizing the importance of progressive learning in refining the model's understanding of the data.

Method	Fashion 200k		
	R@1	R@10	R@50
TIRG	14.1	42.5	63.8
VAL	22.9	50.8	72.7
MAAF	18.9	-	-
CosMo	23.3	50.4	69.3
ARTEMIS	21.5	51.1	70.5
DATIR	21.5	48.8	71.6
Ours	21.7	53.7	74.6

Table 1. Quantitative results for the Fashion 200k dataset. The highest value is colored green. "-" means the result is not reported in the original paper.

Method	Shoes		
	R@1	R@10	R@50
TIRG	12.6	45.5	69.4
VAL	16.5	49.1	73.5
MAAF	16.4	50.0	76.4
CosMo	16.7	48.4	75.6
ARTEMIS	18.7	53.1	79.3
DATIR	17.2	51.1	75.6
Ours	19.2	56.2	73.3

Table 2. Quantitative results for the Shoes dataset. The highest value is colored green.

Method	Fashion 200k	Shoes
	R@10	R@10
No Pre-fine-tuning and no queue-based negative mining	18.2	23.6
No Pre-fine-tuning	30.3	37.4
ProVLA	53.7	56.2

Table 3. Ablation study of removing two key components in the ProVLA.

5. Conclusion

In conclusion, our Progressive Vision-language Alignment and Multimodal Fusion (ProVLA) model introduces a more effective method for image search, especially in contexts where users aim to find items similar to a reference image with specific modifications. The model employs a two-step learning framework, a cross-attention-based fusion encoder, and a momentum queue-based hard negative mining mechanism to tackle inherent challenges. Evaluated on multiple image-text retrieval benchmark datasets, ProVLA outperforms recent state-of-the-art methods, illustrating its potential for robust and precise image search.

References

[1] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1140–1149, 2021. 1

[2] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pages 663–676. Springer, 2010. 3

[3] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020. 1, 2, 3

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[5] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 2

[6] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*, 2022. 2, 3

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2, 4

[8] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. 2, 3

[9] Ming Du, Arnau Ramisa, Amit Kumar KC, Sampath Chanda, Mengjiao Wang, Neelakandan Rajesh, Shasha Li, Yingchuan Hu, Tao Zhou, Nagashri Lakshminarayana, et al. Amazon shop the look: A visual search system for fashion and home. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2822–2830, 2022. 1

[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 1

[11] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1171–1180, 2019. 2

[12] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016. 1, 2

[13] Chunbin Gu, Jiajun Bu, Zhen Zhang, Zhi Yu, Dongfang Ma, and Wei Wang. Image search with text feedback by deep hierarchical attention mutual information maximization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4600–4609, 2021. 2, 3

[14] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31, 2018. 1, 3

[15] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017. 1, 3

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[17] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[18] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021. 2, 3

[19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 3

[20] Wen Li, Lixin Duan, Dong Xu, and Ivor Wai-Hung Tsang. Text-based image retrieval using progressive multi-instance learning. In *2011 international conference on computer vision*, pages 2049–2055. IEEE, 2011. 2

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 2

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[24] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 1, 2

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 4

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[27] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 4

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[29] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 1, 2, 3

[30] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020. 2

[31] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2119–2128, 2018. 1

[32] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12):5412–5425, 2020. 2

[33] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 2, 3

[34] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018. 2

[35] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016. 2

[36] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3536–3545, 2020. 2

[37] Yida Zhao, Yuqing Song, and Qin Jin. Progressive learning for image retrieval with hybrid-modality queries. *arXiv preprint arXiv:2204.11212*, 2022. 2