Learning Similarity Conditions Without Explicit Supervision

Reuben Tan
Boston University
rxtan@bu.edu

Mariya I. Vasileva
University of Illinois at Urbana-Champaign
mvasile2@illinois.edu

Kate Saenko
Boston University
saenko@bu.edu

Bryan A. Plummer
Boston University
bplum@bu.edu

Abstract

Many real-world tasks require models to compare images along multiple similarity conditions (e.g. similarity in color, category or shape). Existing methods often reason about these complex similarity relationships by learning condition-aware embeddings. While such embeddings aid models in learning different notions of similarity, they also limit their capability to generalize to unseen categories since they require explicit labels at test time. To address this deficiency, we propose an approach that jointly learns representations for the different similarity conditions and their contributions as a latent variable without explicit supervision. Comprehensive experiments [1] across three datasets, Polyvore-Outfits, Maryland-Polyvore and UT-Zappos50k, demonstrate the effectiveness of our approach: our model outperforms the state-of-the-art methods, even those that are strongly supervised with pre-defined similarity conditions, on fill-in-the-blank, outfit compatibility prediction and triplet prediction tasks. Finally, we show that our model learns different visually-relevant semantic sub-spaces that allow it to generalize well to unseen categories.

1. Introduction

Reasoning about the similarity between images or data of different modalities is an inherent challenge in computer vision. Beyond its prevalence in fundamental problems such as image-sentence retrieval [41, 38], cross-domain image-matching [32, 16], attribution learning [4, 33] and visual categorization [29], it also has an increasingly prominent role in computer vision problems in the fashion and retail domains like outfit style modeling [14], fashion item retrieval and recommendation [10, 22] and automatic capsule wardrobe generation [15]. Metric learning (the task of learning a distance function between features based on supervised similar/dissimilar pairs) is a common approach used to tackle the above-mentioned problems and is often addressed by learning representations for objects in a unified embedding space, where the distances provide a measure of their similarity. However, this is not naturally representative of the real world. Objects can usually be described with multiple visual attributes such as color, shape or category. Consider the example where a red shirt is similar to a pair of red shoes in color but dissimilar in object category. A single embedding space is unable to learn representations for these contradicting notions of similarity. By discounting such valuable information, these embeddings are not able

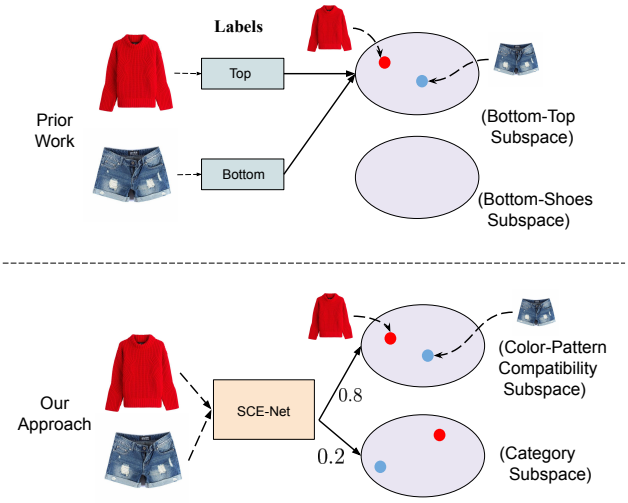


Figure 1: We propose the *SCE-Net* model for learning multi-faceted similarity between images, such as compatibility of two fashion items. Previous work needed user-defined labels to learn multiple feature subspaces for measuring different aspects of similarity, *e.g.*, one for comparing tops and bottoms and another for comparing bottoms and shoes (*e.g.*, [36, 27, 35]). In contrast, our approach learns important subspaces without such labels in a data-driven manner. The concepts and their contributions to a final similarity score are learned together as a single end-to-end trained model.

[1] https://github.com/rxtan2/Learning-Similarity-Conditions

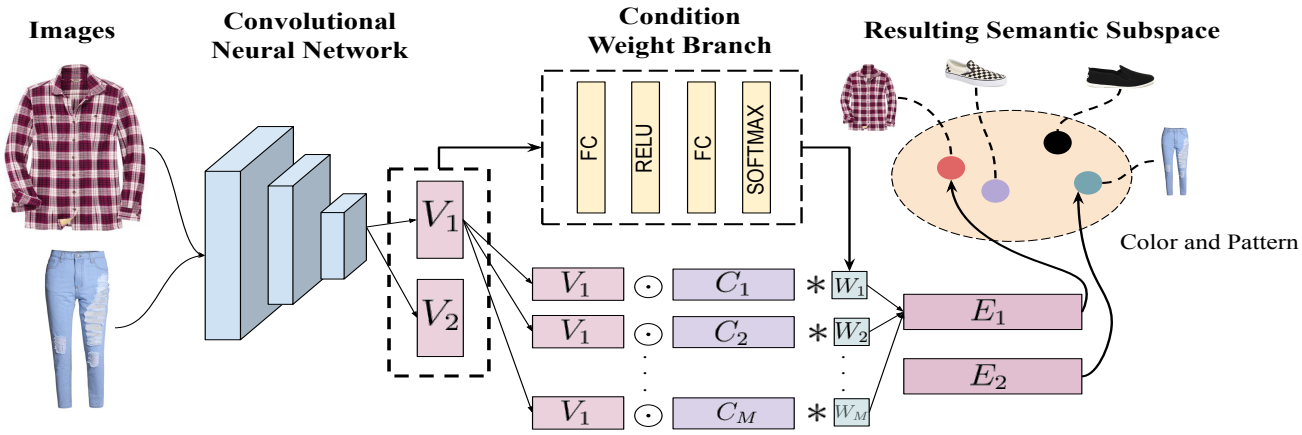


Figure 2: An overview of Similarity Condition Embedding Network (SCE-Net) which is trained end-to-end as a single model. The images are first passed into a Convolutional Neural Network (CNN) to extract their features, denoted V_1 and V_2, in a general embedding space. To determine which semantic subspace is relevant to their comparison, both visual features are passed into the condition weight branch, which is a simple neural network. The output of the weight branch is a feature vector of dimension M, represented by $W_1, \cdots, W_M$. It performs a dynamic assignment of the similarity condition masks, denoted $C_1, \cdots, C_M$, to the pair of images. Each similarity condition mask C_i has the same dimension D as the visual features and applied as a mask via element-wise product. The masked embeddings are multiplied by the weight vector to produce the final representations E_1 and E_2. These final representations induce a relevant semantic subspace within which the similarity between both images are compared. We note that the subspace of 'color and pattern', shown in the figure, provides an example of possible notions of similarity encoded by the subspaces but we do not actually restrict the types of subspaces learned by the model. The arrows from V_2 and to E_2 are removed to prevent the figure from being too crowded.

to reason comprehensively about the relative similarity between objects. There has been a recent trend of training embedding models conditioned on some given axis of similarity such as the object category (*e.g.*, [36, 27, 35]) in order to learn disentangled representations (*i.e.* illustrated on the top part of Figure 1). This helps simplify complex similarity relationships by allowing the model to focus on only one similarity condition at a time for each semantic subspace. However, by relying on such labels, these approaches cannot generalize to unseen categories and attributes, one of the primary advantages of embedding models. As such, we seek to learn multiple notions of similarity jointly without explicit supervision via user-defined labels.

In this paper, we aim to learn how to separate the data, where the different similarity conditions and their contribution are treated as a latent variable and learned in a weakly supervised manner. To obtain richer representations of visual similarity, we propose a Similarity Condition Embedding Network (SCE-Net) model that jointly learns multiple similarity conditions from a unified embedding space. An illustrative overview of our model is provided in Figure 2. To begin, images are projected into a unified embedding space using a convolutional neural network. The core component of our model is a set of parallel similarity condition masks, denoted as $C_1, \cdots, C_M$ in Figure 2. These masks are applied on the image features in the general embedding space. By reweighting dimensions that are relevant to a specific notion of similarity, each similarity condition mask is encouraged to learn representations that encode different semantic subspaces. The relevance of each condition mask to the objects is determined by a weight branch conditioned on their visual representations in the unified embedding space. The condition weight branch can be thought of as a type of attention mechanism [42] that performs a dynamic assignment of each condition mask to the objects being compared.

Our work on learning disentangled representations is motivated by the Conditional Similarity Networks (CSN) of Veit *et al*. [36]. The CSN model pre-defined similarity conditions to supervise the learning of disentangled representations. Our model attempts to learn such representations without explicit supervision via such pre-defined conditions. Plummer *et al*. [27] found that considering the global similarity between items during training a CSN model produced more human-intuitive embedding spaces in addition to improving performance. Vasileva *et al*. [35] adapted the CSN model to learn type-aware embeddings for modeling outfit compatibility. Another drawback of these approaches is that they exhibit linear ([36, 27]) or quadratic [35] growth in the number of conditions per desired similarity condition. In contrast, we found that we can often achieve better per-

formance with far fewer learned subspaces (*e.g*., Vasileva *et al*. learns 66 conditional subspaces for the fashion compatibility task on Polyvore Outfits whereas we obtain better performance with 5 learned subspaces).

The contributions of our paper are summarized below:

- We propose the Similarity Condition Embedding Network (SCE-Net), which learns richer representations of different notions of similarity from images without explicit category or attribute supervision.
- We demonstrate that SCE-Net generalizes well to novel categories and attributes in zero-shot tasks.
- Most importantly, we demonstrate that a dynamic weighting mechanism is integral in aiding a weakly supervised model to learn representations for different notions of similarity.

We perform extensive experiments over three datasets, Polyvore-Outfits [35], Maryland-Polyvore [12] and UT-Zappos50K [43], where our approach outperforms the state-of-the-art in outfit compatibility prediction, fill-in-the-blank outfit completion and triplet prediction tasks, respectively, without requiring strong supervision (via category or attribute labels) used in prior work at test time.

2. Related Work

Metric Learning. Substantial prior work [40, 5, 11] has focused on measuring similarity between images in a single similarity context. To achieve this, images are typically projected into a general embedding space where the respective distances between objects provide a measure of their relative similarity. One notable shortcoming of this approach is that it does not consider different types of visual features. In response to this, there has been a recent trend of comparing images across multiple axes of similarity. As discussed in the introduction, several papers have proposed methods that attempt to learn disentangled representations which capture the different notions of similarity via supervision by predefined similarity conditions [36, 28, 35]. However, since these approaches are trained to only compare items along a known axis of similarity, they cannot make predictions between novel categories at test time. Our idea of overcoming this restriction by using a similarity condition weight branch is similar to the phrase localization approach utilized by Plummer *et al*. [28]. However, their work is primarily focused on measuring similarity between image regions and text, and their conditions are also supervised by text descriptions. Learning distance metrics has also attracted a lot of interest from the computer vision community. Hsieh *et al*. [14] utilizes collaborative filtering with implicit feedback to learn a joint metric which encodes user-user and user-item similarity while Sohn *et al*. [34] introduce a multi-class N-pair loss objective to improve deep metric learning.

Visual Attributes. Visual attributes (*e.g*. color and pattern) entail a lot of information and have been shown to be an effective mode of communication between both humans and artificial agents [7, 2]. For example, Batra *et al*. [2] seeks to improve the performance of agents by using visual attributes as their main mode of communication. Attributes have also been used to address tasks such as image search and classification [19, 18] and scene understanding [31, 25, 20]. However, one major limitation that researchers often face is the sparsity of supervision (*i.e*., a lack of example images and/or labels). To address this, Yu *et al*. [44] trains attribute ranking models on synthetic images to determine the relevance of each attribute for the comparison of a pair of images. Others focus on ways of automatically discovering attributes in images [3, 30, 39, 9]. For example, Ferrari *et al*. [9] introduce a probabilistic generative model of visual attributes as well as an approach to learn its parameters from images.

Recommendation and Retrieval. Similarity learning has also been used extensively to solve computer vision problems in other domains such as fashion and retail (*e.g*., [12, 35, 37]). Using visual attributes is a naturally intuitive way to describe fashion items (*e.g*. color, cut and style). As such, identifying relevant attributes in visual representations of fashion items is essential to reasoning about similarity between them. The deficiency of comparing images by projecting them into a general embedding space as described above is especially apparent in prior work on modeling fashion outfit compatibility [21, 12, 35, 37]. In their approach, Veit *et al*. [37] do not distinguish items by their types but instead attempt to learn the concepts of compatibility and similarity from heterogeneous dyadic co-occurrences of items in user data. These visual attributes also form the basis of many interactive fashion search engines and recommendation systems [45, 1, 15, 27].

3. Similarity Condition Embedding Network

In this section we describe SCE-Net, our model which jointly learns representations for the different similarity conditions that may be present in a dataset by treating them and their contributions as a latent variable. This allows us to train our end-to-end model in a weakly supervised manner where we only know if a pair of images are similar under some unknown condition. To begin, the images are projected via a CNN into a common feature space which we term as the general embedding space. We denote this operation as $g(\mathbf{x}; \theta)$ where $\mathbf{x}$ and θ represent the sets of images and parameters respectively. Our network consists of two components - a set of parallel similarity condition masks which we will discuss in Section 3.1, and a condition weight branch we will discuss in Section 3.2. We discuss variants of our condition weight branch with inputs of different modalities in Section 3.3.

3.1. Learning Similarity Conditions

A core component of our model is a set of M parallel similarity condition masks of dimension D, denoted as $C_1, \cdots, C_M$ in Figure 2. The value of M is determined experimentally using held out data. The similarity condition masks are applied, via elementwise product, to the image features in the general embedding space and their bearing on a similarity relationship is learned during training. By re-weighting relevant dimensions, the similarity condition masks are projecting the image features into secondary semantic subspaces of $\mathbb{R}^D$ which encode different similarity substructures. For each similarity condition mask C_j and general image feature V_i , the masking operation is performed as follows:

$$E_{ij} = C_j \odot V_i, \tag{1}$$

where E_{ij} is the masked embedding and $\odot$ denotes the Hadamard product. The output of the masking operation over all similarity condition masks and an image feature v_i is a matrix of dimensions $M \times D$. Let O represent the output of the masking operation where $O = [E_{i1}, \cdots, E_{iM}]$. Then, the final representation for the image feature is computed as a matrix-vector multiplication operation:

$$E_i = wO^T, \tag{2}$$

where w is the weight vector of dimension M computed by the condition weight branch described below.

3.2. Condition Weight Branch

Instead of pre-defining a set of similarity conditions, we use a condition weight branch to allow the model to automatically determine what concepts to learn. The condition weight branch determines the relevance of each condition mask based on the pair of objects being compared. For a pair of images x_i and x_j, the input feature to the condition weight branch is computed as follows:

$$y = concat\{V_i, V_j\}, \tag{3}$$

where $concat\{...\}$ denotes the concatenation operation. As seen in Figure 2, after concatenating these image features they are fed into a series of fully-connected and ReLU layers. A softmax is used on the final activations resulting in a vector w of dimension M that is used to determine the relevance of each similarity condition mask to the objects being compared.

A triplet loss is a naturally intuitive way to learn representations with complex similarity relationships. We define a triplet of objects as a set $\{x_i, x_j, x_k\}$ where x_i is the reference object and x_j and x_k are positive and negative objects that have been determined by the oracle to be semantically similar and dissimilar to x_i under some unobserved condition c, respectively. In the context of this work, an oracle is defined to be a general entity that has the ground truth measures of similarity between all objects under the set of all possible similarity conditions. Usually, the oracle takes the form of crowd-sourced datasets that are annotated with human labels. The final triplet loss is then given as:

$$l_{triplet}(x_i, x_j, x_k) = max\{0,\ d(E_i, E_j) - d(E_i, E_k) + \mu\}, \tag{4}$$

where $d(E_i, E_j)$ denotes the Euclidean distance between the representations of objects x_i and x_j and the margin μ is a hyper-parameter. The triplet loss requires that $d(E_i, E_j)$ is smaller than $d(E_i, E_k)$ by a margin μ where the final image representations E are computed as described above.

As in Veit *et al*. [36], we impose an l_1 loss on the similarity condition masks to encourage sparsity and disentanglement. In addition, we regularize the learned image representations $g(\mathbf{x}; \theta)$ with an l_2 penalty. As such, the final objective function for our model is given by:

$$l_{final} = l_{triplet}(\mathbf{x}) \ + \lambda_1 \, l_1 + \ + \lambda_2 \, l_2, \tag{5}$$

where λ_1 and λ_2 are scalar hyperparameters.

3.3. Multimodal Variants of SCE-Net

In addition to the vision-only version of the condition weight branch used in our network, we also experiment with variations which leverages multimodal features that may provide some semantic relationship between the different conditions we wish to learn. These variants are:

Text Features. We use the word 'text' to refer to both sentences which may represent either the category labels or natural language descriptions of the images. A sentence is tokenized and each token is represented using a pre-trained word embedding (*e.g*., [26, 24]). For a pair of text features (T_i, T_j) corresponding to image pair (x_i, x_j), the input feature to the condition weight branch is computed according to the formulation above:

$$y = concat\{T_i, T_j\}. \tag{6}$$

Visual-Text Features. For a pair of image features (V_i, V_j) and their text features (T_i, T_j), the condition weight branch determines the relevance of each condition embedding based on the input feature:

$$y = concat\{(V_i \odot T_i),\ (V_j \odot T_j)\}. \tag{7}$$

We note that there are different ways to combine visual and text features such as concatenation and projections of both modalities into the same embedding space but elementwise product performed best in our experiments.

4. Experimental Analysis

We evaluate the capability of the SCE-Net model to capture different notions of similarity as well as how well it

generalizes to novel image categories that are not seen during the training process. To provide a fair comparison [2] of our approach to other baseline models, we perform experiments on the Maryland-Polyvore [12], Polyvore-Outfits [35] and UT-Zappos50k [43] datasets. The Maryland Polyvore and Polyvore Outfits datasets contain two evaluation tasks - outfit compatibility prediction and fill-in-the-blank (FITB). For outfit compatibility prediction, the task is to evaluate the compatibility of a set of fashion items in an outfit. As in Han *et al*. [12], performance on this task is evaluated with the area under a receiver operating characteristic curve (AUC). In the FITB experiment, given a set of candidate items and a subset of items in an outfit, the task is to select the most compatible candidate. The effectiveness of the model is evaluated based on the overall accuracy. Although using a larger final embedding has shown to have performance benefits (*e.g*., [12, 35]), this comes at a higher computational cost at test time. We compare methods with the same final embedding size for a fair comparison. We also evaluate the ability of our model to identify different relative strengths of attributes using the task triplet prediction of [36] on the UT-Zappos50k dataset. We note that the level of supervision indicated in Tables 1 and 5 refers to the amount of supervision required by ours and baseline models during test time (i. e. the models know explicitly which axis of similarity to compare the objects on).

4.1. Datasets

Maryland Polyvore [12]. This dataset collected 21,799 outfits from the social commerce website Polyvore. We use the outfit splits provided by the authors consisting of 17,316 outfits in the training set, 3,076 in the test set and 1,407 in the validation set. In the test set provided by the authors, negatives in both the compatibility prediction and FITB tasks are sampled at random without consideration for item compatibility or category (*i.e*. they could replace a "top" in an outfit with "sunglasses"). As such, we evaluate our model on a much more challenging test set provided by Vasileva *et al*. [35], where the item category is taken into account when sampling for negatives.

Polyvore Outfits [35]. This dataset is much larger than Maryland Polyvore, containing 53,306 outfits for training, 10,000 for testing and 5000 for validation. It is also sourced from the Polyvore website, but unlike the Maryland Polyvore dataset, it contains annotations for fine-grained item types and provides a text description of items.

UT-Zappos50k [43]. This dataset contains 50,000 images of shoes with meta-data labels for annotations. We use the triplets provided by Veit *et al*. [36] which are sampled based on four similarity conditions - type of the shoes, gender of the shoes, height of the shoe heels and the closing mechanism of the shoes. Veit *et al*. generated 200k train, 20k validation and 40k test triplets for each characteristic. When training SCE-Net, we combine all the triplets from each characteristic into a single training set.

[2]Recently, [6] proposed a fashion compatibility model and evaluated on the Maryland Polyvore dataset, but it was published after our submission and thus, should be considered concurrent work. In addition, they use a larger base network, ResNet-50 (theirs) vs. ResNet-18 (ours); we omitted their results since they are not directly comparable.

4.2. Implementation Details

Maryland Polyvore and Polyvore Outfits. For fair comparison, we adopt the implementation as detailed in Vasileva *et al*. [35]. We use an 18-layer deep residual network [13] as a shared feature extractor that has been pretrained on ImageNet [8] and fine-tuned during training on this task. The features in the unified embedding space have an embedding size of 64 dimensions. To represent the text descriptions, we also use the HGLMM Fisher vectors [17] of word2vec [24] which have been PCA reduced to 6000 dimensions. Vaslieva *et al*. also took advantage of additional regularizers on their general embedding space (*i.e*., the output of $g(\mathbf{x};\theta)$ discussed in Section 3) which helped improve performance. These include:

- **VSE**: Visual-semantic loss which requires that an image x_i is embedded closer to its description t_i as compared to the other two images within a triplet.
- **Sim**: A loss which encourages similar images to embed nearby each other (analogously, similar text descriptions should also embed nearby each other).

For our experiments on both of these datasets, we included the VSE and Sim losses into our objective function. As such, our objective function becomes:

$$l_{final} = l_{triplet}(\mathbf{x}) \; + \lambda_1 \, l_1 \; + \lambda_2 \, l_2 + \lambda_3 \, l_{VSE} + \lambda_4 \, l_{Sim}, \tag{8}$$

where λ_3 and λ_4 are scalar hyperparameters. We use the same settings as Vasileva *et al*. for learning rates and hyperparameters for loss functions.

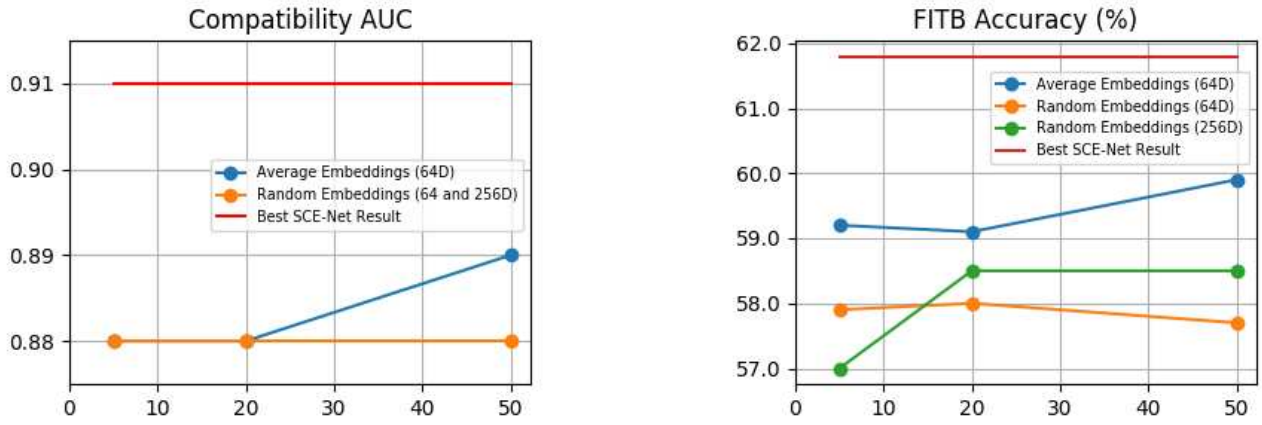
UT-Zappos50k Dataset. An 18-layer ResNet is also used as our base image encoder on this dataset. Due to the triplet format of the dataset, we modify the weight branch to be conditioned on all three images in a triplet. Given a triplet $\{x_i, x_j, x_k\}$, the input to the condition weight branch (at both train and test time) is given as,

$$y = concat\{V_i, V_j, V_k\}, \tag{9}$$

where V_i, V_j and V_k are the representations of images x_i, x_j and x_k respectively. In Section 4.3.2, we demonstrate that conditioning the weight branch on triplet visual representations helps our model to learn the different notions of similarity explicitly defined in the dataset.

Method	Test-time Supervision	Polyvore Outfits		Maryland Polyvore	
		Compat AUC	FITB Acc	Compat AUC	FITB Acc
Siamese Net [35]	None	0.81	52.9	0.85	54.4
Type-Aware Embedding Network [35]	Strong	0.86	55.3	0.90	59.9
SCE-Net	None	**0.91**	**61.6**	**0.90**	**60.8**

Table 1: Comparison of different methods on the outfit compatibility prediction and fill-in-the-blank tasks over the test set for Maryland Polyvore and Polyvore Outfits datasets.



(a) Compatibility AUC results

(b) FITB Accuracy results

Figure 3: We report results of our model with average embeddings and random embeddings of 64 and 256D on the Polyvore-Outfit test set. The values on the x-axis represent the number of similarity condition masks used in SCE-Net. In both plots, the red line denotes the best result obtained by our SCE-Net model for comparison.

4.3. Results

4.3.1 Polyvore Outfits and Maryland Polyvore

Table 1 reports performance on the compatibility prediction and fill-in-the-blank tasks for the Maryland Polyvore and Polyvore Outfits datasets. Across both datasets, our model obtains consistent improvements in both tasks over prior work. In particular, our approach outperforms the state-of-the-art Type-Aware Embedding Network [35] by 5% and 6.3% on the compatibility prediction and FITB tasks respectively, demonstrating that it can better capture the compatibility relationship between items without requiring the type of each item being compared at test time. In addition, we perform better using only 5 similarity conditions, whereas [35] learns 66 similarity conditions for Polyvore Outfits.

To show that our condition weight branch provides meaningful assignments, we compare to making random assignments of image pairs to conditions in Figure 3. We also compare to averaging the embeddings, which demonstrate that the additional parameters (from using multiple conditions for each image pair) in our approach cannot account for most of the improvements we see over [35] (which uses a single condition for each pair). The significant performance gap between SCE-Net and the average or random embeddings demonstrate that our dynamic weighing mechanism is integral to achieving good performance. We also show how the number of similarity conditions affect performance in Table 2, where we find that optimal performance can be obtained using only a few similarity conditions (*e.g.*, 5 for Polyvore Outfits).

Number of Conditions	Compat AUC	FITB Accuracy
1	0.86	53.2
2	0.90	59.7
5	**0.92**	**62.1**
10	0.91	60.8
20	0.89	59.7

Table 2: Ablation studies on how the number of similarity condition masks effect the performance of our model on the validation set of Polyvore Outfits.

To evaluate the capability of our model to generalize to unseen categories based on visual features alone, we remove fashion items that belong to *scarves* and *accessories* categories from the training set. We selected these two categories because they are generally not an essential part of

Unseen Categories (FITB Accuracy)		
Method	Scarves	Accessories
Number of questions	144	248
Siamese Net	46.62	50.82
SCE-Net	**59.46**	**56.55**

Table 3: Comparison of different methods on a subset of FITB questions from the Polyvore-Outfits test set where the candidate choices belong to categories that are unseen during training.

Variants of Condition Weight Branch		
Number of Conditions	Compat AUC	FITB Accuracy
Labels	0.90	60.8
Visuals	0.91	61.6
Visual-Labels	0.90	61.5

Table 4: Results on the Polyvore-Outfit test set obtained by variants of the SCE-Net model with input features of different modalities into the condition weight branch.

outfits and appear in fewer outfits than other categories in the training set. For evaluation purposes, we extract FITB questions from the test set where the candidate choices belong to the removed categories. As a baseline comparison, we train a Siamese network based off the model used by Vasileva *et al*. on the modified training set. The results for both models are reported in Table 3. Our model outperforms Siamese Net by a significant margin in both categories, demonstrating the ability of our model to generalize well to novel categories and attributes.

The performance of our multimodal variants are shown in Table 4. Surprisingly, using the language features of the items' labels alone leads to results that are comparable to those obtained by using the visual features of the items. Using a combination of visual and language features does not lead to a performance gain. However, this could be due to that fact that the language features of item labels do not contain much semantic information. It is possible that we can observe a larger improvement if the language features for the items' descriptions are used instead. However, not all items in this dataset contain a corresponding description.

4.3.2 UT-Zappos50K

We evaluate the effectiveness of our approach on the task of triplet prediction against the strongly-supervised CSN model of Veit *et al*. [36]. Recall that the test set is divided into 4 similarity conditions. In particular, during inference, Veit *et al*. evaluates each triplet with the query $\{x_i, x_j, x_k, c\}$ to determine if the distance between x_i and x_k is smaller than that of x_i and x_k under the similarity condition c. Such explicit supervision during evaluation provides their model with an unfair advantage as compared to our proposed SCE-Net which isn't provided the similarity condition being compared.

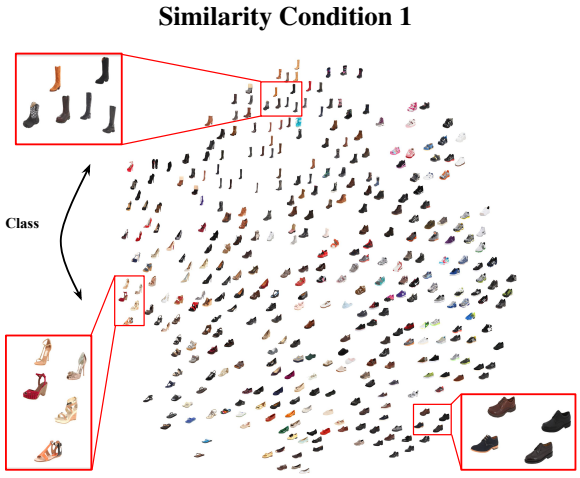
	Method	Error Rate	Test-time Supervision
(a)	CSN fixed disjoint masks [36]	10.79%	Strong
	CSN learned masks [36]	10.73%	Strong
(b)	SCE-Net (2)	11.12%	None
	SCE-Net (3)	8.48%	None
	SCE-Net (4)	**7.53%**	None

Table 5: Results on the UT-Zappos50K test set. **(a)** contains the results reported in prior work [36] and **(b)** reports the results of our model. Numbers in parenthesis indicate the number of similarity condition masks used.

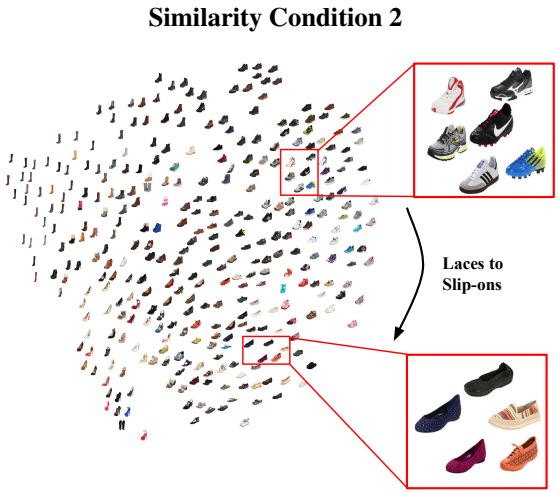
Table 5 shows that when using the concept weight branch to combine our weakly supervised conditions SCE-net outperforms the CSN model, which is provided the exact condition being compared, by approximately 3.2% when using the same number of learned conditions (*i.e*., 4). Reducing the number of learned conditions by 1 for our model, we still outperform the CSN model by 2%. This suggests that it is beneficial to not limit the learning of a notion of similarity to a single subspace. Instead, using a weighted combination of semantic subspaces encourages a model to learn better representations for a similarity condition. In addition, the number of similarity condition masks required for optimal learning increases with the number of similarity conditions present in the dataset.

4.4. Visualizations of Learned Subspaces

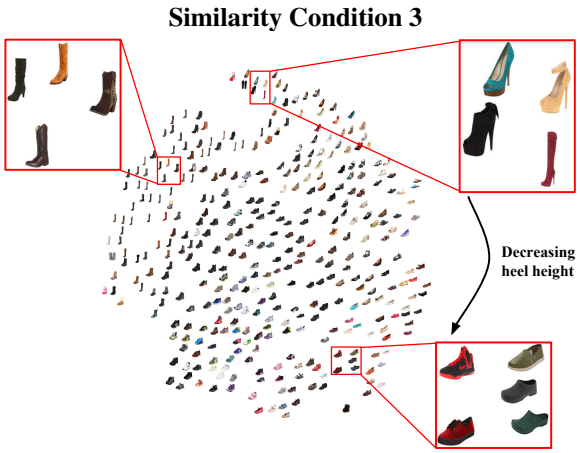
To gain insights into the conditions learned by our model, we provide t-SNE [23] visualizations for all similarity condition masks of learned subspaces for the UT-Zappos50k dataset in Figure 4. The first similarity condition mask learns to differentiate shoes based on their class (*e.g*. boots and high-heels). As we move from the top of the visualization in Figure 4b to the bottom, we can clearly see that the closing mechanism of the shoes gradually changes from laces to slip-ons. Figure 4c displays a subspace that learns the differences in the heel height. In this case, the heel height of the shoes is decreasing from the top of the embedding space to the bottom. From Figure 4d, we see the fourth similarity condition mask has learned to differentiate shoes based on the targeted gender. Women's shoes are embedded at the top of the subspace while men's shoes are mostly embedded at the bottom. This demonstrates that even with just weak supervision during training time, our approach is capable of learning visually-relevant similarity conditions that are explicitly defined in the dataset.



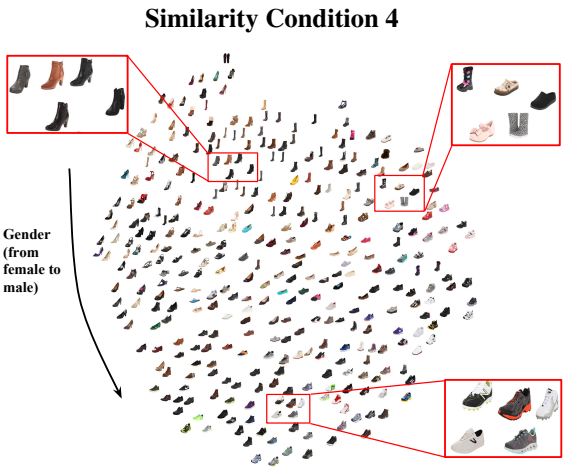
(a) The visualization suggests that shoes are differentiated by class (*e.g.* boots and slippers) in this subspace.



(b) Shoes at the top of this subspace generally have laces while the shoes at the bottom are generally slip-ons, demonstrating that this similarity condition has learned to differentiate between closing mechanisms.



(c) The visualization suggests that shoes are differentiated by heel height in this subspace. The heel height of the shoes decreases as we go from the top to the bottom of the subspace.



(d) The visualization suggests that shoes are differentiated by gender in this subspace. Women's shoes are embedded at the top of the subspace, and men's shoes at the bottom.

Figure 4: Visualizations of the semantic subspaces encoded by our 4 similarity condition masks on the UT-Zappos50k dataset.

5. Conclusion

In this work, we propose an approach that treats the different similarity conditions and their contributions as a latent variable and attempts to learn them in a weakly supervised manner. SCE-Net removes the need for strong supervision via pre-defined similarity conditions by using a condition weight branch conditioned on visual representations of images to determine the context relevance of each similarity condition mask. We demonstrate that our model not only outperforms strongly supervised methods but also generalizes well to novel image categories and attributes.

We show that a dynamic weighting mechanism is essential in training a weakly supervised model to learn different notions of similarity. In particular, our results indicate that restricting the learning of a similarity condition to a single subspace can be disadvantageous to the learning capability of the model. Finally, we demonstrate that a weighted combination of semantic subspaces can learn better representations for a similarity condition. One exciting avenue for future work is to learn to determine the optimal number of similarity condition masks in an unsupervised manner.

Acknowledgements: This work is supported in part by DARPA and NSF awards IIS-1724237, CNS-1629700, CCF-1723379.

References

[1] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 388–397, 2017.

[2] Tanmay Batra and Devi Parikh. Cooperative learning with visual attributes. *arXiv preprint arXiv:1705.05512*, 2017.

[3] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010.

[4] Zhi-Qi Cheng, Xiao Wu, Siyu Huang, Jun-Xiu Li, Alexander G Hauptmann, and Qiang Peng. Learning to transfer: Generalizable attribute learning with multitask neural model search. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 90–98. ACM, 2018.

[5] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *null*, pages 539–546. IEEE, 2005.

[6] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12617–12626, 2019.

[7] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2951–2960, 2017.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Advances in neural information processing systems*, pages 433–440, 2008.

[10] Xiaoling Gu, Yongkang Wong, Lidan Shou, Pai Peng, Gang Chen, and Mohan S Kankanhalli. Multi-modal and multi-domain embedding learning for fashion retrieval and analysis. *IEEE Transactions on Multimedia*, 2018.

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[12] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *ACM Multimedia*, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Wei-Lin Hsiao and Kristen Grauman. Learning the latent look: Unsupervised discovery of a style-coherent embedding from fashion images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4213–4222. IEEE, 2017.

[15] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7161–7170, 2018.

[16] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*, pages 1062–1070, 2015.

[17] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015.

[18] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980. IEEE, 2012.

[19] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[20] Li-Jia Li, Hao Su, Yongwhan Lim, and Li Fei-Fei. Objects as attributes for scene classification. In *European Conference on Computer Vision*, pages 57–69. Springer, 2010.

[21] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, 19(8):1946–1955, 2017.

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[25] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012.

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[27] Bryan A. Plummer, M. Hadi Kiapour, Shuai Zheng, and Robinson Piramuthu. Give me a hint! Navigating Image Databases using Human-in-the-loop Feedback. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[28] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *Proceedings of the*

European Conference on Computer Vision (ECCV), pages 249–264, 2018.

[29] Qi Qian, Rong Jin, Shenghuo Zhu, and Yuanqing Lin. Fine-grained visual categorization via multi-stage metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3716–3724, 2015.

[30] Mohammad Rastegari, Ali Farhadi, and David Forsyth. Attribute discovery via predictable discriminative binary codes. In *European Conference on Computer Vision*, pages 876–889. Springer, 2012.

[31] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4657–4666, 2015.

[32] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Data-driven visual similarity for cross-domain image matching. In *ACM Transactions on Graphics (ToG)*, volume 30, page 154. ACM, 2011.

[33] Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*, pages 753–769. Springer, 2016.

[34] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016.

[35] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 390–405, 2018.

[36] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 830–838, 2017.

[37] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015.

[38] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order embeddings of images and language. In *International Conference on Learning Representations (ICLR)*, 2016.

[39] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. Automatic attribute discovery with neural activations. In *European Conference on Computer Vision*, pages 252–268. Springer, 2016.

[40] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[41] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2019.

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

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

[44] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5570–5579, 2017.

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