Conditioned Image Retrieval for Fashion using Contrastive Learning and CLIP-based Features

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, Alberto Del Bimbo
[name.surname]@unifi.it
Università degli Studi di Firenze - MICC
Firenze, Italy

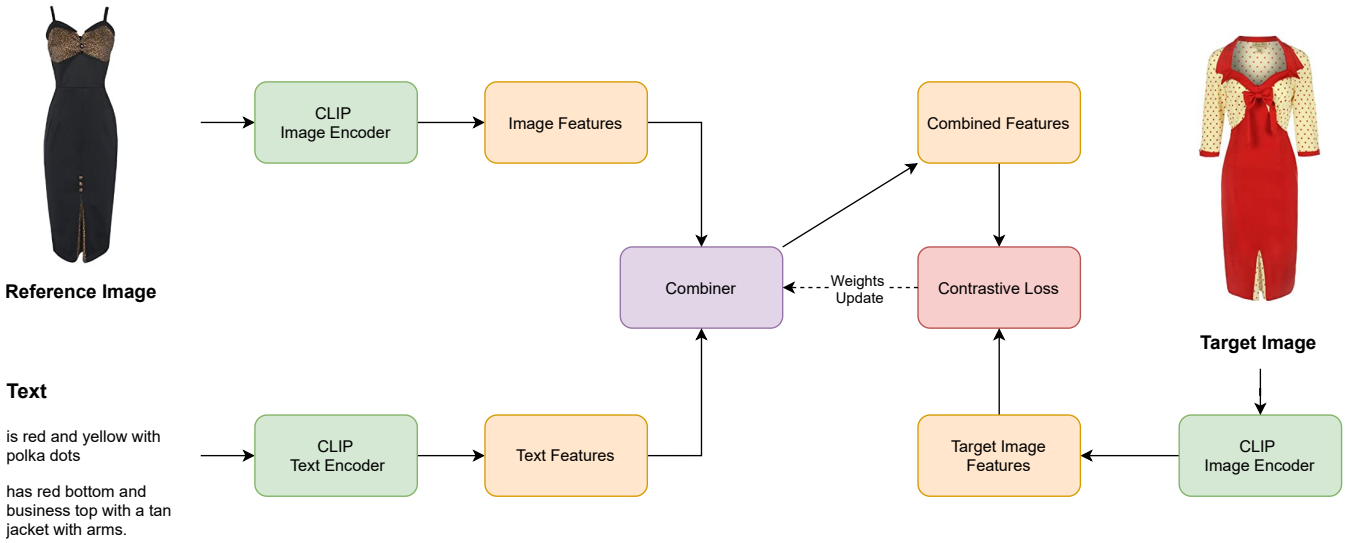

Figure 1: An overview of the system, from the input image and captions on the left, to the target image on the right.

ABSTRACT

Building on the recent advances in multimodal zero-shot representation learning, in this paper we explore the use of features obtained from the recent CLIP model to perform conditioned image retrieval. Starting from a reference image and an additive textual description of what the user wants with respect to the reference image, we learn a Combiner network that is able to understand the image content, integrate the textual description and provide combined feature used to perform the conditioned image retrieval. Starting from the bare CLIP features and a simple baseline, we show that a carefully crafted Combiner network, based on such multimodal features, is extremely effective and outperforms more complex state of the art approaches on the popular FashionIQ dataset.

MMAsia '21, December 1–3, 2021, Gold Coast, Australia

ACM ISBN 978-1-4503-8607-4/21/12…$15.00
https://doi.org/10.1145/3469877.3493593

CCS CONCEPTS

• **Information systems** → **Image search**; *Users and interactive retrieval*; • **Computing methodologies** → **Neural networks**.

KEYWORDS

multimodal retrieval, deep neural networks, contrastive learning

ACM Reference Format:
Alberto Baldrati, Marco Bertini, Tiberio Uricchio, Alberto Del Bimbo. 2021. Conditioned Image Retrieval for Fashion using Contrastive Learning and CLIP-based Features. In *ACM Multimedia Asia (MMAsia '21), December 1–3, 2021, Gold Coast, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3469877.3493593

1 INTRODUCTION

Content-Based Image Retrieval (CBIR) is a fundamental task in multimedia and computer vision and has been applied to many different domains like art [10], commerce [16], medicine [20], security [2], nature [14], landmarks [26], etc. Typically image features of the database images are computed and compared with the features of a query image.

Interactive (i.e. conditioned) image retrieval systems extend CBIR systems to improve their effectiveness, by adding some form of user feedback, e.g. to provide some measure of relevance [3] or requesting constraints on some attributes of the retrieved results [25]. These types of systems can be applied in many different domains

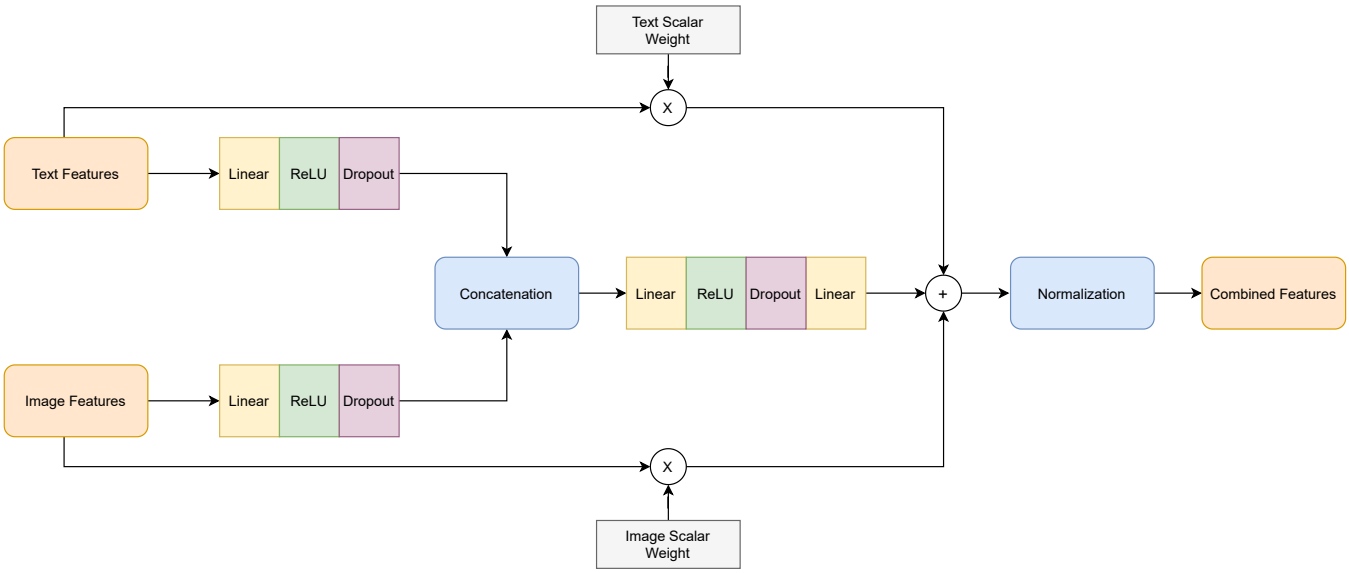

Figure 2: Architecture of the combiner network.

such as web search, e-commerce and surveillance. However, the difficulty in the development of these approaches is the need to incorporate features from the feedback and the intent of the user, in addition to the semantic gap between features and image content.

Very recently, it has been shown that a deep neural network like CLIP [22], trained using an image-caption alignment objective on large-scale internet data, can obtain impressive zero-shot transfer on a myriad of downstream tasks like image classification, text-based image retrieval, object detection and video action recognition.

In this work we show that CLIP-based features can be effectively used to implement a conditioned image retrieval system where user feedback is provided as natural language input to provide additional (or contrasting) requirements with respect to those embedded in the visual features of the image used to query the system. Figure 1 on the preceding page shows how the system works: a user selects a reference image and then provides additional requirements and requests in form of text, e.g. asking to change texture or shape features of the reference image. We apply the system to the fashion domain. Experiments are carried on the challenging FashionIQ dataset [28] obtaining state-of-the-art results.

2 PREVIOUS WORK

Traditional CBIR did not use any kind of user feedback or its intent to refine results. However, within interactive and conditioned CBIR, a lot of work has been done to improve retrieval performance incorporating user's feedback in terms of relevance to the query [23] or by considering relative [18] and absolute attributes [12, 30]. The limiting expressiveness of attributes was successively addressed in [11, 27] by considering purely textual feedback, allowing richer expressiveness. Nonetheless, performance of the textual model can limit the system in understanding and reacting appropriately. At the same time, GPT-2, BERT [21] and GPT-3 [4] models have shown that large amounts of text combined with recent improvements in attention mechanisms enable learning of powerful features that integrate vast knowledge. Adding images to the learning process, CLIP [22] has very recently shown that it is feasible to perform multimodal zero shot learning, obtaining remarkable feature generalization of both images and text. Contrary to standard vision models that are trained on typical datasets and that are good at only one task, this new class of models learn only associations between the abundant images and natural language supervision available on the internet. They are not directly optimized for a benchmark and yet are able to perform consistently well on different tasks. CLIP effectiveness is still subject of study [1], with first applications to art [7], image generation [9] and zero shot video retrieval [8]. Our work builds upon CLIP and further explores its potential in the task of conditioned image retrieval, applying the proposed approach to fashion.

In the growing area of image retrieval with user feedback, our work is related to the recently introduced conditioned fashion image retrieval with text [28]. In [6], a transformer that can be seamlessly plugged in a CNN to selectively preserve and transform the visual features conditioned on language semantics is presented. In [24, 27] they use skip connections and combine them with graph neural networks, reporting improved performance. In [19], image style and content are considered separately by two different neural network modules. In [17] a *Correction Network* is added which explicitly models the difference between the reference and target image in the embedding space.

Differently from these work, our method differs by few factors. It explicitly considers a learned manifold of visual and text features with the goal of learning an additive transformation in the same space. Moreover, our approach does not use any kind of spatial information. Instead, in [19, 27] features extracted from the backbone are 3-dimensional and the composition takes care of spatial information, in [6] the features are extracted at different convolutional layers from the ResNet-50 backbone. In [17] the authors divided the image and the sentence into a set of localized components assigning

a representation module, denoted as *experts*, to each of them. More similar to our work is [24] which trains a combiner directly on flattened image and text features that, differently from our work, are obtained from different embeddings.

3 THE PROPOSED METHOD

The proposed method addresses the multimodal problem of conditional fashion image retrieval. Given as input a reference image (e.g. an image of a black dress) and a text that includes a descriptive request from the user in relation to the image ("red and yellow"), the goal of the retrieval is to retrieving the best matching images that satisfy similarity constraints imposed by both of the input components (an image of a red and yellow dress). To retrieve correct images, the system must be able to understand both the contents of the image and text, and further add the textual comment to the image content.

A schema of the complete system is shown in Figure 1 on page 1. In contrast to previous works like [6, 17, 19, 24] that build from different image and textual model, we start from the hypothesis of having a common embedding of images and text, realized by CLIP. As shown in [22], similar concepts expressed in text and images tend to share similar features, or at least be "near" in the common space.

The input image and text are encoded using their respective CLIP encoders into features in the common space. The task is then cast as a problem of learning a transformation from the reference image feature and input text to a combined feature that includes both the multimodal input information and is as near as possible to the common manifold. We denote this transformation as a *Combiner* function and design a neural network architecture that is trained to learn the correct function. We explore different Combiner functions showing that state of the art performance is obtainable.

The Combiner function, depicted in Figure 2 on the facing page, is simple yet more performing than more complex architectures that we tested. The idea is to build an additive transformation where text, image and the combination of both are all added into the final combined feature. The text and image features are each weighted by a scalar that is trained to balance their contribution. We found these two contributions essential to obtain a new state of the art performance. The third contribution is given from the mixture of image and text. Starting from text and image features, we apply to each feature a linear transformation followed by the ReLU function. Features are then concatenated and the output is fed to another linear layer that is followed by a ReLU and a final linear layer. The three contributions are finally summed and L_2 normalized. Dropout is applied to each layer to reduce overfitting.

Training of the system is performed with triplets of input images, text and target images. Following [24, 27] we employ the DML loss as pairwise contrastive loss using the normalized dot product as similarity kernel. Similarly to CLIP [22], we multiply the logits (i.e. the dot product between predicted and target features) by 100 before computing the loss. This was shown to help the training process by improving the dynamic range of features.

3.1 Implementation Details

We decided to perform experiments using two CLIP models of different size. The smallest one is based on a modified ResNet-50 (RN50) [13] architecture. It takes as input images of 224×224 pixels and outputs features of 1024 dimensions. The biggest one, denoted as RN50x4, follows the EfficientNet-style model scaling and use approximately 4x the computation of the smallest. It takes as input images of 288×288 pixels and outputs features of 640 dimensions. In the experiments, the CLIP encoders have been kept frozen and the only trained part of the model is the Combiner function. The dropout rate was set to 0.5 as commonly done with linear layers. The text and image scalar weights were both initialized to 1. We used PyTorch in our experiments. The learning rate was set to $5e-5$ and we trained the model for a maximum of 300 epochs. The batch size was set to 1024 for the experiments with RN50 and 512 for the experiments with RN50x4, due to memory limits.

4 EXPERIMENTAL RESULTS

	Average	
Model	**R@10**	**R@50**
Sum	19.55	38.40
Weighted sum	19.78	39.04
No skip	23.38	46.81
Linear after skip	23.36	47.42
No Dropout	28.36	51.62
No ReLU & Dropout	28.20	51.10
CLIP fine-tuning	27.91	51.50
Proposed model	**29.67**	**53.41**

Table 1: Recall at K on the validation set, with variations on the architecture. Best score is highlighted in bold.

	Average	
Batch size	**R@10**	**R@50**
64	28.75	51.94
128	29.01	52.41
256	29.10	52.58
512	29.00	53.02
1024	**29.67**	**53.41**

Table 2: Recall at K on the validation set when increasing the batch size. Best score is highlighted in bold.

4.1 Dataset and metrics

We used the popular FashionIQ dataset [28] since it is commonly used to test conditioned image retrieval. FashionIQ provides 77,684 fashion images crawled from the web and split in train, validation and test sets, divided into three different categories: *Dress*, *Toptee* and *Shirt*. Among the 46,609 training images there are 18,000 training triplets made of a candidate image, a pair of user texts and a target image. The texts describe properties to modify in the candidate image to match the target image. Validation and test set have, respectively, 15,537 and 15,538 images with 6,017 and 6,119 triplets.

Method	Shirt		Dress		Toptee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
JVSM [5]	12.0	27.1	10.7	25.9	13.0	26.9	11.9	26.6
TRACE w/BERT [15]	20.80	40.80	22.70	44.91	24.22	49.80	22.57	46.19
VAL w/GloVe [6]	22.38	44.15	22.53	44.00	27.53	51.68	24.15	46.61
CurlingNet [29]	21.45	44.56	26.15	53.24	30.12	55.23	25.90	51.01
RTIC-GCN [24]	22.72	44.16	27.71	53.50	29.63	56.30	26.69	51.32
CoSMo [19]	24.90	49.18	25.64	50.30	29.21	57.46	26.58	52.31
DCNet [17]	23.95	47.30	**28.95**	**56.07**	30.44	58.29	27.78	53.89
Our (CLIP-RN50)	31.41	52.11	25.69	50.64	31.91	57.50	29.67	53.41
Our (CLIP-RN50x4)	**35.76**	**56.20**	27.20	53.57	**36.31**	**61.14**	**33.09**	**56.99**

Table 3: Comparison between our method and current state-of-the-art models on the Fashion-IQ validation set. Best scores are highlighted in bold.

We follow experimental setting as in [17, 19]. We employ the average recall at rank K (Recall@K) as evaluation metric, namely Recall@10 (R@10) and Recall@50 (R@50). Note that for each triplet there is only a positive index image. Hence, each individual query has R@K either of zero or one. All results are on the validation set since at the time of writing the test set ground-truth labels has not been released yet.

4.2 Ablation studies

In this section we show preliminary experiments with variations of the architecture shown in Figure 2 on page 2, and with different batch sizes. All experiments were performed with RN50 as backbone.

We tested the following baselines:

- **Sum**: image and text features are summed;
- **Weighted sum**: a weighted sum between the image and text features, i.e. the model without the mixture contribution of text and image;
- **No skip**: only the mixture contribution of text and image;
- **Linear after skip**: the regular model with an additional linear layer in both text and image contributions;
- **No Dropout**: without dropout layers;
- **No ReLU & Dropout**: without ReLU activations and dropout layers;
- **CLIP fine-tuning**: end-to-end fine-tuned CLIP with the Combiner function;
- **Proposed model**: the proposed model shown in Figure 2 on page 2.

We report the results for each variation in Table 1 on the previous page.

The first interesting thing to notice is that a simple sum of the candidate image features and the relative captions features led to decent results that are not too far from the worst competing state-of-the-art methods. This confirms that text and images in the CLIP embedding reside (approximately) in the same manifold. The weighted sum baseline, where text and image weights are learned, results in little improvement. The two weights stabilize respectively to 1 and 0.80 for images and text, signaling a preference towards image features. Compared to the proposed model, we note that removing the text and image direct contributions lead to a significant drop in performance. Given the effectiveness of the Sum baseline, this is reasonable, since their presence may enable the Combiner function to only learn an offset to an already good starting point. In our experiments, fine-tuning CLIP along Combiner training did not bring any performance improvement.

Regarding the batch size, we tested different value ranging from 64 to 1024. We report the performance obtained in Table 2 on the preceding page. We note that increasing the batch size provides a $\sim$ 3% increase of both recall measures.

4.3 Comparison with SotA

Table 3 shows the quantitative results on Fashion-IQ validation set. Our approach outperforms the state-of-the-art by improving up to $\sim$ 5% in average R@10 and 3% in average R@50 upon the best method, DCNet [17], when using the CLIP RN50x4 backbone. Our method have the highest recall in the Shirt and Toptee categories, with comparable performance in the Dress category, using both backbones. Between the two backbones, we note that bigger RN50x4 obtains the best performance, with an improvement on the smaller RN50 in the range of about 2% to 4% in all categories.

5 CONCLUSIONS

In this paper we tackled the problem of conditioned image retrieval for fashion using the recent CLIP model where we exploited its zero shot transfer features. We developed a Combiner network that is able to compute a combined feature made from reference images integrated with a textual description. Experiments on the FashionIQ dataset show that our approach is able to outperform more complex state of the art methods.

Our future work will deal with the extension of the proposed method to videos and further experimentation with different image domains.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911 - AI4Media.

REFERENCES

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications. *arXiv preprint arXiv:2108.02818* (2021).

[2] Jamil Ahmad, Khan Muhammad, Sambit Bakshi, and Sung Wook Baik. 2018. Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets. *Future Generation Computer Systems* 81 (2018), 314–330.

[3] Imon Banerjee, Camille Kurtz, Alon Edward Devorah, Bao Do, Daniel L Rubin, and Christopher F Beaulieu. 2018. Relevance feedback for enhancing content based image retrieval and automatic prediction of semantic image features: Application to bone tumor radiographs. *Journal of biomedical informatics* 84 (2018), 123–135.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[5] Yanbei Chen and Loris Bazzani. 2020. *Learning Joint Visual Semantic Matching Embeddings for Language-Guided Retrieval.* 136–152. https://doi.org/10.1007/978-3-030-58542-6_9

[6] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image Search With Text Feedback by Visiolinguistic Attention Learning. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR).*

[7] Marcos V Conde and Kerem Turgutlu. 2021. CLIP-Art: Contrastive Pre-Training for Fine-Grained Art Classification. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR).* 3956–3960.

[8] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. *arXiv preprint arXiv:2106.11097* (2021).

[9] Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. 2021. Generating images from caption and vice versa via CLIP-Guided Generative Latent Space Search. *arXiv preprint arXiv:2102.01645* (2021).

[10] Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proc. of European Conference on Computer Vision (ECCV) Workshops.* 0–0.

[11] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Schmidt Feris. 2018. Dialog-based Interactive Image Retrieval. arXiv:1805.00145 [cs.CV]

[12] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. 2017. Automatic Spatially-aware Fashion Concept Discovery. arXiv:1708.01311 [cs.CV]

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[14] Bogdan Ionescu, Henning Müller, Renaud Péteri, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Dzmitri Klimuk, Aleh Tarasau, Asma Ben Abacha, Sadid A Hasan, et al. 2019. ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In *Proc. of International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF).* Springer, 358–386.

[15] Surgan Jandial, Ayush Chopra, Pinkesh Badjatiya, Pranit Chawla, Mausoom Sarkar, and Balaji Krishnamurthy. 2020. TRACE: Transform Aggregate and Compose Visiolinguistic Representations for Image Search with Text Feedback. arXiv:2009.01485 [cs.CV]

[16] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. 2017. Cross-domain image retrieval with attention modeling. In *Proc. of ACM Multimedia (ACMMM).* 1654–1662.

[17] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021. Dual Compositional Learning in Interactive Image Retrieval. In *Proc. of AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 35. 1771–1779. https://ojs.aaai.org/index.php/AAAI/article/view/16271

[18] Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2015. WhittleSearch: Interactive Image Search with Relative Attribute Feedback. *International Journal of Computer Vision* 115, 2 (Apr 2015), 185–210. https://doi.org/10.1007/s11263-015-0814-0

[19] Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR).* 802–812.

[20] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. 2017. Medical image retrieval using deep convolutional neural network. *Neurocomputing* 266 (2017), 8–20.

[21] Yuanbin Qu, Peihan Liu, Wei Song, Lizhen Liu, and Miaomiao Cheng. 2020. A Text Generation and Prediction System: Pre-training on New Corpora Using BERT and GPT-2. In *Proc. of IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC).* IEEE, 323–326.

[22] 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. arXiv:2103.00020 [cs.CV]

[23] Yong Rui, T.S. Huang, M. Ortega, and S. Mehrotra. 1998. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 8, 5 (1998), 644–655. https://doi.org/10.1109/76.718510

[24] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. 2021. RTIC: Residual Learning for Text and Image Composition using Graph Convolutional Network. *arXiv preprint arXiv:2104.03015* (2021).

[25] Haibo Su, Peng Wang, Lingqiao Liu, Hui Li, Zhen Li, and Yanning Zhang. 2020. Where to Look and How to Describe: Fashion Image Retrieval with an Attentional Heterogeneous Bilinear Network. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[26] Federico Vaccaro, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2020. Image Retrieval Using Multi-Scale CNN Features Pooling. In *Proc. of ACM International Conference on Multimedia Retrieval (ICMR)* (Dublin, Ireland) *(ICMR '20).* Association for Computing Machinery, New York, NY, USA, 311–315. https://doi.org/10.1145/3372278.3390732

[27] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2018. Composing Text and Image for Image Retrieval - An Empirical Odyssey. arXiv:1812.07119 [cs.CV]

[28] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2020. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. arXiv:1905.12794 [cs.CV]

[29] Youngjae Yu, Seunghwan Lee, Yuncheol Choi, and Gunhee Kim. 2020. CurlingNet: Compositional Learning between Images and Text for Fashion IQ Data. arXiv:2003.12299 [cs.CV]

[30] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR).* 6156–6164. https://doi.org/10.1109/CVPR.2017.652