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

Rich semantics inside an image result in its ambiguous relationship with others, i.e.*, two images could be similar in one condition but dissimilar in another. Given triplets like "aircraft" is similar to "bird" than "train", Weakly Supervised Conditional Similarity Learning (WS-CSL) learns multiple embeddings to match semantic conditions* without explicit condition labels *such as "can fly". However, similarity relationships in a triplet are uncertain except* providing a condition*. For example, the previous comparison becomes invalid once the conditional label changes to "is vehicle". To this end, we introduce a novel evaluation criterion by predicting the comparison's correctness after assigning the learned embeddings to their optimal conditions, which measures how much WS-CSL could cover latent semantics as the supervised model. Furthermore, we propose the Distance Induced Semantic COndition VERification Network (*DISCOVERNET*), which characterizes the instance-instance and triplets-condition relations in a "decompose-and-fuse" manner. To make the learned embeddings cover all semantics,* DISCOVERNET *utilizes a set module or an additional regularizer over the correspondence between a triplet and a condition.* DISCOVERNET *achieves state-of-the-art performance on benchmarks like UT-Zappos-50k and Celeb-A w.r.t. different criteria.*

1. Introduction

Learning embeddings (a.k.a. representations) from data benefits machine learning and visual recognition systems [1,6,11,25,27,39,42,43,49]. Side information such as triplets [9, 10, 22, 29, 31] indicates the comparison relationship of objects, from which the embedding pulls visually similar objects close while pushing dissimilar ones away.

The linkage between objects conveys rich information about the object itself as well as its relationship with others, which becomes ambiguous when the similarity is measured from different perspectives. As illustrated in Fig. 1 (upper), (a) female with glasses, (b) female without glasses, and (c) male with glasses are organized in triplets.[1] We think (a) and (b) are more similar when we measure based on "gender". In contrast, we also treat (a) and (c) as neighbors since they "wear glasses". Since one embedding space outputs fixed relationships between instances, learning multiple embeddings facilitates discovering rich semantics.

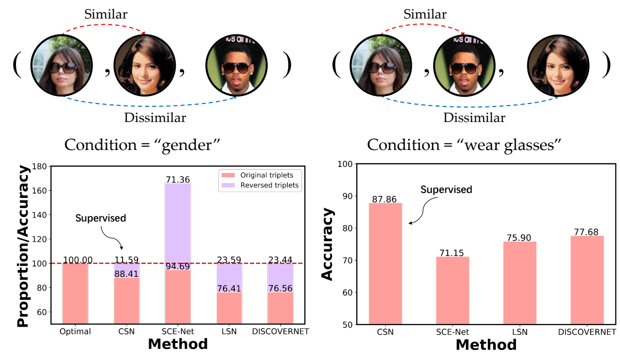


Figure 1. Upper: Two triplets with the same set of instances could be *both* meaningful when we measure similarity from different conditions. Lower Left: Given original (correct) triplets and their reversed variants (invalid w.r.t. the same similarity condition) on UT-Zappos-50k, we compute the proportion a model predicts them as valid ones. Last three are WS-CSL methods. Lower right: Our proposed criterion avoids the issue of reversed triplets naturally, and our DISCOVERNET outperforms other WS-CSL methods.

Given triplets associated with their condition labels, indicating under what kind of similarity the comparisons are made, Conditional Similarity Learning (CSL) learns multiple embeddings to cover latent semantic [36, 38]. During the evaluation, CSL predicts whether a triplet is meaningful or not under a specified condition. Although supervised CSL has been successfully applied in various applications [18, 19, 26, 34], labeling conditions introduces additional costs. As in a recommendation system, users may click relevant items (label item-wise similarities) based on particular preferences, and we only collect diverse comparison relationships *without explicit condition labels*. [24, 32] propose *Weakly Supervised*-CSL (WS-CSL), where multiple embeddings are learned with triplets and the model is *unaware of* their corresponding conditions.

[1] In a valid triplet (a, b, c), (a, b) is more similar than (a, c).

Current WS-CSL borrows the evaluation protocol from supervised CSL [36], which checks the validness of a triplet but *neglects the specified condition*. For example, we predict whether "aircraft" is similar to "bird" than "train" conditioned on "can fly" in the supervised scenario, but ask the model to predict the correctness of the triplet *free from the condition* in WS-CSL. Since a triplet could be ambiguous, WS-CSL may focus on *weighting those embeddings to explain the triplet instead of learning semantically conditional embeddings corresponding to the ground-truth*.

We demonstrate the challenge with the following experiment. After removing all condition labels of correct triplets, we compute the proportion a WS-CSL model predicts those triplets as valid ones via its learned embeddings, which equals the "accuracy" used in WS-CSL evaluation. Then we *reverse* those triplets — changing the order of the second and the third items, which makes them invalid under previous conditions. We further compute the valid ratio over reversed triplets. Results in Fig. 1 (lower left) show an optimal supervised model has 100% and 0% proportion/accuracy in two cases. A WS-CSL method SCE-Net [32] has a higher proportion than the supervised CSN [36] on original triplets as well as a high proportion on reversed ones, which means it treats both original and reversed triplets as valid. The diverse results between WS-CSL and its supervised counterparts indicate the "accuracy" is biased towards predicting all triplets as valid, and the correctness of a triplet without a specified condition is meaningless. To check the coverage of all semantics, we propose to measure the comparison ability of multiple learned embeddings after *assigning them to target conditions*.

We also propose Distance Induced Semantic COndition VERification Network (DISCOVERNET) to balance the ability of *comparison prediction* and *semantic coverage*. DISCOVERNET works in a "decompose-and-fuse" manner, which identifies similarity conditions and captures the ambiguous relationship with discriminative embeddings.

We aim to ensure the learned multiple embeddings in WS-CSL are able to reveal the similarities in the target conditions as much as the supervised methods. In DISCOVERNET, we achieve the goal from two aspects. First, we use a set module to map various triplets with the same set of instances to one condition, avoiding messy training update signals. On the other hand, we add a regularizer to force selecting different conditional embeddings when a model predicts both a triplet and its reversed one as valid. DISCOVERNET demonstrates higher performance on benchmarks like UT-Zappos-50k in our newly proposed criterion, which is shown in Fig. 1 (lower right). The code is available at https://github.com/shiy19/DiscoverNet.

Our contributions could be summarized as:

- We point out the challenges in WS-CSL evaluation and design a novel criterion.
- Based on the proposed DISCOVERNET, we improve the quality of learned embeddings and match the ground-truth semantics from two aspects.
- DISCOVERNET can identify latent rich conditions and works better than others in our criterion.

2. Related Work

Metric Learning. Learning embeddings from data attracts lots of attention in machine learning and computer vision. The learned embeddings encode the relationship between objects well and facilitate downstream tasks [4, 16, 27, 31]. With the guidance of the comparison relationship between object pairs [6], triplets [39, 43], and higher-order statistics [17], visually similar instances are pulled together, and visually dissimilar ones are pushed away. Types of loss functions are proposed [2,9,10,22,25,30,31] to take full advantage of the similarity comparisons between instances in a mini-batch, *e.g.*, the triplet loss [27] and N-Pair loss [29].
Conditional Similarity Learning (CSL). Different from measuring all linkages with a single metric, the relationship between objects could be measured from diverse aspects (under different conditions) [3, 23, 33]. In [36, 46], CSL is investigated by associating conditions with image attributes, and multiple diverse embeddings could be derived from feature masks to capture the semantic of various conditions. CSL has been applied in applications like image-text classification [20,24], fashion retrieval [8,15,32], zero-shot learning [48] and video grounding [28]. Condition labels during training link a particular embedding with a condition, and the learned embeddings are evaluated by predicting the validness of a triplet *under a certain condition*.
Weakly supervised CSL. Explicit condition labels are unavailable in some cases, *e.g.*, we get a comparison tuple once a user selects an item than others without any knowledge of his/her preference (condition). The weakly supervised CSL, *i.e.*, learning conditional embeddings without condition annotations is investigated in [1], where a model infers conditions given a triplet and then decides the right embedding to use for training and deployment. [32] emphasizes the comparison ability of the fused embedding, while [24, 42] select one metric from multiple candidates to explain a triplet. Using the same evaluation protocols as CSL, weakly supervised CSL can get higher performance than supervised CSL without involving the ground-truth condition labels. We analyze the protocol and point out its drawbacks of semantic coverage. We propose a new criterion to reveal the difference between the quality of the learned embedding and its supervised counterpart. We also propose DISCOVERNET to trade-off the diversified embedding and semantic fusion in a "decompose and fuse" manner. A set module and a semantic regularizer are discussed to make DISCOVERNET cover all target conditions.

3. Notations and Background

We organize comparisons into triplets, *i.e.*, $\mathcal{T} = \{\tau = (\mathbf{x}, \mathbf{y}, \mathbf{z})\}$. For three items in τ: the anchor $\mathbf{x} \in \mathbb{R}^D$ has a similar target neighbor $\mathbf{y} \in \mathbb{R}^D$, and an impostor $\mathbf{z} \in \mathbb{R}^D$ is dissimilar with $\mathbf{x}$. Similarity in τ could be determined by categories of instances. The embedding, a.k.a., feature extractor, $\phi : \mathbb{R}^D \to \mathbb{R}^d$ projects an instance $\mathbf{x}$ into a latent space, whose distance with $\mathbf{y}$ is

$$\mathbf{Dis}_L^2(\phi(\mathbf{x}),\ \phi(\mathbf{y})) = \|L^\top(\phi(\mathbf{x}) - \phi(\mathbf{y}))\|_2^2\ . \quad (1)$$

$L \in \mathbb{R}^{d\times d'}$ is a projection. We implement ϕ with deep neural network, and learn embedding by minimizing the violation of comparisons over $|\mathcal{T}|$ triplets. Define the validness of a triplet τ by comparing the distances between "anchor-impostor" and "anchor-target neighbor", *i.e.*,

$$\mathbf{Diff}_\tau = \mathbf{Dis}_L^2(\phi(\mathbf{x}),\ \phi(\mathbf{z})) - \mathbf{Dis}_L^2(\phi(\mathbf{x}),\ \phi(\mathbf{y}))\ . \quad (2)$$

If the anchor has a larger distance with the impostor than with the target neighbor, the embedding is consistent with the relationship in τ. Given a loss $\ell(\cdot)$, *e.g.*, the margin loss or the logistic loss, we optimize $\min_{\phi,L} \sum_{\tau\in\mathcal{T}} \ell(\mathbf{Diff}_\tau - \gamma)$. Then $\mathbf{Diff}_\tau$ would be larger than the margin γ, so that anchor and the target neighbor are pulled while the impostor would be pushed away.

Conditional similarity learning (CSL). Comparisons in triplet τ may vary across environments. For example, when we ask which image in a candidate set is close to a given anchor, different workers may measure the similarities from their own aspects. In CSL, we associate a condition label $k \in \{1, \ldots, K\}$ with each triplet, *i.e.*, $\mathcal{T} = \{\tau = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)\}$, indicating based on the k-th condition the comparison in τ is made. One embedding ϕ fails to capture the diverse relationship when comparisons with the same set of instances are opposite, *e.g.*, $\tau_1 = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)$ vs. $\tau_2 = (\mathbf{x}, \mathbf{z}, \mathbf{y}, k')$. CSL extends ϕ, L, and $\mathbf{Diff}_\tau$ to ϕ_k, L_k, and $\mathbf{Diff}_\tau^k$, respectively. K embeddings Ψ_K covering K conditions are optimized:

$$\min_{\Psi_K=\{\psi_k=L_k\circ\phi_k\}_{k=1}^K} \sum_{\tau\in\mathcal{T}} \ell\left(\sum_{k'=1}^{K} I[k'=k]\left[\mathbf{Diff}_\tau^{k'}\right] - \gamma\right)\ . \quad (3)$$

$I[\cdot]$ outputs 1 when the input is true and 0 otherwise. In Eq. 3, only the distance corresponding to condition k of the triplet τ is activated. In evaluation, ψ_k is used to check whether a triplet from the k-th condition is valid or not.

Weakly supervised CSL (WS-CSL). Due to additional annotation costs and ambiguity of condition labels, the triplet-wise condition labels $\{k\}$ are *unknown* in WS-CSL. The model needs to *infer* condition labels and activates the corresponding embeddings to capture triplet's characteristic. The target of WS-CSL is to learn a set of embeddings Ψ_K, which is able to capture those semantics of target conditions as much as the supervised CSL.

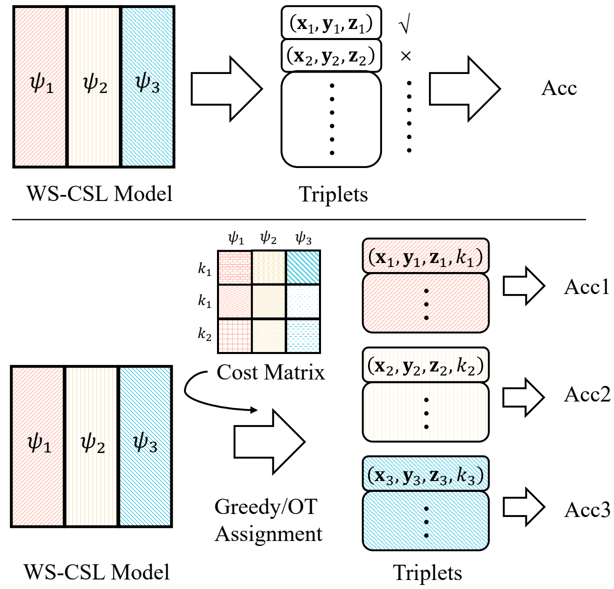


Figure 2. Comparison between previous (upper) and our proposed (lower) WS-CSL criterion. Instead of predicting the validness of triplets indistinguishably, We align the learned embeddings with target conditions and compute condition-specific accuracy.

4. Evaluation of Weakly Supervised CSL

Current WS-CSL follows CSL to measure ψ_k by predicting whether three instances could form a triplet *without a condition label*. We analyze the issues of the current criterion and propose our new evaluation protocol.

4.1. Analysis of Supervised CSL Evaluation

In supervised CSL evaluation, a model is asked to determine whether a triplet $\tau_1 = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)$ is valid or not given the condition k. In detail, we get $\mathbf{Diff}_\tau^k$ with the corresponding conditional embedding ψ_k and predict τ_1 as valid if $\mathbf{Diff}_\tau^k > 0$. Since $\tau_1 = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)$ and $\tau_2 = (\mathbf{x}, \mathbf{z}, \mathbf{y}, k)$ could not co-exist, we sample the same number of valid triplets from each condition in evaluation. The average accuracy (proportion of triplets predicted as correct) over them reveals the quality of a CSL model.

Since the ability of comparison prediction is important, previous WS-CSL methods follow this protocol *without using the test-time condition labels*. In other words, we predict whether a triplet $\tau_3 = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ is correct without its condition label. So in this case, a WS-CSL model predicts the validness of a triplet with *all learned embeddings*.

There are two issues when evaluating with the supervised protocol directly. First, a model tends to find a condition to explain the triplet, which predicts all triplets as true in most cases, *e.g*., even for reversed triplets (demonstrated in Fig. 1). Specifically, given a valid $\tau_1 = (\mathbf{x}, \mathbf{y}, \mathbf{z}, k)$, we construct $\tau_2 = (\mathbf{x}, \mathbf{z}, \mathbf{y}, k)$, and ask a WS-CSL model whether $\tau_4 = (\mathbf{x}, \mathbf{z}, \mathbf{y})$ is valid or not. Since τ_2 is invalid under the same condition k, an optimal supervised model will predict τ_1 as true and τ_2 as false. However, a WS-CSL method SCE-Net [32] predicts most original (τ_3) and reversed (τ_4) triplets as valid, which makes the evaluation biased.

Moreover, the current evaluation predicts the validness of the triplet while neglecting from which condition the WS-CSL model determines the relationship. We'd like a WS-CSL model works similarly to a supervised one, so 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. Therefore, we propose a new criterion to meet the previous requirements.

4.2. Condition Alignment for WS-CSL Evaluation

We claim that predicting the validness of a triplet is only meaningful *given a specific condition*. Based on the ground-truth conditions and the validness of triplets *during the evaluation*, we propose a *two-step* new criterion. First, we map target conditions to WS-CSL embeddings Ψ_K. Then, we evaluate triplets accuracy with corresponding ψ_k as in the supervised scenario, which reveals how much a WS-CSL model covers the target conditions as the supervised model.

We assume there is the same number of embeddings and conditions.[2] Given triplets from the k-th condition for evaluation, we compute the triplet prediction accuracy with $\{\psi_{k'}\}_{k'=1}^{K}$. We collect accuracy for all conditions and form a cost matrix $C \in \mathbb{R}^{K\times K}$, whose element $C_{k'k}$ is the error (100 minus the accuracy) using k'-th embedding to predict the triplets from the k-th condition. The alignment could be obtained from C with the following two strategies.

Greedy Alignment. We use a greedy strategy to find the most suitable embedding, *i.e*., $\arg\min_{k'} C_{k'k}$, for a condition k. So one embedding may handle multiple conditions.

OT Alignment. We optimize the following Optimal Transport (OT) objective [37] to obtain a mapping $T \in \mathbb{R}^{K\times K}$ from the embedding set to the condition set

$$\min_{T\geq 0} \langle T, C\rangle \qquad \text{s.t.} \quad T\mathbf{1} = \frac{1}{K}\mathbf{1},\ T^\top \mathbf{1} = \frac{1}{K}\mathbf{1}\,. \tag{4}$$

Eq. 4 minimizes the total cost when we use an embedding to predict triplets from another condition. We set the marginal distribution of the transportation T as uniform. By minimizing Eq. 4 we obtain T as a map. Element $T_{k'k}$ reveals how much the k'-th embedding is related to the k-th condition. We can further obtain a one-to-one mapping based on T under assumptions [5]. Fig. 2 shows a comparison between our proposed and previous criteria.

[2]Our analysis could be extended when their numbers do not match.

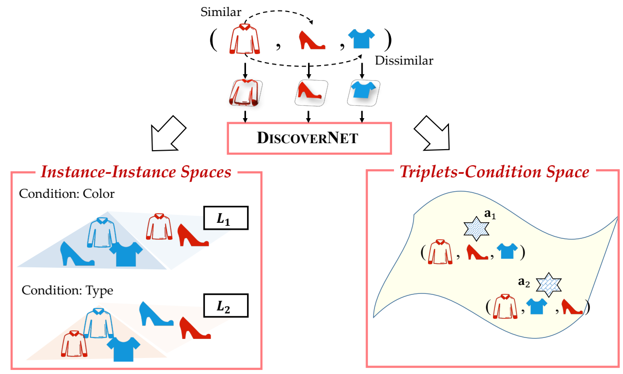


Figure 3. Illustration of DISCOVERNET with two types of space. The embedding space could be decomposed based on two different projections (L_1, L_2) and each of them has their own "instance-instance" linkage preference. For example, the first similarity condition is about "Color" and the second one focuses on "Type". Besides, we identify the latent similarity condition through the "triplets-condition" space, where a triplet is summarized as a vector by $g(\cdot)$ and we match this triplet with all condition anchors (a_1, a_2). The ability of DISCOVERNET to learn decomposed embeddings are demonstrated in Section 6.3.

Discussions. Based on our criterion, a WS-CSL model gets high accuracy only when all conditions can be explained with certain learned embeddings, and the model predicts triplets from diverse aspects well with different embeddings. Thus our criterion depicts the "semantic coverage" of Ψ_K, which also reveals the gap between Ψ_K in WS-CSL and its supervised counterpart. We may compute the cost C over the validation data with a small amount of ground-truth condition labels, and use the obtained alignment to compute the final accuracy on the test set. Benefiting from this condition-embedding alignment, our criterion naturally avoids the issue from the reversed triplets shown in Fig. 1.

5. DISCOVERNET for Weakly-Supervised CSL

We'd like to learn the embedding set Ψ_K to cover all rich semantics, and the comparison relationship under each target condition could be revealed by a certain conditional embedding ψ_k. We propose Distance Induced Semantic Condition VERification Network (DISCOVERNET) for WS-CSL (illustrated in Fig. 3). DISCOVERNET introduces a "triplets-condition" space to match condition scores of a triplet, which automatically selects or fuses multiple pairwise distances measured by "instance-instance" spaces. As we discussed before, a WS-CSL model could violate the semantic constraints over artificially reversed triplets, which influences the coverage of semantics. We consider a set module to avoid those cases during training, which maps various triplets with the same set of instances to one condition. Furthermore, we add a regularizer to force the model to select diverse conditions if both original and reversed triplets are predicted as valid ones.

5.1. Decompose and Fuse Hierarchical Spaces

We assume condition-specific embeddings Ψ have different projections $\mathcal{L}_K$ but share the same ϕ.[3] Then we have

$$\left\{\mathbf{Diff}_{\tau}^{k} = \mathbf{Dis}_{L_k}^{2}\left(\phi(\mathbf{x}), \phi(\mathbf{z})\right) - \mathbf{Dis}_{L_k}^{2}\left(\phi(\mathbf{x}), \phi(\mathbf{y})\right)\right\}_{k=1}^{K} . \tag{5}$$

The space measured by projection L_k biases towards a "local" view of the embedding ϕ, and allows inconsistent comparisons across different "instance-instance" spaces.

The overall linkages between objects are measured based on a fused distance over condition-specific comparisons. Different from the indicator in Eq. 3 selecting the distance, we use a multinomial distributed variable $c_\tau \in \{0,1\}^K$ to denote the latent condition label of a triplet τ. There are K binary elements in c_τ. We optimize ϕ and $\mathcal{L}_K$ jointly:

$$\min_{\phi, \mathcal{L}_K} \sum_{\tau \in \mathcal{T}} \ell\left(\mathbb{E}_{c_\tau}\left[\mathbf{Diff}_{\tau}^{k}\right] - \gamma\right) . \tag{6}$$

The condition of the triplet is revealed by the *expected distance* w.r.t. the distribution of c_τ [44], which *selects or fuses* the related conditions of the triplet. If a triplet could only be valid under the k-th condition, we expect c_τ has the k-th element $c_\tau^k = 1$ and other elements equal 0. Then the expected distance activates the k-th metric, and the target neighbor (resp. impostor) is pulled close (resp. pushed away) with L_k. The k-th embedding becomes more specific for the condition. If the triplet is related to more conditions, we expect τ to fuse the semantics from multiple distances together. Therefore, values in τ determine whether to select or fuse conditional embeddings, which emphasizes the semantic coverage or comparison prediction, respectively.

Given $\mathbb{E}_{c_\tau}[\mathbf{Diff}_\tau^k] = \sum_{k=1}^{K} \Pr(c_\tau^k = 1)\mathbf{Diff}_\tau^k$, the variable c_τ indicates the influence of the posterior probability $\Pr(c_\tau^k = 1)$ that a triplet belongs to the k-th condition. We introduce a "triplets-condition" space, where a triplet is embedded to a point with mapping $g(\cdot)$. g summarizes the triplet τ and maps the set of embeddings in τ into a d-dimensional vector. We expect triplets with similar conditions will be close. Furthermore, K learnable anchors $\{\mathbf{a}_k \in \mathbb{R}^d\}_{k=1}^{K}$ captures those similarity conditions, and we match a triplet to its closest anchor:

$$\begin{aligned} \Pr(c_\tau^k = 1) &\sim \mathbf{Sim}(g(\tau),\ \mathbf{a}_k) \\ &= \frac{\exp\left(\cos(g(\tau),\ \mathbf{a}_k)/\varsigma\right)}{\sum_{k'}^{K} \exp\left(\cos\left(g(\tau),\ \mathbf{a}_{k'}\right)/\varsigma\right)} . \end{aligned} \tag{7}$$

We implement the similarity with cosine. $\varsigma > 0$ is the temperature. The larger the ς, the more uniform $\Pr(c_\tau^k = 1)$ is. If a triplet is related to the k-th condition, then its transformed vector $g(\tau)$ is close to the anchor $\mathbf{a}_k$. Then a larger $\mathbf{Sim}(g(\tau),\ \mathbf{a}_k)$ emphasizes the k-th condition when computing the expected distance. Benefited from the "decompose and fuse" manner in Eq. 7, a triplet that has a clear condition tendency will be close to a particular anchor, which makes c_τ becomes one-hot, and the embedding is highly correlated with the corresponding condition. Then embeddings will cover more semantics. Those ambiguous triplets will have a nearly uniform c_τ. Although hard to differentiate semantics in this case, their fused distances take advantage of all embeddings to predict the validness of triplets and improve the comparison prediction ability.

[3] Since conditions are not independent, it is not necessary to match each condition with an embedding ψ_k. More discussions are in experiments.

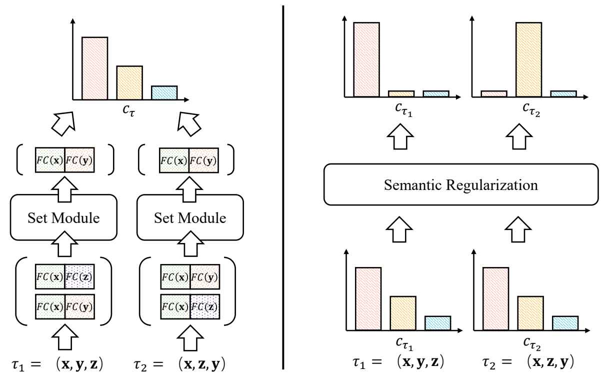


Figure 4. Illustration of the set module (left) and semantic regularization (right) to deal with artificially reversed triplets.

Instance Space for Semantic Discovery. Given $L_k \in \mathbb{R}^{d\times d}$, we define the k-th conditional embedding $\psi_k(\mathbf{x})$ as

$$\psi_k(\mathbf{x}) = \phi(\mathbf{x})(I + L_k) = \phi(\mathbf{x}) + \phi(\mathbf{x})L_k . \tag{8}$$

We obtain the local embedding ψ_k in a residual form, where L_k encodes the *similarity bias* of the k-th condition based on the general embedding ϕ. If ϕ is discriminative enough for a particular similarity condition, we do not need to over-allocate the local metric. In other words, strong ϕ makes L_k degenerate to zero. In summary, distances for different similarity conditions in Eq. 5 are calculated based on ψ_k.

5.2. Condition Space for Semantic Fusion

Based on our analyses in Section 4, a WS-CSL model could violate the semantic constraints in vanilla training. A *too flexible* mapping g will make the model treat both a triplet and its reversed version as correct ones while activating almost the same conditional embedding ψ_k. In other words, the model may focus on fusing those conditions with g and collapse all semantics into one conditional space.

Since semantic constraints on a WS-CSL model are limited, we address the issue with the help of *artificially reversed triplets*. For a pair of original triplet and its reversed version, we restrict the model's flexibility by either weakening the representation ability of g with a set module or adding another semantic regularizer (illustrated in Fig. 4).

Set Module. A natural way to implement g is to capture the order of instances in a triplet with a sequential model [14, 32]. As we mentioned, a triplet $\tau_1 = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ and its reversed version $\tau_2 = (\mathbf{x}, \mathbf{z}, \mathbf{y})$ should belong to different conditions. However, without any constraints, we find a sequential module has the ability to weight conditions in different manners for τ_1 and τ_2, but it is likely to work in a "lazy" way. In detail, a model may learn a strong conditional embedding ψ_k and map both τ_1 and τ_2 to similar distributions that have larger weights on ψ_k and small weights on others. Then, rather than making learned embeddings associated with corresponding conditional semantic meanings, a model tends to improve the fusion module and use partial embeddings to cover all semantics.

Therefore, we consider a set module to make g agnostic with the order in the triplet. In particular, we use the pairwise concatenation of instance embeddings as the input of g, *i.e.*, for a triplet τ_1, we obtain its embeddings, $\phi(\tau) = (\phi(\mathbf{x}), \phi(\mathbf{y}), \phi(\mathbf{z}))$ and get the augmented set:

$$\mathcal{P} = \{[\phi(\mathbf{x}), \phi(\mathbf{y})], [\phi(\mathbf{x}), \phi(\mathbf{z})]\} ,$$

which encodes the pairs in τ_1. Since $[\phi(\mathbf{y}), \phi(\mathbf{z})]$ is not included in $\mathcal{P}$, we get the same representation for τ_1 and the reversed τ_2. Moreover, to take a holistic view of τ, we transform with element-wise maximization [47]:

$$g(\tau) = \mathbf{FC} \max_{\mathbf{p}\in\mathcal{P}} \mathbf{FC}(\mathbf{p}) . \tag{9}$$

FC is a fully connected network with one hidden layer and ReLU activation, which projects the given input to a same-dimensional output. With Eq. 9, the pairwise relationship in the triplet is evaluated first, and the most evident one is activated to represent the whole triplet. In summary, the set module restricts the ability of g and makes one triplet and its reversed version have the same representation. Although it is a bit counter-intuitive, the special configuration of g avoids the ambiguous update directions of the model so that the conditional embeddings should be more discriminative to cover the target semantics. We also investigate the Transformer [35] implementation of g in the supplementary.

Semantic Regularizer. We can also think from another aspect by regularizing a representative g. We set Eq. 9 to

$$\mathcal{P} = \{[\phi(\mathbf{x}), \phi(\mathbf{y})], [\phi(\mathbf{x}), \phi(\mathbf{z})], [\phi(\mathbf{y}), \phi(\mathbf{z})]\} ,$$

covering the sequential information of instances. Due to an additional pair in $\mathcal{P}$, different orders of instances in τ_1 and τ_2 do not share the same output. In this case, it is abnormal when the model predicts τ_1 and τ_2 based on the same condition k. We construct a semantic regularization to avoid this case. If $\mathbf{Diff}^k_{\tau_1} > 0$ and $\mathbf{Diff}^k_{\tau_2} > 0$ (the model treats two triplets as valid ones),[4] we add the following semantic regularizer to make the activated conditions diverse:

$$\mathrm{Reg}(g) = \lambda \sum_{k=1}^{K} \min\left(\mathbf{Sim}(g(\tau_1),\ \mathbf{a}_k),\ \mathbf{Sim}(g(\tau_2),\ \mathbf{a}_k)\right) . \tag{10}$$

In other words, if τ_1 and τ_2 are predicted as valid ones, we minimize the similarity between their condition distributions c_{τ_1} and c_{τ_2}. The similarity between these two multinomial distributions are measured with Histogram Intersection Kernel (HIK) [12, 40, 41], which is the sum of element-wise minimum of two distributions. Therefore, with the help of Eq. 10, we explicitly enforce the WS-CSL model considers different conditional embeddings to explain triplets, which further improve the semantic coverage of Ψ_K.

Discussions. There are two implementations of DISCOVERNET, using the set module or the semantic regularizer. The two strategies are designed from different aspects to make the WS-CSL model contain rich semantics as much as the supervised model. Since the set module avoids the diverse selection for reversed triplet naturally, it satisfies the regularizer directly. Thus, it does not help if we combine two strategies together.

6. Experiments

We verify the effectiveness of DISCOVERNET over benchmarks based on our new criterion. Ablation studies and visualization results demonstrate DISCOVERNET learns conditions successfully as the supervised methods. Detailed setups and more results are in the supplementary.

6.1. Experimental Setups

Datasets. **UT-Zappos-50k Shoes** contains 50,025 images of shoes with four similarity conditions collected online [45, 46]. Following [24, 32, 36], we discretize the "height" condition and resize all images to 112 by 112. There are 200,000, 20,000, and 40,000 triplets following splits in [36] for training, validation and test, respectively. **Celeb-A Faces** has 202,599 face images of different identities [21]. 8 of the 40 attributes (conditions) are selected for analysis [24]. We resize all images to 112 by 112. 400,000/80,000/160,000 triplets are used for model training/validation/test. We construct a more difficult **Celeb-A**† by combining related binary attributes in Celeb-A together, where each multi-choice condition has 5-7 discrete values, We apply the same configuration for Celeb-A variants.

Splits. All models are trained and evaluated over triplets in [36]. Since there are no published triplets over Celeb-A, we randomly sample triplets by ourselves. Equal number of triplets for each attribute are sampled from the standard training, validation, and test split of Celeb-A [21]. For each triplet, we organize instances with the same attribute label into a similar pair. Otherwise, we think they are dissimilar.

[4] We check this condition with the current model during training. Parameters in $\mathbf{Diff}^k_\tau$ here are detached without gradient back propagation.

Table 1. GR accuracy and OT accuracy on UT-Zappos-50k. We investigate two cases that training the model from scratch and fine-tune the model with pre-trained weights. CSN is the fully supervised CSL method which utilizes condition labels during training. We make the best WS-CSL results in each case in bold.

Setups →	w/ pretrain		w/o pretrain	
Criteria →	GR Acc.	OT Acc.	GR Acc.	OT Acc.
CSN [36]	87.86	87.86	82.14	82.14
LSN [24]	76.26	75.90	71.49	68.49
SCE-Net [32]	72.21	71.15	64.08	61.03
DiscoverNet$_{\mathrm{Set}}$	76.98	75.68	**74.67**	**74.13**
DiscoverNet$_{\mathrm{Reg}}$	**77.84**	**77.68**	72.99	71.46

Criterion. We utilize our proposed new criterion to evaluate WS-CSL methods. After constructing a mapping from condition to learned embeddings with greedy or OT strategies, we compute the average triplet prediction accuracy as in the supervised case. We denote the results with two strategies as GR Accuracy and OT Accuracy, respectively.

Comparison Methods. We compare DiscoverNet with both supervised method Conditional Similarity Networks (CSN) [36] and two WS-CSL methods, *i.e*., Latent Similarity Networks (LSN) [24] and Similarity Condition Embedding Network (SCE-Net) [32].

Implementation Details. Following [24, 32, 36], we use ResNet-18 [13] to implement ϕ. Different from previous literature fine-tuning the backbone based on the weights pre-trained on ImageNet [7], we also consider the case that we train the full model from scratch. The model with the best accuracy over the validation set is selected for the final test.

6.2. Benchmark Evaluations

UT-Zappos-50k. The results of DiscoverNet and comparison methods over UT-Zappos-50k are listed in Table 1, which contains the greedy accuracy and OT accuracy over both pre-trained and non-pre-trained weights. We re-implement all comparison methods. DiscoverNet$_{\mathrm{Set}}$ and DiscoverNet$_{\mathrm{Reg}}$ denote the variant using set module and semantic regularization, respectively.

By observing both setups and criteria, CSN becomes the "upper-bound" for WS-CSL methods. The main reason is that ground-truth condition labels in CSN associate an embedding with a particular condition during training. The supervised CSN gets the same greedy and OT accuracy. The explicit supervision in CSN makes those embeddings biased towards different semantics, which gets the same one-to-one OT mapping with the greedy strategy. WS-CSL methods get lower OT accuracy than the corresponding greedy accuracy. The main reason is that OT accuracy requires a one-to-one mapping between embeddings and conditions, where those conditions would be related. While greedy accuracy allows one embedding to handle multiple conditions.

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 trained from scratch. More results are in the supplementary.

	Celeb-A		Celeb-A†	
Criteria →	GR Acc.	OT Acc.	GR Acc.	OT Acc.
CSN [36]	83.41	83.41	67.72	67.72
LSN [24]	70.29	69.89	57.02	56.33
SCE-Net [32]	69.47	51.81	51.49	50.64
DiscoverNet$_{\mathrm{Set}}$	**78.45**	**77.98**	**64.57**	**63.98**
DiscoverNet$_{\mathrm{Reg}}$	78.04	76.96	57.43	56.94

Table 3. Influence of the embedding number (the number of projections in $\mathcal{L}_K$) for DiscoverNet$_{\mathrm{Set}}$ and DiscoverNet$_{\mathrm{Reg}}$ on Celeb-A. Models are fine-tuned with pre-trained weights.

	DiscoverNet$_{\mathrm{Set}}$		DiscoverNet$_{\mathrm{Reg}}$	
# Projections	GR Acc.	OT Acc.	GR Acc.	OT Acc.
2	78.28	77.74	71.92	71.82
4	79.58	77.60	73.19	71.32
6	80.41	77.69	75.53	73.87
8	80.65	78.81	75.19	74.07
10	81.35	78.68	76.09	73.77

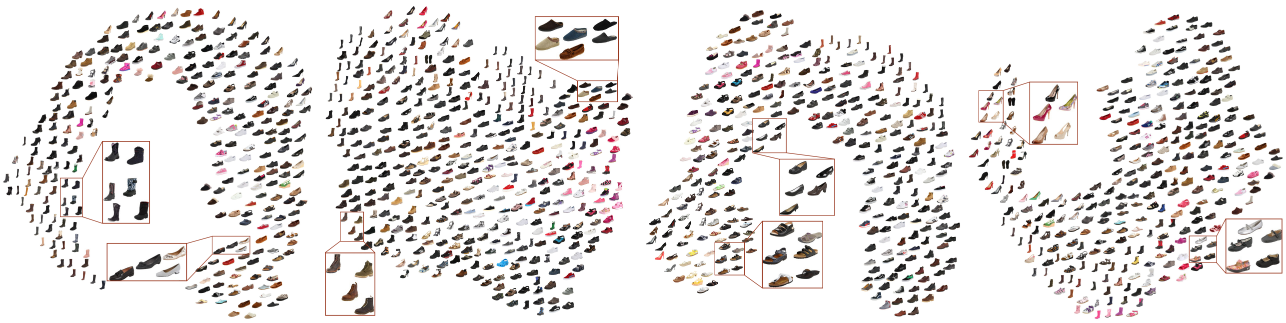
As we discussed before, SCE-Net tends to fuse the semantic meaning of embeddings with its self-attention module, so each of its learned embeddings is hard to cover a specific condition. By contrast, LSN performs better in our criteria, benefiting from its multi-choice learning paradigm. Our DiscoverNet can get the best performance among WS-CSL methods. In detail, DiscoverNet$_{\mathrm{Set}}$ works better when training from scratch and DiscoverNet$_{\mathrm{Reg}}$ performs well with the pre-trained weights. One possible reason is that the pre-trained weights are strong and make the model (especially the mapping function g) too flexible, so an explicit regularization helps more. In summary, our criterion reveals how much similar a WS-CSL model performs to a supervised one.

Celeb-A. In Table 2, we investigate (the 8 attribute) Celeb-A and its variant Celeb-A† with model trained from scratch. DiscoverNet$_{\mathrm{Set}}$ and DiscoverNet$_{\mathrm{Reg}}$ still get better performance than other WS-CSL counterparts, while CSN is still the 'upper bound' due to the help of condition labels.

6.3. Ablation Studies and Visualizations

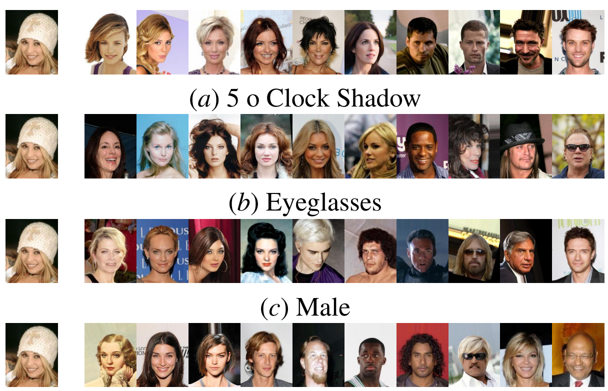
We analyze the properties of DiscoverNet and show the visualization results. More analysis such as the help of the WS-CSL embeddings given limited conditional supervisions are in the supplementary.

Influence of condition number configuration. We set



(a) Functional Types (b) Closing Mechanism (c) Suggested Gender (d) Height of Heels

Figure 5. TSNE of the learned embeddings for each of the four conditions (*i.e.*, functional types, closing mechanism, suggested gender and height of heels) on UT-Zappos-50k dataset based on DiscoverNet$_{\text{Reg}}$.



(a) 5 o Clock Shadow

(b) Eyeglasses

(c) Male

(d) Wearing Lipstick

Figure 6. Visualization of the image retrieval results for each of the four conditions on Celeb-A dataset with the learned embedding of DiscoverNet$_{\text{Reg}}$. The first image in each row is the anchor, and faces are ranked by distances to the anchor in an ascending order.

the number of embeddings in benchmarks the same as the number of ground-truth conditions, which is 8 in Celeb-A. We show the change of the two criteria along with the increase of the embedding number in Table 3. Both DiscoverNet$_{\text{Set}}$ and DiscoverNet$_{\text{Reg}}$ get higher OT accuracy when the number of embeddings increases from two to eight, and decreases from eight to ten. We conjecture that allocating too many local metrics interfere with each other and their related semantics partially overlap. Besides, greedy accuracy shows a generally incremental trend since it allows one embedding to handle multiple conditions.

Visualization of semantic embeddings. To better illustrate how our DiscoverNet learns for different semantics, we provide TSNE visualizations for each of the learned semantic spaces $\{\psi_k\}_{k=1}^K$ on UT-Zappos-50k in Fig. 5. DiscoverNet captures the variety of conditions and learns different embeddings for the dataset with good interpretability. Typically, on the condition "height of heels", the heel height of the shoes is decreasing from the left-top of the embedding space to the bottom, then to the right-top.

Visualization of conditional image retrieval. We provide image-retrieval visualizations for four conditions on Celeb-A in Fig. 6. We keep the anchor image, and retrieve its neighbor from a set of randomly collected candidates with different local embeddings. Benefiting from the correspondence obtained when computing the OT accuracy, we can qualitatively measure whether a local embedding could reveal the corresponding semantic via its ranking of images. For example, on the "Male" condition, since the anchor image has the label "Female", DiscoverNet makes all images labeled "Female" close while pushing images related to "Male" away. The results indicate our DiscoverNet can cover latent semantics of data and each of its learned conditional embedding ψ_k corresponds to a particular meaningful semantic.

7. Conclusion

We revisit WS-CSL and observe that evaluating the quality of the model without specifying concrete conditions produces biased accuracy. Thus, we match multiple learned embeddings with ground-truth conditions in advance before predicting the correctness of given triplets, which simultaneously considers the ability of triplet prediction and semantic coverage. We also utilize a set module or a semantic regularizer in our proposed DiscoverNet to emphasize the correspondence between a conditional embedding and a semantic condition. DiscoverNet outperforms other WS-CSL methods on benchmarks with different criteria.

Limitations. Our new criterion evaluates WS-CSL in a "supervised" manner by assigning multiple learned embeddings to target conditions. The criterion does not fit the case when the goal is not to learn a model similar to its supervised counterpart, *e.g.*, to distinguish whether triplets are correct and explain their validness as much as possible.

Acknowledgments. This research was supported by National Key R&D Program of China (2020AAA0109401), NSFC (62006112, 61921006, 62176117), Collaborative Innovation Center of Novel Software Technology and Industrialization, NSF of Jiangsu Province (BK20200313).

References

[1] Ehsan Amid and Antti Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *ICML*, pages 1472–1480, 2015. 1, 2

[2] Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics*, 34(4):98:1–98:10, 2015. 2

[3] Soravit Changpinyo, Kuan Liu, and Fei Sha. Similarity component analysis. In *NIPS*, pages 1511–1519, 2013. 2

[4] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, 2010. 2

[5] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *TPAMI*, 39(9):1853–1865, 2017. 4

[6] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007. 1, 2

[7] 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. 7

[8] Jianfeng Dong, Zhe Ma, Xiaofeng Mao, Xun Yang, Yuan He, Richang Hong, and Shouling Ji. Fine-grained fashion similarity prediction by attribute-specific embedding learning. *CoRR*, abs/2104.02429, 2021. 2

[9] Yueqi Duan, Jiwen Lu, Wenzhao Zheng, and Jie Zhou. Deep adversarial metric learning. *TIP*, 29:2037–2051, 2020. 1, 2

[10] Ismail Elezi, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixé. The group loss for deep metric learning. In *ECCV*, pages 277–294, 2020. 1, 2

[11] Boyuan Feng, Yuke Wang, Zheng Wang, and Yufei Ding. Uncertainty-aware attention graph neural network for defending adversarial attacks. *CoRR*, abs/2009.10235, 2020. 1

[12] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005. 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 6

[15] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *ICCV*, pages 12147–12157, 2021. 2

[16] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge J. Belongie, and Deborah Estrin. Collaborative metric learning. In *WWW*, pages 193–201, 2017. 2

[17] Marc T Law, Nicolas Thome, and Matthieu Cord. Learning a distance metric from relative comparisons between quadruplets of images. *IJCV*, pages 1–30, 2016. 2

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 212–228, 2018. 1

[19] Yen-Liang Lin, Son Tran, and Larry S. Davis. Fashion outfit complementary item retrieval. In *CVPR*, pages 3308–3316, 2020. 1

[20] Weiyang Liu, Zhen Liu, James M. Rehg, and Le Song. Neural similarity learning. In *NeurIPS*, pages 5026–5037, 2019. 2

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 6

[22] R. Manmatha, Chao-Yuan Wu, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *ICCV*, pages 2859–2867, 2017. 1, 2

[23] Roland Memisevic and Geoffrey E. Hinton. Multiple relational embedding. In *NIPS*, pages 913–920, 2004. 2

[24] Ishan Nigam, Pavel Tokmakov, and Deva Ramanan. Towards latent attribute discovery from triplet similarities. In *ICCV*, pages 402–410, 2019. 1, 2, 6, 7

[25] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Deep metric learning with BIER: boosting independent embeddings robustly. *TPAMI*, 42(2):276–290, 2020. 1, 2

[26] Bryan A. Plummer, Paige Kordas, M. Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *ECCV*, pages 258–274, 2018. 1

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 1, 2

[28] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *CVPR*, pages 10444–10452, 2019. 2

[29] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, pages 1849–1857, 2016. 1, 2

[30] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. Deep metric learning via facility location. In *CVPR*, pages 2206–2214, 2017. 2

[31] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016. 1, 2

[32] Reuben Tan, Mariya I. Vasileva, Kate Saenko, and Bryan A. Plummer. Learning similarity conditions without explicit supervision. In *ICCV*, pages 10373–10382, 2019. 1, 2, 4, 6, 7

[33] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55, 2012. 2

[34] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David A. Forsyth. Learning type-aware embeddings for fashion compatibility. In *ECCV*, pages 405–421, 2018. 1

[35] 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. 6

[36] Andreas Veit, Serge J. Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *CVPR*, pages 1781–1789, 2017. 1, 2, 6, 7

[37] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 4

[38] Xiaofang Wang, Kris M. Kitani, and Martial Hebert. Contextual visual similarity. *CoRR*, abs/1612.02534, 2016. 1

[39] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009. 1, 2

[40] Jianxin Wu. A fast dual method for HIK SVM learning. In *ECCV*, pages 552–565, 2010. 6

[41] Jianxin Wu, Nini Liu, Christopher Geyer, and James M. Rehg. ${\rm C}^{4}$: A real-time object detection framework. *TIP*, 22(10):4096–4107, 2013. 6

[42] Han-Jia Ye, De-Chuan Zhan, Yuan Jiang, and Zhi-Hua Zhou. What makes objects similar: A unified multi-metric learning approach. *TPAMI*, 41(5):1257–1270, 2019. 1, 2

[43] Han-Jia Ye, De-Chuan Zhan, Nan Li, and Yuan Jiang. Learning multiple local metrics: Global consideration helps. *TPAMI*, 42(7):1698–1712, 2020. 1, 2

[44] Han-Jia Ye, De-Chuan Zhan, Xue-Min Si, and Yuan Jiang. Learning mahalanobis distance metric: Considering instance disturbance helps. In *IJCAI*, pages 3315–3321, 2017. 5

[45] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, pages 192–199, Jun 2014. 6

[46] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *ICCV*, pages 5571–5580, 2017. 2, 6

[47] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *NIPS*, pages 3391–3401, 2017. 6

[48] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015. 2

[49] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *CVPR*, 2022. 1