Supplemental Material of
Identifying Ambiguous Similarity Conditions via Semantic Matching

Han-Jia Ye Yi Shi De-Chuan Zhan
State Key Laboratory for Novel Software Technology, Nanjing University
{yehj, shiy, zhandc}@lamda.nju.edu.cn

Abstract

There are four parts in this supplementary:

- *The details to reproduce Fig. 1 (bottom left) in the main paper and the details of our proposed criterion.*
- *A probabilistic view of* DISCOVERNET *which decomposes the ambiguous distance between objects.*
- *Detailed configurations of the experiments and implementation details.*
- *Additional experimental results to show the superiority of the proposed* DISCOVERNET.

1. Details of Our Proposed Criterion

1.1. Supervised CSL

A supervised CSL model learns multiple embedding Ψ_K, revealing the specific characteristics of each similarity condition.

Given Ψ_K, we can determine the validness of a triplet. Specifically, given a (supervised) triplet $\tau = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)$ from the k-th condition, we use the corresponding embedding ψ_k and compute the value

$$\begin{aligned}\mathbf{Diff}_\tau^k &= \mathbf{Dis}^2_{L_k}(\phi(\mathbf{x}),\ \phi(\mathbf{z})) - \mathbf{Dis}^2_{L_k}(\phi(\mathbf{x}),\ \phi(\mathbf{y}))\\ &= \|\psi_k(\mathbf{x}) - \psi_k(\mathbf{z})\|_2^2 - \|\psi_k(\mathbf{x}) - \psi_k(\mathbf{y})\|_2^2\ . \quad (1)\end{aligned}$$

We use $\mathbf{Diff}_\tau^k$ to predict whether a triplet is valid or not. if $\mathbf{x}$ (the anchor) and $\mathbf{y}$ (the target neighbor) are more similar than $\mathbf{x}$ (the anchor) and $\mathbf{z}$ (the impostor), *i.e.*, the distance between $\mathbf{x}$ and $\mathbf{y}$ based on ψ_k is smaller than the distance between $\mathbf{x}$ and $\mathbf{z}$, then we have $\mathbf{Diff}_\tau^k > 0$.

Evaluation Criterion of Supervised CSL. To evaluate a supervised CSL model, we transform the task into measuring the prediction ability of multiple embeddings, *i.e.*, we use the learned Ψ_K to determine whether a given triplet is valid or not. In detail, we sample the same number of *valid* triplets from each condition during evaluation. Then for each condition k, we use ψ_k to predict whether triplets from that condition are correct or not based on the value of $\mathbf{Diff}_\tau^k$. The average accuracy (proportion of triplets predicted as correct) over them is used as the final criterion, which reveals the quality of multiple CSL embeddings.

Denote $\mathbf{Map}[k]$ as a mapping from ground-truth condition k to a particular learned embedding $\psi_{\mathbf{Map}[k]}$ in Ψ_K. In the supervised scenario, we learn the same number of conditional embeddings with the number of ground-truth conditions, and we set $\mathbf{Map}[k] = k$ as an identity mapping. The evaluation steps are listed in Algorithm 1.

It is notable that given a particular ψ_k, we have either $\mathbf{Diff}_\tau^k > 0$ or $\mathbf{Diff}_\tau^k \leq 0$, so one triplet $\tau_1 = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)$ and its reversed version $\tau_2 = (\mathbf{x}, \mathbf{z}, \mathbf{y}, k)$ could not exist simultaneously under the same condition. Therefore, only using the *valid* triplets during evaluation (all of them with the "correct" label) produces reasonable results.

Algorithm 1 Evaluation of Supervised CSL

```
Require: Triplets $\{\tau_i\}_{i=1}^T$, the learned embeddings $\Psi_K$,
    the condition-embedding mapping $\mathbf{Map}[k]$
Initialize: $t = 0$.
for $i = 1 \to T$ do
    get $\tau_i = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)$ with ground-truth condition $k$
    compute $\mathbf{Diff}_{\tau_i}^k$ with $\psi_{\mathbf{Map}[k]}$ and Eq. 1
    if $\mathbf{Diff}_{\tau_i}^k > 0$ then
        $t = t + 1$
    end if
end for
return $Acc = t/T$
```

Discussions of the Supervised CSL Criterion. One direct question of the criterion is *whether it could reveal all conditional similarities or not.* Due to the fact that a triplet τ_1 and its reversed version τ_2 could not be satisfied by the same conditional embedding ψ_k, the evaluation of those valid triplets excludes the relationship revealed by the reversed triplets and characterize a particular similarity condition. The learned conditional embeddings could achieve high accuracy only when they characterize the latent relationship of all target conditions.

1.2. Weakly Supervised CSL

The current WS-CSL evaluation utilizes the same criterion as the supervised CSL [6, 10] — predicting the correctness of a triplet *without using the test-time condition labels*. We first demonstrate the drawback of the current criterion via a synthetic experiment.

Synthetic Experiment Setups. We investigate various methods on UT-Zappos-50k dataset [14, 15], where we extract 40,000 triplets from four conditions. We use conditional labels for supervised methods, and neglect them for WS-CSL methods during evaluation. There are various methods in our synthetic experiment.

- Optimal. We show the optimal *Supervised* CSL method as a reference. Which achieves the highest performance with the supervised evaluation criterion.

- CSN [12]. CSN learns multiple embedding masks during training in a fully *supervised* way, and applies the corresponding mask over the backbone as a special conditional measure. Note that the condition labels are required in both training and test processes.

- LSN [6]. LSN is a multi-attribute model which discovers latent attributes by choosing the one with the minimum loss in a *Weakly Supervised* way. Following the multi-choice learning [2], it learns disentangled embeddings Ψ_K with hard condition assignments.

- SCE-Net [10]. SCE-Net is a *Weakly Supervised* method, which consists of an attention module and multiple embeddings Ψ_K. Given a triplet, SCE-Net generates a weight vector over multiple conditions, with which it fuses multiple embeddings Ψ_K to predict the correctness of a triplet.

- DISCOVERNET. Our proposed *Weakly Supervised* CSL method with semantic regularization. Details can be found in the main paper.

We introduce another task to predict the correctness of reversed triplets, *i.e.*, we transform $\tau_1 = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)$ to $\tau_2 = (\mathbf{x}, \mathbf{z}, \mathbf{y}, k)$ by exchanging the position of last two instances in the triplets. So in a supervised evaluation, the ground-truth of all reversed triplets must be false, and a supervised method should predict them as correct ones as fewer as possible (the lower the proportions of triplets be predicted as valid, the higher the accuracy). However, in a WS-CSL evaluation, we do not use the condition label k, so a method predicts the correctness of $\tau_3 = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ and $\tau_4 = (\mathbf{x}, \mathbf{z}, \mathbf{y})$ in the original and reversed cases directly.

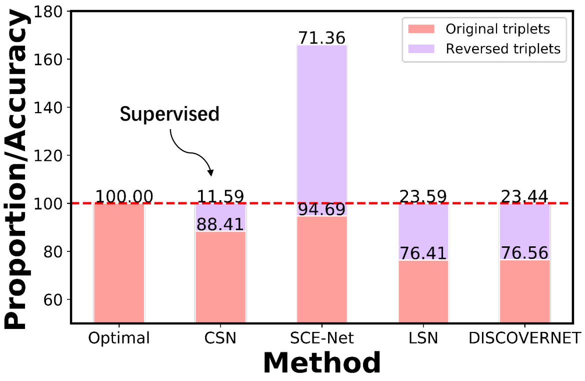


Figure 1. Given original (correct) triplets and their reversed variants on UT-Zappos-50k, we compute the proportion a model that predicts them as valid ones. The higher the proportion of *original* triplets be predicted as correct ones, the higher the accuracy. In contrast, the higher the proportion of *reversed* triplets be predicted as correct ones, the lower the accuracy. The first two are supervised CSL methods, and the last three are WS-CSL methods.

Analysis of the Experiment. The results are shown in Fig. 1. We show the proportion of original and reversed triplets with red and blue respectively, and accumulate their values together. For the optimal supervised model at the leftmost, it predicts all original triplets as right ones and all reversed triplets as invalid (achieves 100% accuracy in both cases). For the supervised CSN, it has high proportion (accuracy) on the original triplets and low proportion (also high accuracy) on the reversed ones.

The phenomenon is different for WS-CSL methods, especially for SCE-Net. SCE-Net fuses all embeddings with attention for each triplet. The results show SCE-Net gets much higher accuracy on the original triplets (even better than the supervised CSN), but also predicts a lot of reversed triplets as correct ones. Fig. 1 indicates that SCE-Net tends to treat most original and reversed triplets as right ones, which is different from the supervised case.

The main reason is that the original and reversed triplets could possess different conditions, so a WS-CSL method explains them from two diverse aspects. Using the supervised criterion (*i.e.*, the red part) could be *biased* in this case. For example, SCE-Net is able to learn good embeddings and powerful attentions, but the current criterion demonstrates that it prefers valid triplets, which is explained by combined embeddings. Thus, we are unaware of whether we learn meaningful embeddings or a strong fusion module.

Our Proposed WS-CSL Criterion. Based on our analysis, the supervised criterion is able to evaluate the quality of all learned embeddings, while the WS-CSL's criterion may fall into the scenario using all conditional embeddings to explain the triplets. We follow an intuitive way to measure a WS-CSL model — whether the learned WS-CSL model performs similarly to the supervised CSL model. Then, not only the fusion of conditional embeddings Ψ_K should cover all target semantics, but also the behavior of each ψ_k reveals the relationship w.r.t. a specific condition.

We propose a new evaluation criterion (details could be found in the main paper). Using the conditional labels during evaluation, we find an alignment $\mathbf{Map}[k]$ between a particular condition and one of the learned embeddings in Ψ_K. The condition labels are only utilized to choose a good alignment. Then we can use the supervised criterion for a better evaluation. The details to evaluate with our criterion are listed in Algorithm 2.

Algorithm 2 Our evaluation steps for WL-CSL methods

```
Require: Learned embeddings $\Psi_K$, the number of ground-truth condition $K'$, the number of triplets of different ground-truth conditions $\{N_{k'}\}_{k'=1}^{K'}$.
I. Compute condition-embedding cost $C$.
for $k' = 1 \rightarrow K'$ do
    for $k = 1 \rightarrow K$ do
        Initialize: $t = 0$.
        get $\tau_i = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k')$
        compute $\mathbf{Diff}_{\tau_i}^k$ with $\psi_k$ and Eq. 1
        if $\mathbf{Diff}_{\tau_i}^k > 0$ then
            $t = t + 1$
        end if
        $Acc_{k'k} = t / N_{k'}$
    end for
end for
$C_{k'k} = 1 - Acc_{k'k}$
II. Match a condition $k'$ with embedding $\psi_k$.
i. Greedy Alignment:
for $k' = 1 \rightarrow K'$ do
    $\mathbf{Map}[k'] = \arg\min_k C_{k'k}$
end for
ii. OT Alignment:
compute transportation matrix T via
$\min_{T \geq 0} \langle T, C \rangle$   s.t.   $T\mathbf{1} = \frac{1}{K}\mathbf{1},\ T^\top \mathbf{1} = \frac{1}{K}\mathbf{1}$
for $k' = 1 \rightarrow K'$ do
    $\mathbf{Map}[k'] = \arg\max_k T_{k'k}$
end for
III. Compute accuracy with Alg. 1 and $\mathbf{Map}[k]$.
```

Discussions. A natural question for obtaining the conditional alignment is why we use OT instead of the Hungarian method. OT is a more *general* method to solve a matching problem. OT can deal with the case when there exists a number mismatch between the conditional embeddings and ground-truth conditions. We can show that when we use uniform marginal distributions and a square cost matrix in OT, the transportation will degenerate to the same solution as Hungarian's output with special conditions [7].

2. Probabilistic View of DISCOVERNET

In Weakly Supervised Conditional Similarity Learning (WS-CSL), only triplets $\{\tau = (\mathbf{x},\ \mathbf{y},\ \mathbf{z})\}$ from multiple conditions are provided, and the condition label k in each triplet is *unknown*. This is the usual case that we can explicitly describe the similarity of two images but difficult to point out from which aspect we measure them clearly. The model needs to infer the right condition label and learn discriminative embeddings to capture objects' characteristics simultaneously. Here we provide a detailed description of the probabilistic view of the basic version of DISCOVERNET. The final DISCOVERNET is equipped with a set module or a semantic regularizer.

DISCOVERNET characterizes both the *instance-instance* and *triplets-condition* relations in a "decompose-and-fuse" manner, which could be interpreted in a probabilistic aspect. With a bit abuse of the notation, we use $\tau = 1$ to represent the comparison relationship in the triplet is true, and $\tau = 0$ otherwise. With the help of the latent variable c_τ, we can measure the validness of a triplet τ via a holistic consideration of all similarity conditions:

$$\Pr(\tau = 1) = \prod_{k=1}^{K} \Pr(\tau = 1 | c_\tau^k = 1)^{c_\tau^k} . \tag{2}$$

$c_\tau \in \{0,1\}^K$ is a multinomial random variable, and we use $c_\tau^k = 1$ to denote the k-th component is selected. $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function, which squashes a variable into the range [0,1]. In Eq. 2, whether the relationship in the triplet is true or not depends both on the activated latent condition c_τ^k and the probability of the triplet in that view $\Pr(\tau = 1 | c_\tau^k = 1)$. Thus, the probability is related to both the concept prior $c_\tau^k = 1$ (the "triplets-condition" relationship) and the validness of the triplet conditioned on a particular concept $\Pr(\tau = 1 | c_\tau^k = 1)$ (the "instance-instance" relationship). The influence of all similarity conditions is aggregated together to determine the triplet in expectation.

Based on the projection L_k, we can define the probability that a given triplet is true based on their distance difference in the specific instance-instance embedding space

$$\Pr(\tau = 1 | c_\tau^k = 1) = \sigma\left(\mathbf{Diff}_\tau^k - \gamma\right) . \tag{3}$$

where the distance difference $\mathbf{Diff}_\tau^k$ is defined in Eq. 1. If the distance between $\mathbf{x}$ and $\mathbf{y}$ in this embedding space based on ϕ is smaller than the distance between $\mathbf{x}$ and $\mathbf{z}$, the input to σ is large such that the triplet will have a large probability to be valid. $\gamma > 0$ is a threshold. By subtracting γ from $\mathbf{Diff}_\tau^k$, we require the distance with the impostor should not

Table 1. The detailed configurations of conditions in our synthesized Celeb-A†. We synthesize 5 attributes over Celeb-A via combining similar binary attributes together, so each condition has 5-7 possible values.

Combined attributes	# of values	Original attributes included
Hair color	5	black-hair,blond-hair,brown-hair,gray-hair
Hair type	7	bangs,receding-hairline,straight-hair,wavy-hair
Eye and eyebone	6	arched-eyebrows,bags-under-eyes,bushy-eyebrows,narrow-eyes
Accessories	6	wearing-earrings,wearing-hat,wearing-necklace,wearing-necktie
Nose and mouth	7	big-lips,big-nose,mouth-slightly-open,pointy-nose

only be large but also larger than the distance with the target neighbor plus a margin.

For all given triplets $\mathcal{T}$, we optimize the embedding by maximizing the log-likelihood:

$$\mathcal{O} = \sum_{\tau \in \mathcal{T}} \log \Pr(\tau = 1) \ , \tag{4}$$

which has a similar form of the large margin loss [8,9] when only the general projection L and ϕ is used. The overall objective which minimizes the negative log likelihood of the joint probability over all triplets is:

$$\begin{aligned} \mathcal{O} = & -\sum_{\tau \in \mathcal{T}} \log \prod_{k=1}^{K} \sigma(\mathbf{Diff}_\tau^k - \gamma)^{c_\tau^k} \\ = & -\sum_{\tau \in \mathcal{T}} \sum_{k=1}^{K} c_\tau^k \log \sigma(\mathbf{Diff}_\tau^k - \gamma) \\ = & -\sum_{\tau \in \mathcal{T}} \mathbb{E}_{c_\tau} \log \sigma(\mathbf{Diff}_\tau^k - \gamma) \\ = & \sum_{\tau \in \mathcal{T}} \mathbb{E}_{c_\tau} \ell(\mathbf{Diff}_\tau^k - \gamma) \\ \approx & \sum_{\tau \in \mathcal{T}} \ell\left(\mathbb{E}_{c_\tau}\left[\mathbf{Diff}_\tau^k\right] - \gamma\right) \ . \end{aligned} \tag{5}$$

Here $\ell(x) = \log(1 + \exp(-x))$ is the logistic loss function, which can be replaced by other general losses. The approximation transforms the expectation over the loss function to the *expected distance over the conditions*, which fuses and reveals the preference over multiple similarity conditions. Therefore, the optimization in Eq. 5 requires the relationship in the selected instance-wise metric space to reveal the corresponding similarity condition, which makes the distance between the anchor and the impostor larger than the distance between the anchor and the target neighbor.

3. Experimental Setups

We describe the dataset, the comparison methods, and the implementation details in this section.

3.1. Datasets

Celeb-A† is a more complicated version of Celeb-A [5] via combining similar binary attributes in Celeb-A together. In particular, there are 202,599 face images from different identities in Celeb-A, and 40 binary visual attributes for each image, *e.g.*, "Eyeglasses" or "Wearing Hat". Each attribute corresponds to a condition, and we sample triplets randomly based on the binary values for each condition. To increase the difficulty of the dataset, we combine related binary attributes in Celeb-A together. The attributes related to "Hair color", "Hair type", "Eye and eye-bone", "Accessories", "Nose and mouth" are merged, and each of them has 5-7 possible discrete values. Details can be found in Table 1. Thus, different from the vanilla version with 40 binary attributes, the smaller number of attributes in the synthesized new dataset **Celeb-A**† has multi-choice conditions. We apply the same model configurations (such as the learning rate and the architecture) for Celeb-A† as Celeb-A.

3.2. Implementation Details

Following [6, 10, 12], we use ResNet-18 [3] to implement the embedding backbone ϕ. Different from the previous literature fine-tuning the backbone based on the weights pre-trained on ImageNet [1], we also consider the case that we train the full model from scratch. We find although the pre-trained weights make the model predict triplet well, it losses the coverage of semantics among conditions. The last downsampling layer in the backbone is removed to accommodate the smaller image size, and an additional fully connected layer is appended to project the embeddings to specified dimensions. We set the embedding dimension 64 and the temperature ς in Eq. 10 in the main paper as 1.0 for both UT-Zappos-50k and Celeb-A. We use the Adam optimizer [4] in our experiments and train our model for 90 epochs totally. The initial learning rate is 0.01 and annealed to 10% every 30 epochs. When the model is fine-tuned from the pre-trained weights, we set the initial learning rate as 5e-4, and the learning rate of the last layer is 10 times faster. In the case of training a model from scratch, the initial learning

Table 2. Greedy accuracy and OT accuracy on 8-condition Celeb-A (binary conditions) and its attribute merged variant Celeb-A† with five multi-choice conditions, respectively. All methods are fine-tuned with pre-trained weights. We make the best WS-CSL results in bold.

	Celeb-A		Celeb-A†	
Criteria →	GR Acc.	OT Acc.	GR Acc.	OT Acc.
CSN [12]	84.88	84.88	73.04	73.04
LSN [6]	72.89	71.95	63.73	63.67
SCE-Net [10]	69.91	68.73	60.26	59.73
DiscoverNet$_{\text{Set}}$	**80.65**	**78.81**	64.23	63.49
DiscoverNet$_{\text{Reg}}$	79.31	78.79	**65.32**	**64.71**

Table 3. Influence of the number of training triplets for DiscoverNet$_{\text{Set}}$ and DiscoverNet$_{\text{Reg}}$ on Celeb-A. Models are fine-tuned with pre-trained weights.

	DiscoverNet$_{\text{Set}}$		DiscoverNet$_{\text{Reg}}$	
# Number	GR Acc.	OT Acc.	GR Acc.	OT Acc.
1×10^5	79.08	78.07	78.18	75.66
2×10^5	79.83	77.55	78.28	76.14
4×10^5	80.65	78.81	79.31	78.79

rate is 0.01. Other configurations are the same as the scenario optimized over the pre-trained weights. Margin loss is used to optimize the embedding, and each mini-batch contains 64 triplets.

4. Additional Experiments

In this section, we further investigate our proposed DiscoverNet, including additional benchmark evaluations, some ablation studies, and several visualizations omitted in the main paper.

4.1. Additional Benchmark Evaluations

Celeb-A. The results of all methods over Celeb-A variants are listed in Table 2. We observe that DiscoverNet variants get the best accuracy with greedy or OT alignments among other WS-CSL methods.

4.2. Ablation Studies

Another choice of the set module. In DiscoverNet$_{\text{Set}}$, we use the maximum operator to make the output of the pair sets in a triplet become permutation invariant [16]. Inspired by [13], we can also investigate the mapping function g with Transformer [11]. The multi-head self-attention mechanism keeps the set property of the mapping but improves the learning ability. After replacing g with the Transformer, we name the variants of set module as DiscoverNet$_{\text{Tra}}$. We compare the performance of DiscoverNet variants on the Zappos dataset. As shown in Table 4, DiscoverNet$_{\text{Tra}}$ outperforms DiscoverNet$_{\text{Set}}$ by a large margin if the model is optimized without the pre-trained weights.

Table 4. Performance comparison between DiscoverNet$_{\text{Tra}}$ (the Transformer [11] implementation of g) and DiscoverNet$_{\text{Set}}$ on UT-Zappos-50k. We investigate two cases that training the model from scratch and fine-tune the model with pre-trained weights. Both GR accuracy and OT accuracy are measured.

Setups →	w/ pretrain		w/o pretrain	
Criteria →	GR Acc.	OT Acc.	GR Acc.	OT Acc.
DiscoverNet$_{\text{Set}}$	76.98	75.68	74.67	74.13
DiscoverNet$_{\text{Reg}}$	**77.84**	**77.68**	72.99	71.46
DiscoverNet$_{\text{Tra}}$	77.54	77.30	**75.72**	**75.46**

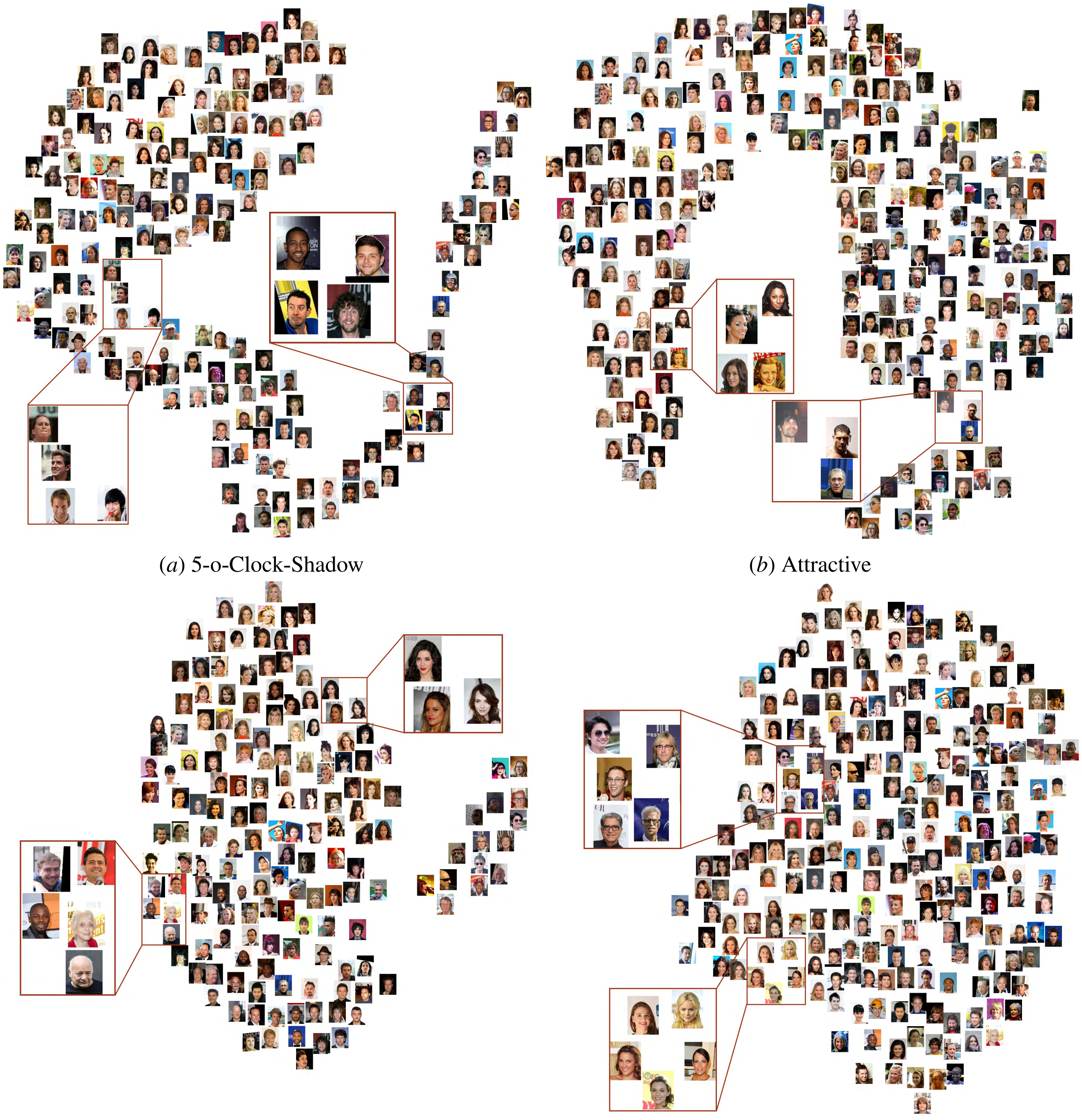
Table 5. Influence of the balance weight λ over the training regularizer for DiscoverNet$_{\text{Reg}}$ on Celeb-A. Models are fine-tuned with pre-trained weights.

λ	0	0.0001	0.001	0.01	0.1
GR Acc.	73.81	77.51	77.70	75.19	50.87
OT Acc.	69.52	76.19	77.09	74.07	50.46

Influence of training triplets. We show the influence of the triplets number in Table 3, where we vary the number of triplets during the training progress of DiscoverNet. More training triplets make it easier for a model to infer the latent conditions of triplets and shape the various spaces. As shown in Table 3, when there are more training triplets, both the GR accuracy and OT accuracy increase. The results also indicate that DiscoverNet is able to learn discriminative conditional embeddings and identify the latent conditions given a relatively small number of training triplets.

Influence of the balance weight λ of semantic regularization. Table 5 shows the influence of λ in DiscoverNet$_{\text{Reg}}$. We find that DiscoverNet$_{\text{Reg}}$ can get higher accuracy when λ increases around 0.001, which indicates the regularization indeed helps.

Visualization of semantic embeddings. To better illustrate the ability that DiscoverNet learns meaningful semantics given the triplets, we show the TSNE visualizations for each of the learned semantic spaces by DiscoverNet$_{\text{Reg}}$ on Celeb-A in Fig. 2 and Fig. 3. Obviously, DiscoverNet captures the variety of conditions and learns different embeddings for the dataset with good interpretability. Typically, in Fig. 2 (d), on the "eyeglasses"



(*a*) 5-o-Clock-Shadow

(*b*) Attractive

(*c*) Bags-Under-Eyes

(*d*) Eyeglasses

Figure 2. TSNE of the learned embeddings for each of the four conditions (*i.e*., 5-o-Clock-Shadow, attractive, bags-under-eyes and eyeglasses) on Celeb-A dataset based on DISCOVERNET$_{\text{Reg}}$.

condition, DISCOVERNET gathers all faces with eyeglasses in the center-left part of the image. Similarly, in Fig. 3 (b), on the "smiling" condition, DISCOVERNET gathers all faces without smile in the upper-left part of the image. Note that DISCOVERNET learns the semantic metric spaces without any condition labels.

Visualization of conditional image retrieval. To gain insights into the conditions learned by our model (DISCOVERNET$_{\text{Reg}}$), we provide image-retrieval visualizations for four conditions on UT-Zappos-50k in Fig. 4 and Fig. 5. Generally speaking, DISCOVERNET can learn the distance between images based on a certain semantic. For example, on the "suggested gender" condition in Fig. 5 (a), our model can make all images related to Male (resp. Female) closer to the anchor related to Male (resp. Female) while pushing images related to Female (resp. Male) away.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4

[2] Abner Guzmán-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *NIPS*, pages 1808–1816, 2012. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 4

[6] Ishan Nigam, Pavel Tokmakov, and Deva Ramanan. Towards latent attribute discovery from triplet similarities. In *ICCV*, pages 402–410, 2019. 2, 4, 5

[7] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. 3

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 4

[9] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016. 4

[10] Reuben Tan, Mariya I. Vasileva, Kate Saenko, and Bryan A. Plummer. Learning similarity conditions without explicit supervision. In *ICCV*, pages 10373–10382, 2019. 2, 4, 5

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 5

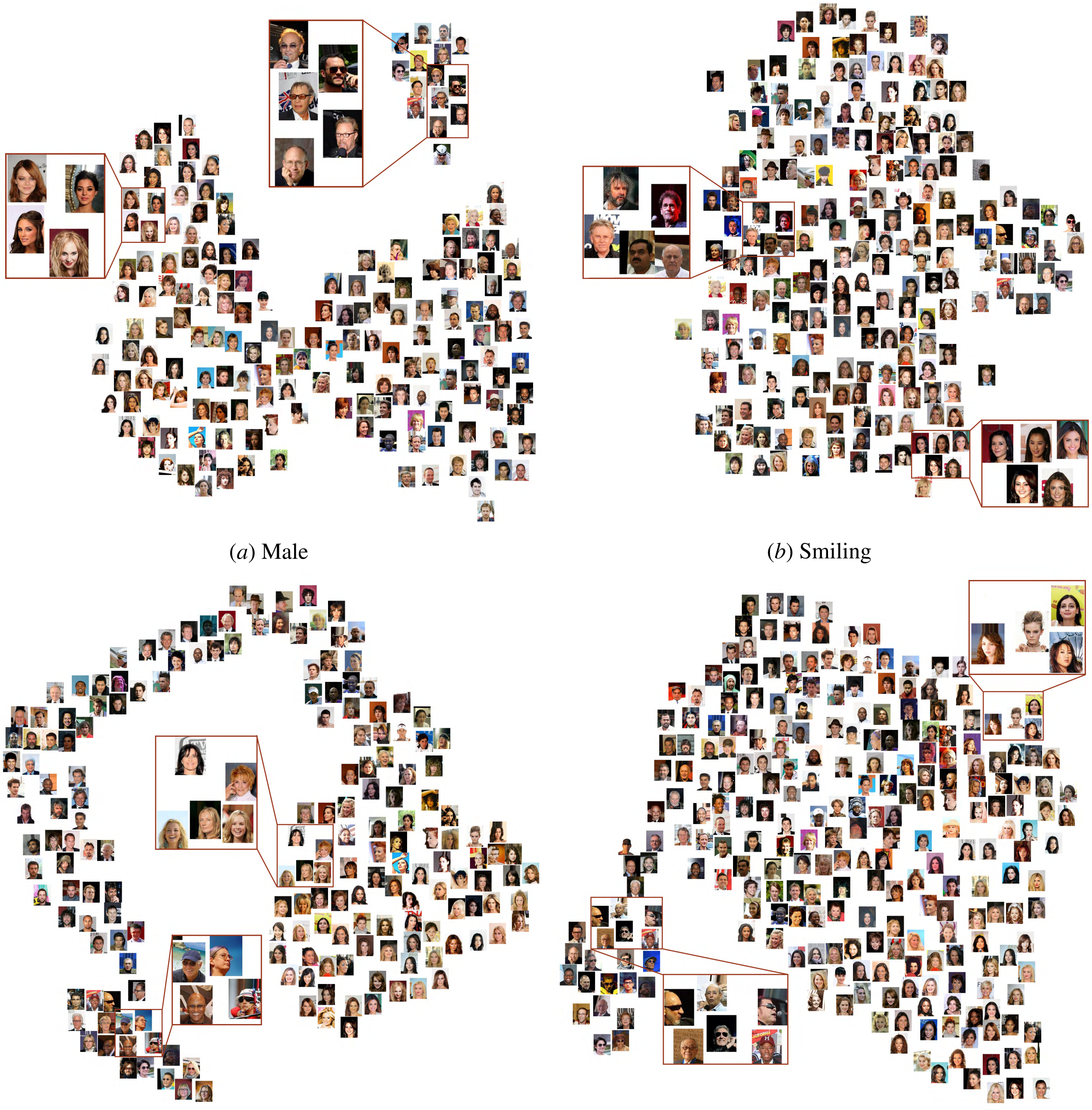
[12] Andreas Veit, Serge J. Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *CVPR*, pages 1781–1789, 2017. 2, 4, 5

[13] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, pages 8805–8814, 2020. 5

[14] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, pages 192–199, Jun 2014. 2

[15] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *ICCV*, pages 5571–5580, 2017. 2

[16] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *NIPS*, pages 3391–3401, 2017. 5



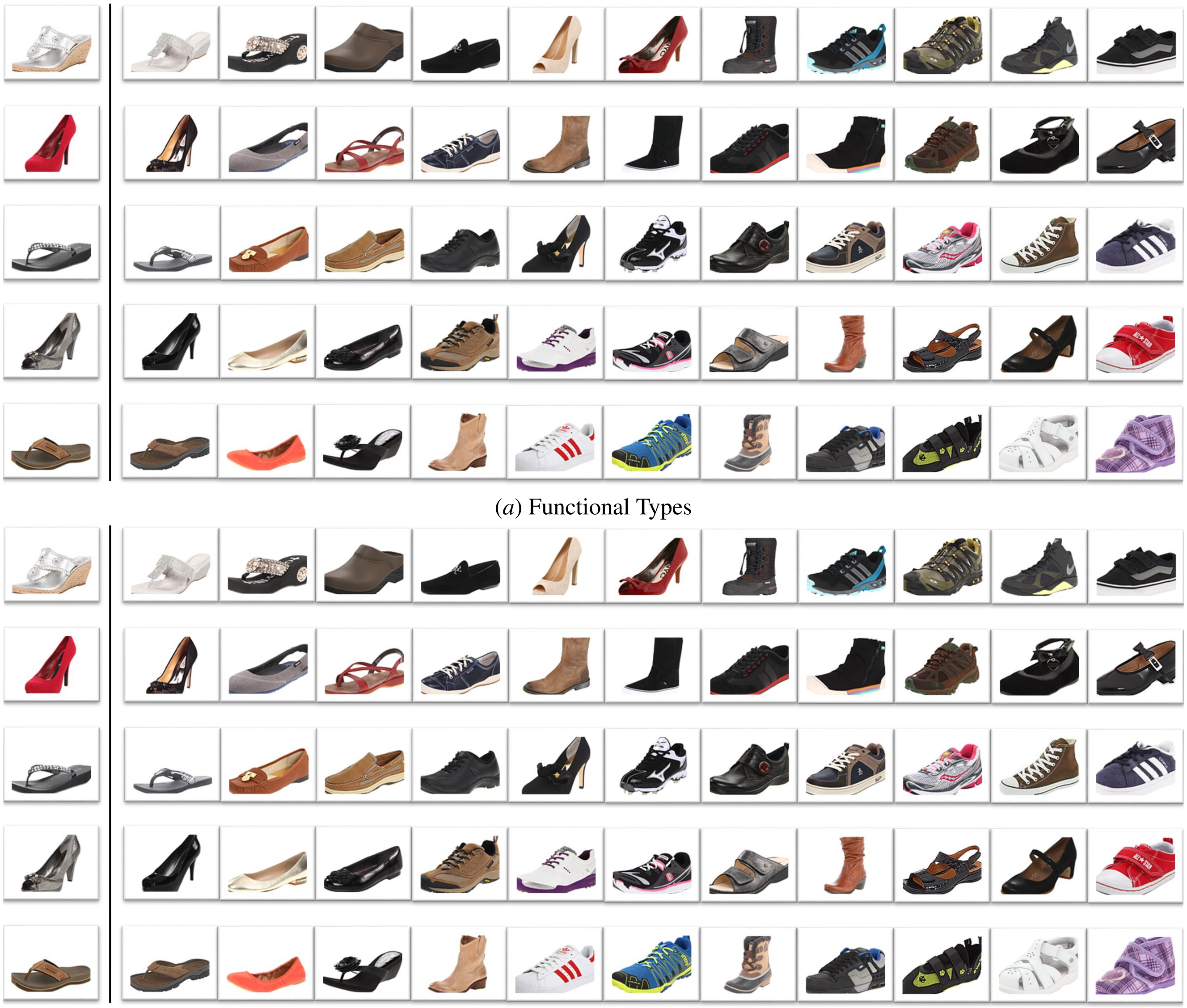
(*a*) Male

(*b*) Smiling

(*c*) Wearing-Lipstick

(*d*) Young

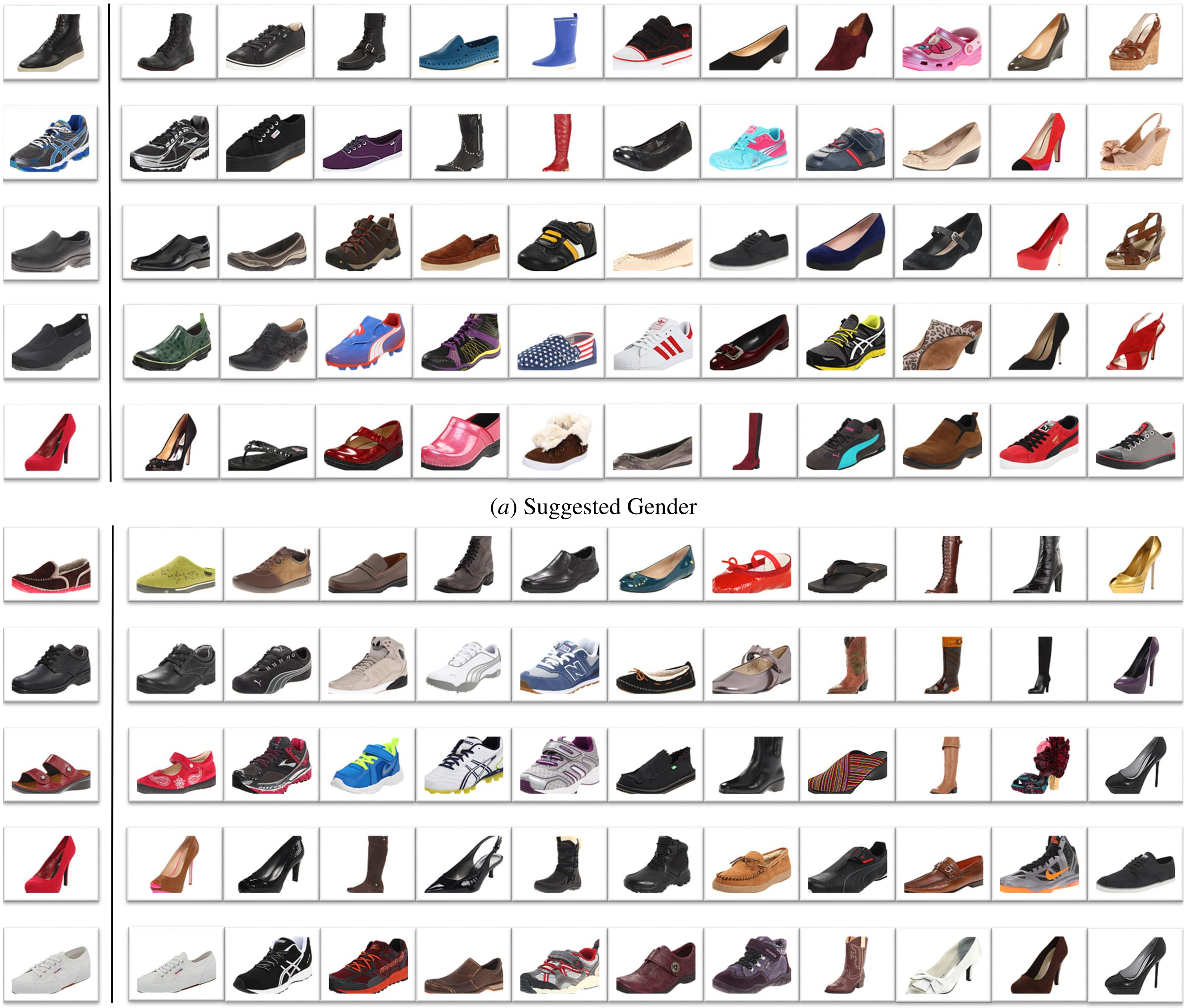
Figure 3. TSNE of the learned embeddings for each of the four conditions (*i.e.*, male, smiling, wearing-lipstick, and young) on Celeb-A dataset based on DISCOVERNET$_{\text{Reg}}$.



(*a*) Functional Types

(*b*) Closing Mechanism

Figure 4. Visualization of the image retrieval results for each of the two conditions (*i.e.*, functional types and closing mechanism) on UT-Zappos-50k dataset with the learned embedding of DISCOVERNET$_{\text{Reg}}$. The first image in each row is the query item, and shoes are ranked by distances to the query item in ascending order.



(*a*) Suggested Gender

(*b*) Height of Heels

Figure 5. Visualization of the image retrieval results for each of the two conditions (*i.e*., suggested gender and height of heels) on UT-Zappos-50k dataset with the learned embedding of DISCOVERNET$_{\text{Reg}}$. The first image in each row is the query item, and shoes are ranked by distances to the query item in ascending order.