Boosting Fine-grained Fashion Retrieval with Relational Knowledge Distillation

Ling Xiao
The University of Tokyo
ling@cvm.t.u-tokyo.ac.jp

Toshihiko Yamasaki
The University of Tokyo
yamasaki@cvm.t.u-tokyo.ac.jp

Abstract

Fine-grained fashion retrieval (FGFR) aims to retrieve fashion items from a database that match specific and detailed attributes of a query image. This task requires a model to discern subtle variations, which is more challenging than general recognition tasks. To improve retrieval accuracy, we propose an online Knowledge Distillation (KD) framework that leverages KD's advantages in feature extraction. We also introduce a novel relational knowledge distillation (RKD) strategy that outperforms conventional KD by focusing on relational information. The proposed KD framework and RKD strategy can be easily applied to existing state-of-the-art FGFR models to significantly improve retrieval accuracy, such as a +7.72% increase in mAP on the FashionAI dataset for ASENet_V2. The source code is available in https://github.com/Dr-LingXiao/RKD.

1. Introduction

Fashion modeling and analysis are crucial for understanding consumer preferences. Similarity-based retrieval [2, 5, 6, 14, 15, 23], especially in-shop and cross-domain fashion retrieval [1, 10, 12, 15, 17], is a key research area. However, most methods focus on whole image similarity [5, 7, 11, 12, 19, 30], while fine-grained fashion retrieval (FGFR) remains underexplored. FGFR identifies specific regions and features within an image to distinguish between fashion items [3, 16, 20, 21, 24, 26]. Incorporating images and attributes in modeling enhances accuracy and matches user preferences. FGFR also plays a vital role in fashion copyright protection by detecting design plagiarism.

However, FGFR is a complex task with three main difficulties: 1) **Intra-attribute variation**, where a single attribute (e.g., skirt-length) can have several sub-classes, requiring precise discernment by the model. 2) **Subtle variation**, where items within the same attribute have subtle differences that are challenging for conventional computer vision techniques. 3) **Viewpoint variation**, where fashion items can appear differently based on viewpoint, pose, or orientation in images, necessitating robust retrieval models

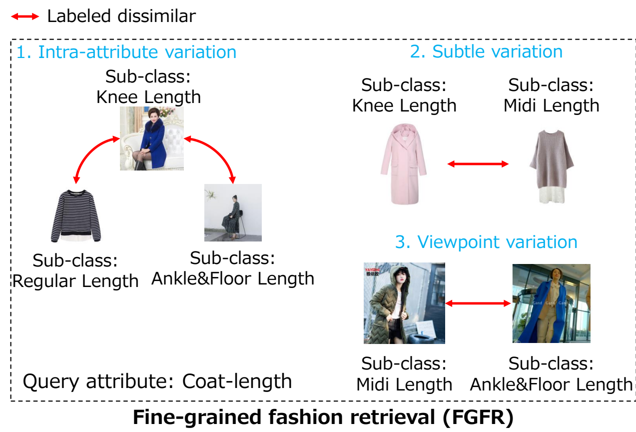

Figure 1. The **intra-attribute variation**, **subtle variation**, and **viewpoint variation**.

(see Figure 1).

To address FGFR, various methods have been proposed, including learning an overall embedding space with a fixed mask (Veit *et al.* [20]), multiple attribute-specific embedding spaces (Ma *et al.* [16]), and fusing attribute-aware spatial and channel attention (Wan *et al.* [21]). Yan *et al.* [26] employed iterative learning, Jiao *et al.* [13] incorporated instance and cluster level supervisions, and Xiao *et al.* [24] proposed a contrastive learning method. However, these methods focus on complex attribute-guided embedding modules, neglecting the discrimination of Convolution Neural Network (CNN) extracted image features. Two open questions remain: how to extract more discriminative image representations and how to better learn relational information in FGFR.

In this work, we propose a general online knowledge distillation (KD) framework and a novel relational knowledge distillation (RKD) strategy to improve FGFR, addressing the aforementioned problem. Our paper has the following technical contributions:

- We propose a general online KD framework that can be adopted in existing FGFR methods to boost their performance by extracting more powerful image embeddings.
- We present an innovative RKD strategy that transfers knowledge at the relational level, outperforming the tra-

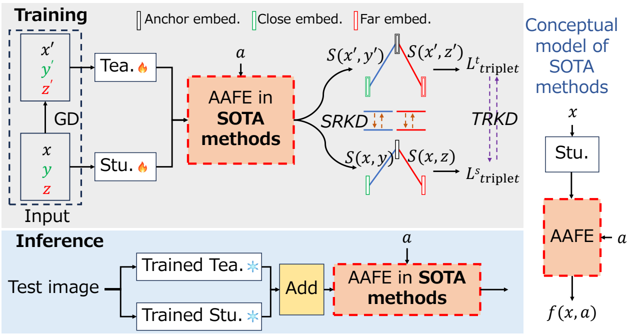

Figure 2. Applying proposed methods to existing FGFR methods

ditional KD strategy that transfers knowledge at the individual item level.

- Experiments show that the proposed methods can consistently significantly improve the state-of-the-art FGFR methods in fine-grained fashion retrieval on the FashionAI dataset. It also shows effectiveness in relation prediction on the Zappos50k dataset.

2. Method

2.1. Preliminaries

Conventional online KD. Conventional online KD [29] methods transfer knowledge at the item level [4, 9, 18, 28], using a peer network to provide training experience. For example, in an image classification task, network Θ_1 predicts labels, while peer network Θ_2 provides its posterior probability p_2. The match between predictions p_1 and p_2 is measured using Kullback-Leibler (KL) Divergence, as shown in Eq. 1.

$$p_1^m(x_i) = \frac{\exp(z_1^m)}{\sum_{m=1}^{M} \exp(z_1^m)}, \tag{1a}$$

$$D_{\mathrm{KL}}(p_2||p_1) = \sum_{i=1}^{N} \sum_{m=1}^{M} p_2^m(x_i) \log \frac{p_2^m(x_i)}{p_1^m(x_i)}, \tag{1b}$$

where x_i is the input image, m is the class number, and z^m is the output of the cohort network. The symmetric Jensen-Shannon Divergence loss can commonly be expressed as:

$$L_{\mathrm{KD}} = \frac{1}{2}\left(D_{\mathrm{KL}}(p_1||p_2) + D_{\mathrm{KL}}(p_2||p_1)\right). \tag{2}$$

Essentially, conventional online KD transfers individual outputs of different networks. They can not directly improve relational information learning.

SOTA FGFR methods. The basic pipeline of state-of-the-art (SOTA) FGFR methods [16, 20, 21, 24, 26] is given below. In the training phase, triplet inputs (with attributes) are sampled. Taking one triplet $\{x_i, y_i, z_i|a\}$ with $i \in \{1, 2, \ldots, N\}$ as an example, a CNN backbone (Resnet50) is used to extract image representations $\{f(x_i), f(y_i), f(z_i)\}$, where x_i is an anchor image, y_i and z_i denote far and close images, respectively. Then, an attribute-aware feature extraction (AAFE) module is used to extract attribute-aware embeddings from image representations under the guidance of attribute a, resulting in $\{f(x_{i,a}), f(y_{i,a}), f(z_{i,a})\}$. Afterwards, the similarities between $(f(x_{i,a}), f(y_{i,a}))$ and $(f(x_{i,a}), f(z_{i,a}))$ are calculated, denoted as $S(x_{i,a}, y_{i,a})$ and $S(x_{i,a}, z_{i,a})$ respectively. Finally, a triplet loss L_{triplet} is calculated and used to update the model.

$$L_{\mathrm{triplet}} = \max\{0, m + S(x_{i,a}, z_{i,a}) - S(x_{i,a}, y_{i,a})\}, \tag{3}$$

where m denotes a margin and is set as 0.2.

2.2. Proposed methods

Online KD framework. Our online KD framework is motivated by two key points: 1) Using two backbones with varied inputs and architectures enhances feature diversity, while combining their outputs preserves information. 2) To handle minor differences between attribute sub-classes, we employ soft geometrical distortion (GD) techniques, boosting learning without major differences between two backbones' features.

Figure 2 shows the proposed KD framework. For easier expression, we denote the higher capacity backbone as the teacher and the other as the student. During training, 1) the input images $\{x_i, y_i, z_i|a\}$ are transformed with soft GD, and $\{x_i', y_i', z_i'\}$ are obtained; 2) the original and distorted inputs are then passed into the student and teacher backbones respectively, and $\{f(x_i), f(y_i), f(z_i)\}$ and $\{f(x_i'), f(y_i'), f(z_i')\}$ are obtained; 3) the student and teacher extracted image features and attribute a are processed with the AAFE module in SOTA methods separately, and $\{f(x_{i,a}), f(y_{i,a}), f(z_{i,a})\}$ and $\{f(x_{i,a}'), f(y_{i,a}'), f(z_{i,a}')\}$ are obtained; 4) the obtained embeddings are then used for calculating the RKD loss and the triplet loss. During inference, the test image is passed through both student and teacher backbones to extract feature embeddings, which are then fused by addition. The fused output is processed by an AAFE module to obtain attribute-aware embeddings for similarity evaluation.

RKD. Previous KD strategies primarily focused on transferring knowledge at the item level. However, this approach may not be as effective for FGFR, which prioritizes similarities. In this paper, we introduce a RKD strategy that transfers knowledge at the relational level. This strategy encompasses Triplet Relational Knowledge Distillation (TRKD) and Similarity Relational Knowledge Distillation (SRKD), aiming to transfer the mutual relations between data examples at different levels. The TRKD is based on the output of Eq. 3, expressed as Eq. 4. The SRKD is expressed as Eq. 5. We use an Mean Squared Error (MSE) loss for RKD.

$$L_{\mathrm{TRKD}} = D_{\mathrm{mse}}(L^s_{\mathrm{triplet}}, L^t_{\mathrm{triplet}}), \tag{4}$$

Table 1. Performance comparison. Numbers in bold indicate the best performance for each attribute.

Methods	KD strategies	Skirt -length	Sleeve -length	Coat -length	Pant -length	Collar -design	Lapel -design	Neckline -design	Neck -design	MAP ↑
		MAP for each attribute (Fashion AI) ↑								
ASENet_V2 [16]	w/o	64.57	54.96	51.76	64.50	71.93	66.72	60.29	60.83	60.76
	Conventional KD	66.82	59.39	56.21	71.15	75.12	71.83	65.97	64.92	65.35 (+4.59)
	Ours	**69.28**	**62.13**	**59.72**	**73.08**	**80.11**	**74.08**	**68.98**	**70.04**	**68.48 (+7.72)**
AttnFashion [21]	w/o	62.22	47.05	46.15	63.10	**72.87**	**65.05**	**52.87**	58.76	56.47
	Conventional KD	61.31	44.45	44.11	64.28	71.84	50.78	49.48	59.04	53.86 (-2.61)
	Ours	**64.53**	**50.53**	**48.07**	**65.74**	71.23	57.26	50.71	**59.50**	**56.78 (+0.31)**
ISLN [26]	w/o	56.35	24.90	35.34	59.50	37.51	29.93	22.37	22.72	35.14
	Conventional KD	55.47	28.23	35.25	59.24	36.89	29.32	20.88	23.69	35.24 (+0.1)
	Ours	**58.41**	**33.62**	**37.82**	**62.62**	**42.41**	**31.35**	**29.91**	**25.43**	**39.87 (+4.73)**
ASENet_V2+PT [24]	w/o	67.50	60.52	55.20	70.58	77.35	72.31	68.31	67.28	66.29
	Conventional KD	**70.13**	61.34	58.54	72.20	77.94	73.79	69.71	66.35	67.80 (+1.51)
	Ours	68.94	**62.13**	**60.88**	**73.56**	**78.20**	**77.77**	**69.94**	**69.32**	**69.14 (+2.85)**

Table 2. Ablation studies on FashionAI dataset: Performance changes when removing GD, TRKD or SRKD from the whole framework.

Methods	Ours GD	Ours TRKD	Ours SRKD	Skirt -length	Sleeve -length	Coat -length	Pant -length	Collar -design	Lapel -design	Neckline -design	Neck -design	MAP ↑
				MAP for each attribute (Fashion AI) ↑								
ASENet_V2 [16] +ours				64.57	54.96	51.76	64.50	71.93	66.72	60.29	60.83	60.76
			✓	69.42	59.15	57.55	72.24	76.54	74.24	66.19	68.29	66.56
		✓		68.74	59.40	56.54	71.11	77.07	73.34	65.32	68.30	66.04
	✓	✓	✓	69.28	62.13	59.72	73.08	80.11	74.08	68.98	70.04	68.48

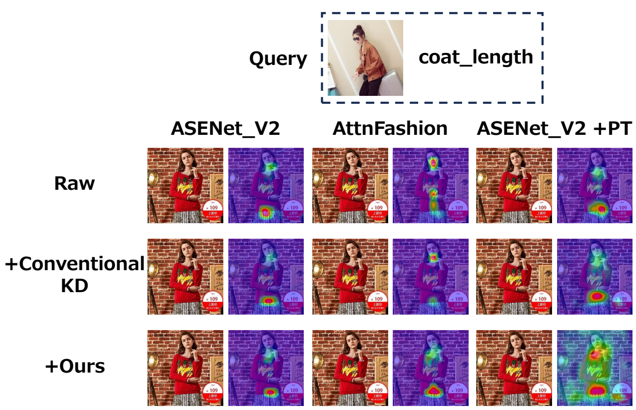

Figure 3. Visualization of the spatial attention based on a specified query attribute, depicted above the original query image. Our methods enhance the baseline's ability in localizing related region.

$$L_{\text{SRKD}} = \frac{1}{2}(D_{\text{mse}}(S(x_{i,a}, y_{i,a}), S(x'_{i,a}, y'_{i,a})) + D_{\text{mse}}(S(x_{i,a}, z_{i,a}), S(x'_{i,a}, z'_{i,a}))), \tag{5}$$

where L^s_{triplet} is triplet loss output of student network, which is also the loss of SOTA methods. $S(x_{i,a}, y_{i,a}) = f^s(x_{i,a}) \cdot f^s(y_{i,a})$ and $S(x'_{i,a}, y'_{i,a}) = f^t(x'_{i,a}) \cdot f^t(y'_{i,a})$.

Fused model learning. The detailed training process when applying our KD framework and RKD to SOTA methods are given in Algorithm 1. The final loss is a weighted combination of L_{TRKD}, L_{SRKD}, and L^s_{triplet}, denoted as:

$$L = \frac{1}{n\sigma_1^2} L^s_{\text{triplet}} + \frac{1}{n\sigma_2^2} L_{\text{TRKD}} + \frac{1}{n\sigma_3^2} L_{\text{SRKD}} + \log(1+\sigma_1^2) + \log(1+\sigma_2^2) + \log(1+\sigma_3^2), \tag{6}$$

where n denotes numbers of losses, σ_{1-3} are learned weight parameters and are initialized as 1.0. All weighted loss used in this paper are fused in this way.

3. Experiments

3.1. Experimental settings

We used two datasets: FashionAI [20] and Zappos50k [27]. For single-backbone models, we used ResNet50 [8]. When applying our methods, we paired SE-ResNext50-32x4d [25] with ResNet50 to avoid capacity gaps [22] and align with existing FGFR methods. We evaluated our KD framework and RKD on SOTA FGFR methods: ASENet_V2, AttnFashion, ISLN, and ASENet_V2+PT. For available source codes (ASENet_V2, ASENet_V2+PT), we used official implementations; otherwise, we re-implemented based on papers. Experiments were conducted

Algorithm 1 Training of SOTA methods when adopting our KD framework and RKD.

A teacher network SE-ResNext50-32x4d, named Tea..
A student network Resnet50, named Stu..
A set of attributes A, labeled image set I.
A batch of image triplets $\{x_i, y_i, z_i | a\} \in I$.
GD method.
if in training stage **then**
  for $i = 0$ **to** $N - 1$
    Obtain $\{x_i', y_i', z_i'\}$ with GD.
    Obtain image embeddings $\{f(x_i'), f(y_i'), f(z_i')\}$ and $\{f(x_i), f(y_i), f(z_i)\}$ using Tea. and Stu..
    Obtain $\{f(x_{i,a}'), f(y_{i,a}'), f(z_{i,a}')\}$ and $\{f(x_{i,a}), f(y_{i,a}), f(z_{i,a})\}$ using AAFE module in SOTA methods with guidance of attribute a.
  end for
  Update the whole model with Eq. 6.

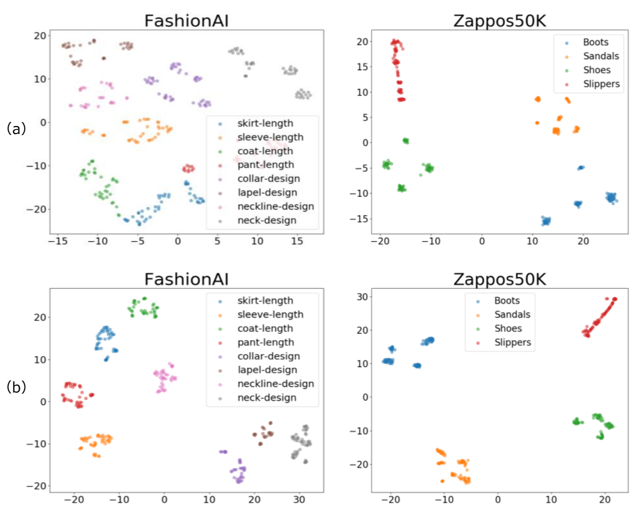

Figure 4. The t-SNE visualization of learned attribute-aware embeddings with (a) ASENet_V2 and (b) ASENet_V2 +ours.

on a V100 GPU with PyTorch 1.1.0, using a batch size of 16, embedding dimension of 1024, and a StepLR scheduler with an initial learning rate of 1×10^{-4}. All settings were consistent for fairness, except ISLN's learning rate was 1×10^{-5} for convergence. We used mAP as the main metric and applied geometrical distortions with shear and rotation degrees from $-15°$ to $15°$, and perspective transformation degrees from 0 to 0.1.

3.2. Experimental results

Main results. Our method outperforms conventional KD on FashionAI dataset (Table 1). Conventional KD sometimes degrades the baseline, but the combination of our KD framework and RKD consistently enhance it. Our methods also further improve ASENet_V2+PT, demonstrating better feature representations for FGFR.

Table 3. Triplet relation prediction on Zappos50k.

Methods	Our RKD	Zappos50k Average loss (%) ↓	Zappos50k Prediction Accuracy (%) ↑
ASENet_V2 [16]	w/o	0.0430	92.54
	w	**0.0305**	**95.02**
AttnFashion [21]	w/o	0.0664	91.37
	w	**0.0443**	**93.97**
ISLN [26]	w/o	0.0761	86.87
	w	**0.0689**	**89.06**
ASENet_V2+PT [24]	w/o	0.0414	92.95
	w	**0.0302**	**95.14**

We visualized some attention maps (Figure 3). For example, in length-related attributes, our methods enhance the ability to locate fashion item endpoints, contributing to performance improvement. Figure 4 also demonstrates more discriminative feature representations with our method.

Ablation experiments. Ablation experiments evaluated the effectiveness of different components of our proposed methods: GD in our KD framework, TRKD, and SRKD, using ASENet_V2 as the baseline (Table 2). When only TRKD or SRKD is adopted, the teacher and student have same inputs ($\{x_{i,a}, y_{i,a}, z_{i,a}\}$). Experiments show that using TRKD or SRKD individually improves the baseline significantly, with the best performance achieved when all components are combined.

Performance on relation prediction task. Experiments on the Zappos50k dataset [27] evaluated the effectiveness of our proposed RKD in improving existing FGFR methods for triplet relation prediction (Table 3). The results show that RKD consistently enhances FGFR methods on Zappos50k, with an average improvement of 2.5% points in prediction accuracy for all baseline methods. This confirms RKD's effectiveness in relation prediction.

4. Conclusions

This paper proposed a general online KD framework and a novel RKD strategy to enhance existing FGFR methods. The RKD strategy outperforms conventional item-level KD and enhances triplet relation prediction on the Zappos50k dataset. These methods can be applied to recommendation and other retrieval tasks, offering valuable insights for future research in FGFR and KD.

5. Acknowledgments

This paper is partially financially supported by JSPS KAKENHI Grant Number 24K20787 and Institute for AI and Beyond of The University of Tokyo.

References

[1] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV 2018)*, pages 1671–1679, 2018. 1

[2] Eric Dodds, Jack Culpepper, and Gaurav Srivastava. Training and challenging models for text-guided fashion image retrieval. *arXiv preprint arXiv:2204.11004.*, 2022. 1

[3] J. Dong, Z. Ma, X. Mao, X. Yang, Y. He, R. Hong, and S. Ji. Fine-grained fashion similarity prediction by attribute-specific embedding learning. *IEEE Transactions on Image Processing*, 30:8410–8425, 2021. 1

[4] S. Fan, F. Zhu, Z. Feng, Y. Lv, M. Song, and F.Y. Wang. Conservative-progressive collaborative learning for semi-supervised semantic segmentation. *IEEE Transactions on Image Processing*, 31:949–961, 2023. 2

[5] Bojana Gajic and Ramon Baldrich. Cross-domain fashion image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 1869–1871, 2018. 1

[6] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 14105–14115, 2022. 1

[7] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM International Conference on Multimedia (ACMMM 2017)*, pages 1078–1086, 2017. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 770–778, 2016. 3

[9] Mengshun Hu, Kui Jiang, Zheng Wang, Xiang Bai, and Ruimin Hu. Cycmunet+: Cycle-projected mutual learning for spatial-temporal video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13376–13392, 2023. 2

[10] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015)*, pages 1062–1070, 2015. 1

[11] Sarah Ibrahimi, Nanne van Noord, Zeno Geradts, and Marcel Worring. Deep metric learning for cross-domain fashion instance retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2019)*, pages 3165–3168, 2019. 1

[12] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. Cross-domain image retrieval with attention modeling. In *Proceedings of the 25th ACM International Conference on Multimedia (ACMMM 2017)*, pages 1654–1662, 2017. 1

[13] Yang Jiao, Ning Xie, Yan Gao, Chien-Chih Wang, and Yi Sun. Fine-grained fashion representation learning by online deep clustering. In *Proceedings of the European Conference on Computer Vision (ECCV 2022)*, pages 19–35, 2022. 1

[14] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2019)*, pages 3066–3075, 2019. 1

[15] 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 (CVPR 2016)*, pages 1096–1104, 2016. 1

[16] Zhe Ma, Jianfeng Dong, Zhongzi Long, Yao Zhang, Yuan He, Hui Xue, and Shouling Ji. Fine-grained fashion similarity learning by attribute-specific embedding network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020)*, pages 11741–11748, 2020. 1, 2, 3, 4

[17] Chen Ning, Yang Di, and Li Menglu. Survey on clothing image retrieval with cross-domain. *Complex & Intelligent Systems*, pages 1–14, 2022. 1

[18] Changhwa Park, Junho Yim, and Eunji Jun. Mutual learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2023)*, pages 2675–2684, 2023. 2

[19] Vivek Sharma, Naila Murray, Diane Larlus, Saquib Sarfraz, Rainer Stiefelhagen, and Gabriela Csurka. Unsupervised meta-domain adaptation for fashion retrieval. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV 2021)*, pages 1348–1357, 2021. 1

[20] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 830–838, 2017. 1, 2, 3

[21] Yongquan Wan, Kang Yan, Cairong Yan, and Bofeng Zhang. Learning attribute-guided fashion similarity with spatial and channel attention. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–17, 2022. 1, 2, 3, 4

[22] Maorong Wang, Hao Yu, Ling Xiao, and Toshihiko Yamasaki. Bridging the capacity gap for online knowledge distillation. In *Proceedings of the IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR 2023)*, pages 1–4, 2023. 3

[23] Ling Xiao and Toshihiko Yamasaki. Sat: Self-adaptive training for fashion compatibility prediction. In *Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP 2022)*, pages 2431–2435, 2022. 1

[24] Ling Xiao, Xiaofeng Zhang, and Toshihiko Yamasaki. Toward a more robust fine-grained fashion retrieval. In *Proceedings of the IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR 2023)*, pages 1–4, 2023. 1, 2, 3, 4

[25] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 1492–1500, 2017. 3

[26] Cairong Yan, Kang Yan, Yanting Zhang, Yongquan Wan, and Dandan Zhu. Attribute-guided fashion image retrieval by iterative similarity learning. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2022)*, pages 1–6, 2022. 1, 2, 3, 4

[27] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, pages 192–199, 2014. 3, 4

[28] Miao Zhang, Li Wang, David Campos, Wei Huang, Chenjuan Guo, and Bin Yang. Weighted mutual learning with diversity-driven model compression. *Advances in Neural Information Processing Systems*, 35:11520–11533, 2022. 2

[29] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pages 4320–4328, 2018. 2

[30] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 1520–1528, 2017. 1