Ontology Guided Supervised Contrastive Learning for Fine-grained Attribute Extraction from Fashion Images

Shubham Paliwal[1,*], Bhagyashree Gaikwad[1], Mayur Patidar[1], Manasi Patwardhan[1], Lovekesh Vig[1], Meghna Mahajan[1], Bagya Lakshmi V[1] and Shirish Karande[1]

[1]*TCS Research, India*

Abstract

Fashion attributes are key to many downstream tasks in e-commerce such as product recommendation, fashion captioning, item matching, fashion image retrieval and generation. Generally, fashion attributes are arranged in an ontology where one fashion attribute may be assigned one or more values. Most state-of-the-art (SOTA) approaches model attribute extraction as a multi-label classification problem and do not consider attribute-value relatedness information during training which leads to poor performance on fine-grained attribute extraction. To address this issue, we propose Ontology Guided Supervised Contrastive Learning For Fine-grained Fashion Attribute Extraction (*OGSCL-FAE*) where we leverage a fashion ontology to create strong negative pairs, model attribute extraction as a matching problem, and fine-tune a pre-trained CLIP on attribute extraction. The proposed approach outperforms existing SOTA approaches on two public datasets DeepFashion and FashionAI by 11.65% top-5 recall rate and 0.93 mAP respectively.

Keywords

Fashion attribute classification, Ontology guided training, Supervised contrastive loss, CLIP

1. Introduction

Extraction of appropriate fine-grained attributes from fashion images is a pre-requisite for automation of multiple e-commerce tasks, including product copy generation [1], catalog search [2], product recommendation [3], fashion image generation [4] and retrieval [5, 6]. Accurately extracted attributes provide control for downstream tasks like product copy. Instead of directly generating copy from an image, attribute extraction allows controlled generation, covering selective attributes (depending on brand, product uniqueness) and reducing hallucination. As depicted in Figure 1, fashion attributes are arranged in an ontology, readily available with every retailer, where there are product category specific attribute types, which may take one or more values. For example, the sleeve length attribute type for product category 'blouse' may have values such as sleeveless, cup sleeves, short sleeves, etc. This makes, extraction of attributes from fashion images non-trivial given its fine-granular nature. The chosen approach has to

eCom'23: ACM SIGIR Workshop on eCommerce, July 27, 2023, Taipei, Taiwan

*Corresponding author.

shubham.p3@tcs.com (S. Paliwal); bhagyashree.gaikwad@tcs.com (B. Gaikwad); patidar.mayur@tcs.com (M. Patidar); manasi.patwardhan@tcs.com (M. Patwardhan); lovekesh.vig@tcs.com (L. Vig); meghna.mahajan@tcs.com (M. Mahajan); bagyalakshmi.v@tcs.com (B. L. V); shirish.karande@tcs.com (S. Karande)

learn to focus on an appropriate part of the image which depicts the position of attribute type such as Sleeve Length and then has to perform the hard-to-distinguish task of differentiating between fine-grained attribute values for that attribute type.

Most off-the-shelf SOTA approaches [7, 8, 9, 10, 11, 12, 13, 14, 15] treat attribute recognition as multi-stage, hierarchical, multi-label (or multi-class) classification task. Some of these approaches use multi-task learning by using product category classification, landmark and/or key-point detection as auxiliary task(s) to improve the performance. However, these approaches do not leverage the relatedness of attribute values, embedded in the attribute ontology. For example, an image representation (e.g. I1 in Figure 1) should be closer to the attribute value representations with which it is labelled (e.g. text prompt T1 for attribute value 'wrist length Sleeves'). Thus, indirectly bringing the image representations of two images having the same attribute value (e.g. 'wrist length sleeves') for an attribute type (e.g. 'Sleeve Length') closer and image representations farther when the two images hold distinct values (e.g. 'turtle neck' and 'Ruffle Semi-High Collar neck') for an attribute type (e.g. 'neck design'). More importantly, to embed the attribute relatedness depicted by the ontology, the image representation should be farther from the attribute values which are siblings (other values of the same attribute type) of the attribute value the image is annotated with. For example, image representation I1 in figure 1 should be farther to the text prompt representation T2 for attribute value 'log length sleeves', which is the sibling of (belongs to the same attribute type 'sleeve length') the attribute value 'wrist length sleeves', with which the I1 is labelled. Such attribute values belonging to same attribute type are hard to distinguish.

In this work, we model attribute extraction as a matching problem [16]. We fine-tune the pre-trained CLIP model, contrasting image-attribute representations. We address the aforementioned limitation of the prior work by a novel supervised contrastive learning-based training mechanism by leveraging fashion ontology to create hard negative pairs. [17] use contrastive learning with object-level supervision to align pre-trained language and vision model by increasing the difficulty of mini-batches over training epochs based on object level ontology. On the other hand, our approach exploits ontology in-built for fashion domain for more fine-grained task of fashion attribute extraction. The proposed approach outperforms existing SOTA approaches on two public datasets, viz. DeepFashion [18] and FashionAI [19] by 11.65% top-5 recall rate and 0.93 mAP, respectively. The main contributions of this work are:

- We have modeled the fine-granular multi-label fashion attribute classification as a matching problem to address the relatedness of attribute values embedded in the attribute ontology.
- We propose a novel ontology-guided supervised contrastive learning approach for fashion attribute extraction.
- The proposed approach outperforms existing baselines on two public datasets, viz. DeepFashion and FashionAI.

2. Problem Description

Fashion ontology ($\mathcal{O}$) [18, 20, 19] consists of fashion-related concepts (e.g., Product Category (PC), Attribute Type (AT), and Attribute Values (AV), etc.) which are arranged in the form of a

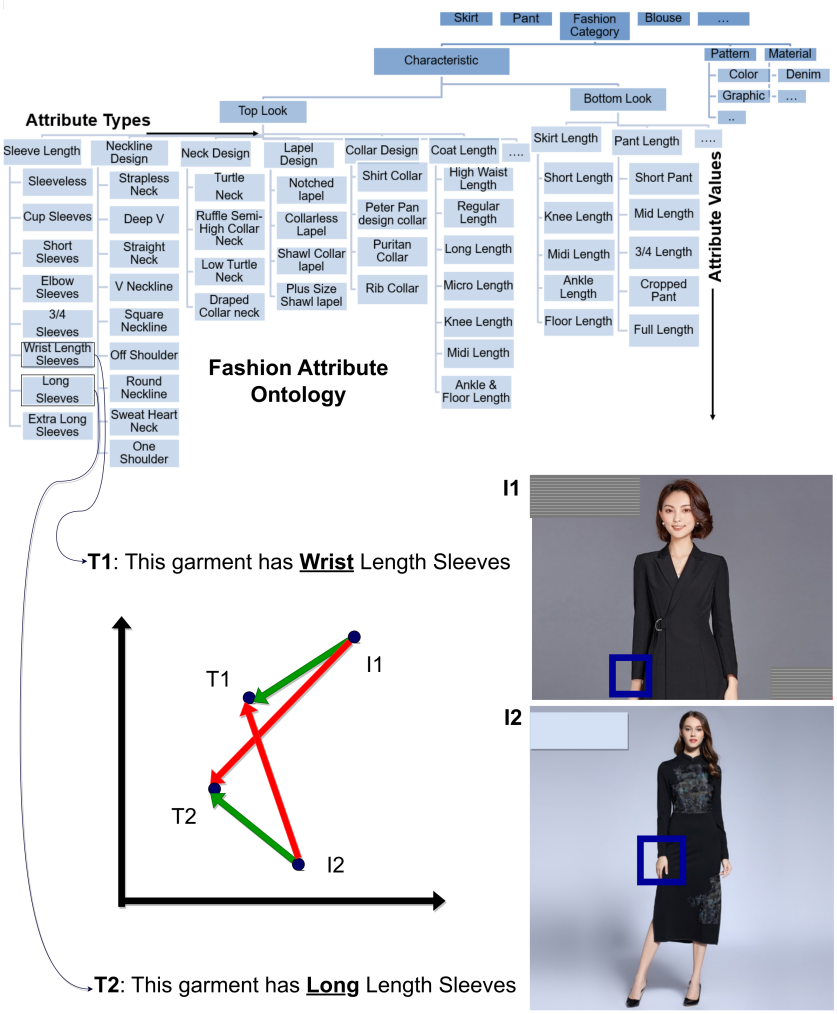

Figure 1: Motivation

hierarchy and are connected to each other via appropriate relationships. For e.g., 'Blouse' is an instance of a product category with Neck, Sleeve, Print, etc., as attribute types and turtle, draped collar, etc., are some of the attribute values of attribute type Neck (Figure 1). Fashion images are annotated w.r.t an ontology $\mathcal{O}$ by domain experts at all levels i.e., product category, valid attribute types, and corresponding attribute values.

Given a fashion ontology $\mathcal{O}$ and corresponding annotated dataset i.e., $\mathcal{D} = \{(I_1, A_1), ..., (I_n, A_n)\}$, where i^{th} image (I_i) is annotated with $A_i = \{PC_i^j, AT_j^l, ..., AT_j^m\}$, product category PC_i^j, valid attribute types $\{AT_j^l, ..., AT_j^m\}$ and corresponding valid attribute values $AT_j^l = \{AV_l^1, ..., AV_l^k\}$, objective is to automatically annotate test image w.r.t $\mathcal{O}$.

3. Proposed Approach

We model fine-grained fashion attribute extraction as a matching problem where we fine-tune CLIP [21] via supervised contrastive loss (SupCon) [22] by minimizing the cosine-similarity between the image and attribute representation. To handle the class imbalance we augment SupCon with asymmetric contrastive focal loss [23] during the training. During inference, we choose valid attribute types and values based on the cosine-similarity between the image and attribute representations.

3.1. Training

3.1.1. Supervised Contrastive Language Image Pre-training Fine-tuning (SCLIP-F)

Similar to CLIP, we obtain the multimodal image representation by first passing it to CLIP's image encoder (f^{IE}_{CLIP}) and then through a multimodal image projection layer (W_I) i.e., $I^e = W_I.f^{IE}_{CLIP}(I)$. We use a textual prompt (T) to verbalize the attribute value and get the corresponding multimodal representation via CLIP text encoder (f^{TE}_{CLIP}) and multimodal text projection layer (W_T) i.e., $T^e = W_T.f^{TE}_{CLIP}(T)$.

CLIP is pre-trained on (image, text) pairs by maximizing the cosine similarity between representations of $\mathscr{B}$ (image, text) pairs and minimizing the cosine similarity for $\mathscr{B}^2 - \mathscr{B}$ invalid pairs in a batch of size $\mathscr{B}$, as shown in Eq. 1.

$$\mathscr{L}_{CLIP} = -\frac{1}{\mathscr{B}}\sum_{i=1}^{\mathscr{B}} log\left[\frac{exp(\langle I_i^e, T_i^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle I_i^e, T_a^e\rangle/\tau)}\right] - \frac{1}{\mathscr{B}}\sum_{j=1}^{\mathscr{B}} log\left[\frac{exp(\langle I_j^e, T_j^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle I_a^e, T_j^e\rangle/\tau)}\right] \quad (1)$$

In a fashion domain, an image might share the attribute values with other images based on the category to which they belong or due to visual similarity among them. For example, a fashion image of shirt and blouse may both share attribute value 'rib collar' for attribute type 'collar design'. Unlike CLIP, all (image, attribute) pairs in $\mathscr{B}$, which share the same attribute values are referred to as positive pairs and others as negative. During the fine-tuning of CLIP, we maximize the cosine similarity between representation of positive (image, attributes) and minimize the cosine similarity between negative pairs, as shown in Eq. 2.

$$\mathscr{L}^{SupCon+}_{CLIP} = -... + -\frac{1}{\mathscr{B}}\sum_{i=1}^{\mathscr{B}}\frac{1}{|P(i)|}\sum_{p\in P(i)} log\left[\frac{exp(\langle I_i^e, T_p^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle I_i^e, T_a^e\rangle/\tau)}\right] \quad (2)$$

3.1.2. Ontology Guided Supervised Contrastive Learning

Class imbalance in a dataset makes it harder to learn good representation for rare classes via supervised contrastive learning, due to absence of positive pairs for low frequency attribute values in $\mathscr{B}$. To alleviate this issue, inspired by [23], we augment $\mathscr{L}^{SupCon}_{CLIP}$ with focal loss as

shown in Eq. 3 and directly minimize the cosine similarity among negative (image, attribute) pairs ($\mathscr{L}_{CLIP}^{SupCon-}$) as shown in Eq. 4.

$$\mathscr{L}_{CLIP}^{SupCon+} = -\frac{1}{\mathscr{B}}\sum_{i=1}^{\mathscr{B}}\frac{1}{|P(i)|}\sum_{p\epsilon P(i)}\left[1-\frac{exp(\langle I_i^e, T_p^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle I_i^e, T_a^e\rangle/\tau)}\right]^{\gamma} \frac{exp(\langle I_i^e, T_p^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle I_i^e, T_a^e\rangle/\tau)} - \frac{1}{\mathscr{B}}\sum_{j=1}^{\mathscr{B}}\frac{1}{|P(j)|} \sum_{p\epsilon P(j)}\left[1-\frac{exp(\langle T_i^e, I_p^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle T_i^e, I_a^e\rangle/\tau)}\right]^{\gamma}\frac{exp(\langle T_i^e, I_p^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle T_i^e, I_a^e\rangle/\tau)} \quad (3)$$

$$\mathscr{L}_{CLIP}^{SupCon-} = -\frac{1}{\mathscr{B}}\sum_{i=1}^{\mathscr{B}}\frac{1}{|N(i)|}\sum_{n\epsilon N(i)} log\left[1-\frac{exp(\langle I_i^e, T_n^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle I_i^e, T_a^e\rangle/\tau)}\right] - \frac{1}{\mathscr{B}}\sum_{j=1}^{\mathscr{B}}\frac{1}{|N(j)|}\sum_{n\epsilon N(j)} log\left[1-\frac{exp(\langle I_n^e, T_j^e\rangle/\tau)}{\sum_{a=1}^{\mathscr{B}} exp(\langle I_a^e, T_j^e\rangle/\tau)}\right] \quad (4)$$

It has been shown that hard-negatives are key to get better representations via contrastive learning [24]. For selecting in-batch hard-negatives (image, attribute) pairs, for fine-grained attribute classification, we use the fashion ontology. For a given (image, attribute) pair in batch $\mathscr{B}$, we treat all those (image,attribute) pairs as hard-negatives, which are siblings of each other in the fashion ontology. As shown in Fig 1, text prompt T1 from attribute value"wrist length Sleeves" is a hard negative pair for image I2 with attribute value "Long length Sleeves", if they appear in the same batch $\mathscr{B}$.

$$\mathscr{L}_{OGSCL} = \mathscr{L}_{CLIP}^{SupCon+} + \eta\mathscr{L}_{CLIP}^{SupCon-} \quad (5)$$

To improve the robustness of $OGSCL - FAE$, we also perform data augmentation by applying attribute-invariant transformations [25] over images present in $\mathscr{D}$.

3.1.3. Training

We fine-tune pre-trained CLIP on a feature extraction dataset $\mathscr{D}$ by minimizing ontology-guided supervised contrastive focal loss as shown in Eq. 5.

3.2. Inference

Given a test image (I_{test}), we obtain its multimodal representation (I_{test}^e) via fine-tuned image encoder f_{OGSCL}^{IE} i.e., $I_{test}^e = W_I \cdot f_{OGSCL}^{IE}$
(I_{test}). In order to predict the product category, we calculate the cosine similarity between I_{test}^e and the multimodal representation of the textual prompt corresponding to each product category present in $\mathscr{O}$ and choose the one (PC_i) with maximum cosine similarity, $\underset{i}{\text{argmax}}\langle I_{test}^e, T_i^e\rangle$. For

attribute prediction, we calculate cosine similarity between I^e_{test} and the multimodal representation of the textual prompt corresponding to each attribute value for an attribute type (applicable to that product category) and choose the one (AV_i) with maximum cosine similarity, we repeat this for all attribute types independently.

Table 1
Comparison of different approaches on DeepFashion for attribute classification using recall-rate@k.

Approach	Attributes									
	Texture		Fabric		Shape		Part		Style	
	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5	Top-3	Top-5
WTBI	24.21	32.65	25.38	36.06	23.39	31.26	26.31	33.24	49.85	58.68
DARN	36.15	48.15	36.64	48.52	35.89	46.93	39.17	50.17	66.11	71.36
FashionNet	37.46	49.52	39.30	49.84	39.47	48.59	44.13	54.02	66.43	73.16
BCRNN	50.31	65.48	40.31	48.23	53.32	61.05	40.65	56.32	**68.70**	**74.25**
TS-FashionNet	58.52	68.19	46.44	57.02	61.86	70.81	49.82	60.36	34.40	43.44
HABP	**60.87**	**70.54**	49.40	59.88	61.97	70.80	51.39	61.82	38.61	46.99
MLC	53.81	61.76	42.61	52.59	39.82	49.62	25.43	38.14	47.60	48.31
CLIP-P	27.73	34.43	11.75	19.42	31.66	44.94	9.983	16.03	41.51	45.57
SCLIP-F	46.24	53.92	46.39	55.09	62.92	74.98	45.08	53.68	57.01	63.41
OGSCL-FAE	52.42	58.76	**52.71**	**62.43**	**68.99**	**77.20**	**56.33**	**63.93**	65.90	70.73

Table 2
Comparison of different approaches on DeepFashion for Overall Performance using recall-rate@k.

Approach	Overall	
	Top-3	Top-5
DARN	42.35	51.95
FashionNet	45.52	54.61
TS-FashionNet	50.58	60.43
HABP	52.82	62.49
TwoStreamMN	59.83	77.91
MTL w/ RNN+VA	53.01	66.40
STL w/ HLS	66.19	73.73
MLC	65.57	71.22
CLIP-P	64.50	71.33
SCLIP-F	81.27	85.42
OGSCL-FAE	**86.31**	**91.22**

4. Experimental Setup

4.1. Dataset Details

DeepFashion [18]: It consists of 289,222 fashion images (Train: 209,222, Validation: 40,000 and Test: 40,000), each belonging to one of 50 different categories and annotated w.r.t the ontology consists of 5 attribute types and 1000 attribute values.
FashionAI [19]: It consists of 180,335 fashion images (Train: 144,335, Validation: 18,000 and Test: 18,000) which belong to 6 different categories and are annotated w.r.t the ontology consists

Table 3
Performance of different approaches on FashionAI using mAP

Approach	Length				Design				Overall
	Skirt	Sleeve	Coat	Pant	Collar	Lapel	Neckline	Neck	
Triplet Network	48.38	28.14	29.82	54.56	62.58	38.31	26.64	40.02	38.52
CSN	61.97	45.06	47.30	62.85	69.83	54.14	46.56	54.47	53.52
ASEN	66.34	57.53	55.51	68.77	72.94	66.95	66.81	67.01	64.31
CAMNET	**68.23**	58.08	60.86	68.74	**78.32**	**73.63**	65.96	68.02	66.37
MLC	66.07	50.81	51.57	68.33	78.17	66.77	46.64	66.16	59.20
CLIP-P	19.45	14.71	13.46	23.43	24.68	20.40	12.10	20.51	17.52
SCLIP-F	63.36	54.72	57.34	64.65	64.75	61.48	60.84	56.98	60.20
OGSCL-FAE	66.30	**63.48**	**63.86**	**69.25**	72.34	71.29	**67.55**	**69.26**	**67.30**

Table 4
Ablation Study for effectiveness of OGSCL-FAE components on DeepFashion

Approach	Overall	
	Top-3	Top-5
OGSCL-FAE	86.31	91.22
OGSCL-FAE w/o DA & OGSCL	80.24	84.50
OGSCL-FAE w/o OGSCL	82.83	87.75
OGSCL-FAE w/o DA	80.67	85.98
OGSCL-FAE w/o OG	83.36	88.06

of 8 design-specific attribute types and 54 attribute values.

4.2. Baselines

Approaches such as CSN[26], ASEN[27], DARN[28], CAMNet [29] are designed for fashion image retrieval by learning fine-grained attribute specific embedding for fashion images with metric learning. In our approach, instead we take attribute relatedness into consideration for learning image representation by ontology guided training using contrastive setting. WTBI [30], FashionNet [18], BCRNNs[31], TS-FashionNet[32], STL w/ HLS, MTL w/ (RNN + VA)[33] and TwoStreamMN[8] treats the attribute extraction as a multi-class classification task and takes help of auxiliary task(s) such as pose estimation, landmark prediction, category identification and/or object type detection by either jointly learning the model or following a staged pipeline, leading to improvement in the performance of the attribute extraction. As opposed to these approaches, instead of multi-class classification, we treat the attribute extraction task as a matching problem. HABP [34] addresses the problem of class imbalance for fashion attribute extraction, by adaptively focusing on training hard data (attributes with very less tagged samples) followed by a method to synthesize complementary samples for such hard attributes. In our approach, we take care of the class imbalance by using focal loss, data augmentation and ontology guided hard negative sampling.

Contrastive Language-Image Pre-Training- Pretrained (CLIP-P) [21] is our baseline, where we use the pre-trained version of the CLIP model without any task specific fine-tuning. Whereas, Supervised Contrastive Language-Image Pre-Training Finetuned (SCLIP-F) [21] is where we perform task specific finetuning of CLIP for domain adaptation. Multilabel Classification (MLC) is where we use the same base model, which is used as the image encoder in the CLIP setting and fine-tune it for multi-label attribute classification.

4.3. Training Details

We use pre-trained ViT/B-16 as our CLIP image encoder implemented in Pytorch. For all of our experiments, the models are trained on an Nvidia A-100, using batch size of 96 and the learning rate of 3e-6. For the MLC baseline, we have appended linear layers of size 512, 1024 and the dimension of attribute classes to the end of the pre-trained image encoder, and fine-tuned using asymmetric focal loss [35]. For Deepfashion we use a sigmoid activation layer per attribute, while for FashionAI we use grouped (as per attribute type) softmax activation distributed over attribute values. We use validation set assistance in training, in which for each epoch, the negative pair sampling frequency is set in proportion to the non-diagonal validation set confusion matrix values, for each attribute pair, helping in better distinguishing confusing attribute pairs.

4.4. Evaluation

4.4.1. Top-k Recall

For a given attribute type, it refers to the fraction of test images for which the true attribute value is present in the top-k predicted attribute values. Also, for a dataset, Top-k recall is the mean of Top-K recall for each attribute type. To compute this metric We use the official code[1] provided by authors of [18].

4.4.2. Mean Average Precision (mAP)

For a given attribute type, it refers to the fraction of test images for which predicted attribute value matches with the ground truth. And for a dataset, mAP is the mean of mAP for each attribute type.

5. Results And Discussion

Pre-training Vs Fine-tuning for fine-grained fashion attribution extraction: as shown in Table 1, 2 and 3, for both the datasets, SCLIP-F outperforms CLIP by a significant margin, suggesting the need for fine-tuning pre-trained CLIP for attribute extraction in fashion domain.

Multilabel classification vs Matching: As depicted in Table 1, 2 OGSCL-FAE outperforms MLC on DeepFashion by 20.74% and 20% in terms of Top-3 and Top-5 recall, respectively. Similarly, in Table 3, it also outperforms MLC on FashionAI by 8.1% in terms of mAP. This suggests that fashion ontology is a key component to achieve better performance on fine-grained attribute classification. During the training contrasting an attribute value with all its sibling (OGSCL-FAE) is more important as compared to maximizing the likelihood of an attribute value in isolation (MLC).

OGSCL-FAE vs. Baselines In terms of overall performance, OGSCL-FAE outperforms the best baseline SCLIP-F by 5.04% (Top-3) and 5.8% (Top-5) on DeepFashion and baseline CAMNET by 0.93% mAP on FashionAI. We use the best-performing variant of CAMNET as a baseline

[1] attr_predict_eval.py @ https://github.com/open-mmlab/mmfashion/

where HRNet [36] with two-step attention layers is used as the backbone as opposed to ViT-B/16 [37] in the proposed approach. But still, OGSCL-FAE outperforms the best baseline i.e. CAMNET on FashionAI, in terms of overall mAP (Table 3) for 5 out of 8 attribute types. Since the code for CAMNET is not publicly available, it's not possible to test CAMNET with ViT-B/16 as a backbone. For DeepFashion, except for Style and Texture, OGSCL-FAE outperforms all baselines for all attribute types.

Discussion about ablations Data augmentations and ontology-guided supervised contrastive learning are key components of OGSCL-FAE because there is a drop in performance of 6.07% (Top-3) and 6.72% (Top-5), as shown in Table 4 (**OGSCL-FAE** w/o DA & OGSCL). CLIP fine-tuning with self-supervised contrastive loss and data augmentation performs very poorly as compared to supervised contrastive loss with data augmentation (**OGSCL-FAE** w/o OGSCL). Ontology guided negative sampling over random sampling improves performance of OGSCL-FAE by 2.95% (Top-3) and 3.16% (Top-5) (**OGSCL-FAE** w/o OG). Data augmentation also affects the overall performance of OGSCL-FAE by 5.64% (Top-3) and 5.24% (Top-5) (**OGSCL-FAE** w/o DA).

6. Conclusion

This paper proposes a novel approach to fine-granular fashion attribute extraction by exploiting an ontology-guided negative sampling strategy for supervised contrastive learning of pre-trained CLIP. The proposed method outperforms existing state-of-the-art results on DeepFashion and FashionAI datasets, achieving 11.65% top-5 recall rate and 0.93 mAP respectively. Future work will include using the attribute extraction module for attribute-guided product copy generation.

References

[1] X. Guo, Q. Zeng, M. Jiang, Y. Xiao, B. Long, L. Wu, Automatic controllable product copywriting for e-commerce, Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2022).

[2] X. Yang, X. Song, X. Han, H. Wen, J. Nie, L. Nie, Generative attribute manipulation scheme for flexible fashion search, Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020).

[3] L. D. Divitiis, F. Becattini, C. Baecchi, A. Bimbo, Disentangling features for fashion recommendation, ACM Transactions on Multimedia Computing, Communications and Applications 19 (2022) 1 – 21.

[4] Z. Azizi, C.-C. J. Kuo, Pager: Progressive attribute-guided extendable robust image generation, ArXiv abs/2206.00162 (2022).

[5] L. Liao, X. He, B. Zhao, C.-W. Ngo, T.-S. Chua, Interpretable multimodal retrieval for fashion products, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1571–1579.

[6] C. Yan, K. Yan, Y. Zhang, Y. Wan, D. Zhu, Attribute-guided fashion image retrieval by iterative similarity learning, 2022 IEEE International Conference on Multimedia and Expo (ICME) (2022) 1–6.

[7] S. Papadopoulos, C. Koutlis, M. Sudheer, M. Pugliese, D. Rabiller, S. Papadopoulos, I. Kompatsiaris, Attentive hierarchical label sharing for enhanced garment and attribute classification of fashion imagery, 2021.
[8] P. Li, Y. Li, X. Jiang, X. Zhen, Two-stream multi-task network for fashion recognition, 2019 IEEE International Conference on Image Processing (ICIP) (2019) 3038–3042.
[9] M. Shajini, A. Ramanan, Multi-staged feature-attentive network for fashion clothing classification and attribute prediction, ELCVIA Electronic Letters on Computer Vision and Image Analysis (2022).
[10] H. S. Arslan, K. Sirts, M. Fishel, G. Anbarjafari, Multimodal sequential fashion attribute prediction, Inf. 10 (2019) 308.
[11] M. Shin, Semi-supervised learning with a teacher-student network for generalized attribute prediction, ArXiv abs/2007.06769 (2020).
[12] V. Parekh, K. Shaik, S. Biswas, M. Chelliah, Fine-grained visual attribute extraction from fashion wear, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2021) 3968–3972.
[13] B. Kolisnik, I. Hogan, F. H. Zulkernine, Condition-cnn: A hierarchical multi-label fashion image classification model, Expert Syst. Appl. 182 (2021) 115195.
[14] H. Cho, C. Ahn, K. M. Yoo, J. Seol, S. goo Lee, Leveraging class hierarchy in fashion classification, 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) (2019) 3197–3200.
[15] J. Dong, Z. Ma, X. Mao, X. Yang, Y. He, R. Hong, S. Ji, Fine-grained fashion similarity prediction by attribute-specific embedding learning, IEEE Transactions on Image Processing 30 (2021) 8410–8425.
[16] K. Goei, M. Hendriksen, M. de Rijke, et al., Tackling attribute fine-grainedness in cross-modal fashion search with multi-level features, in: SIGIR 2021 Workshop on eCommerce. ACM, 2021.
[17] T. Srinivasan, X. Ren, J. Thomason, Curriculum learning for data-efficient vision-language alignment, arXiv preprint arXiv:2207.14525 (2022).
[18] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[19] X. Zou, X. Kong, W. K. Wong, C. Wang, Y. Liu, Y. Cao, Fashionai: A hierarchical dataset for fashion understanding, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019) 296–304.
[20] M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, S. Belongie, Fashionpedia: Ontology, segmentation, and an attribute localization dataset, in: European Conference on Computer Vision (ECCV), 2020.
[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.
[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan,

Supervised contrastive learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 18661–18673. URL: https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

[23] V. Vito, L. Y. Stefanus, An asymmetric contrastive loss for handling imbalanced datasets, Entropy 24 (2022).

[24] J. Robinson, C.-Y. Chuang, S. Sra, S. Jegelka, Contrastive learning with hard negative samples, International Conference on Learning Representations (2021).

[25] S. G. Müller, F. Hutter, Trivialaugment: Tuning-free yet state-of-the-art data augmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 774–782.

[26] A. Veit, S. Belongie, T. Karaletsos, Conditional similarity networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[27] J. Dong, Z. Ma, X. Mao, X. Yang, Y. He, R. Hong, S. Ji, Fine-grained fashion similarity prediction by attribute-specific embedding learning, IEEE Transactions on Image Processing 30 (2021) 8410–8425. doi:10.1109/TIP.2021.3115658.

[28] J. Huang, R. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1062–1070. doi:10.1109/ICCV.2015.127.

[29] C. H. Song, H. Joo Han, Convolutional attribute mask with two-step attention for fashion image retrieval, in: 2022 26th International Conference on Pattern Recognition (ICPR), 2022, pp. 2093–2099. doi:10.1109/ICPR56361.2022.9955640.

[30] H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: Proceedings of the 12th European Conference on Computer Vision - Volume Part III, ECCV'12, Springer-Verlag, Berlin, Heidelberg, 2012, p. 609–623. URL: https://doi.org/10.1007/978-3-642-33712-3_44. doi:10.1007/978-3-642-33712-3_44.

[31] W. Wang, W. Wang, Y. Xu, J. Shen, S.-C. Zhu, Attentive fashion grammar network for fashion landmark detection and clothing category classification, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4271–4280. doi:10.1109/CVPR.2018.00449.

[32] Y. Zhang, P. Zhang, C. Yuan, Z. Wang, Texture and shape biased two-stream networks for clothing classification and attribute recognition, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13535–13544. doi:10.1109/CVPR42600.2020.01355.

[33] S.-I. Papadopoulos, C. Koutlis, M. Sudheer, M. Pugliese, D. Rabiller, S. Papadopoulos, I. Kompatsiaris, Attentive hierarchical label sharing for enhanced garment and attribute classification of fashion imagery, in: N. Dokoohaki, S. Jaradat, H. J. Corona Pampín, R. Shirvany (Eds.), Recommender Systems in Fashion and Retail, Springer International Publishing, Cham, 2022, pp. 95–115.

[34] Y. Ye, Y. Li, B. Wu, W. Zhang, L. Duan, T. Mei, Hard-aware fashion attribute classification, arXiv preprint arXiv:1907.10839 (2019).

[35] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 82–91.

[36] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ICLR (2021).