Semi-Supervised Visual Representation Learning for Fashion Compatibility

Ambareesh Revanur*
ambareesh.r@gmail.com
Carnegie Mellon University
USA

Vijay Kumar
Walmart Global Tech Bangalore
India

Deepthi Sharma
Walmart Global Tech Bangalore
India

ABSTRACT

We consider the problem of complementary fashion prediction. Existing approaches focus on learning an embedding space where fashion items from different categories that are visually compatible are closer to each other. However, creating such labeled outfits is intensive and also not feasible to generate all possible outfit combinations, especially with large fashion catalogs. In this work, we propose a semi-supervised learning approach where we leverage large unlabeled fashion corpus to create *pseudo* positive and negative outfits on the fly during training. For each labeled outfit in a training batch, we obtain a pseudo-outfit by matching each item in the labeled outfit with unlabeled items. Additionally, we introduce consistency regularization to ensure that representation of the original images and their transformations are consistent to implicitly incorporate colour and other important attributes through self-supervision. We conduct extensive experiments on Polyvore, Polyvore-D and our newly created large-scale Fashion Outfits datasets, and show that our approach with only a fraction of labeled examples performs on-par with completely supervised methods.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings**; **Neural networks**.

KEYWORDS

fashion compatibility, semi-supervised learning, self-supervision, product recommendation

ACM Reference Format:
Ambareesh Revanur, Vijay Kumar, and Deepthi Sharma. 2021. Semi-Supervised Visual Representation Learning for Fashion Compatibility. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3460231.3474233

*Work done during an internship at Walmart Global Tech Bangalore.

RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands

ACM ISBN 978-1-4503-8458-2/21/09…$15.00
https://doi.org/10.1145/3460231.3474233

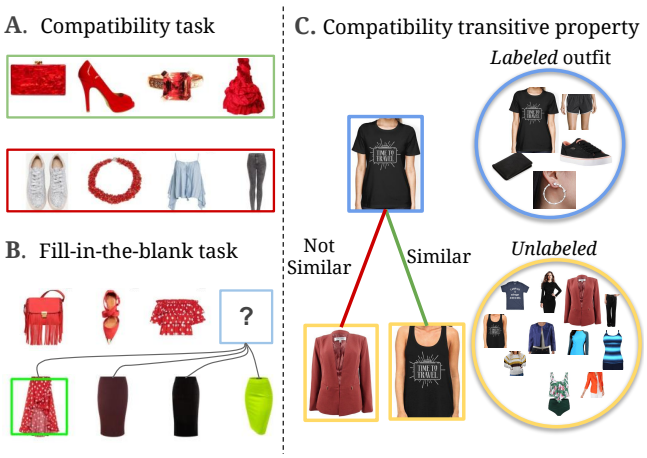

Figure 1: (A) Compatibility task. Compatible (green) and a non-compatible (red) outfit. (B) Fill in the blank task. Given an outfit, the objective here is to pick the most compatible item from the given choices. (C) We generate pseudo-outfits based on visual similarity of labeled outfits with unlabeled examples.

1 INTRODUCTION

Recent advancements in computer vision have led to their several practical applications in fashion such as similar product recommendation [24, 26, 41, 42], shop-the-look [14, 25], virtual try-ons [31, 47] and 3D avatars [27, 36]. In this work, we focus on *fashion compatibility* where the objective is to compose matching clothing items that are appealing and complement well, as shown in the Fig. 1A and Fig. 1B. This could have potential application in online retail industry to recommend complementary products to the user based on their previously purchased product(s). For example, a formal *shoe* can be recommended to a customer who purchased a office-wear *trouser*.

What constitutes a good "complementary" similarity differs from visual similarity. Learning representations for compatibility requires reasoning about color, shape, product category and other high level attributes to ensure that the items from "different" categories are closer. Most existing works [24, 40, 43] achieve this through careful labeling of dataset to identify items that go well together. A simple metric learning approach can then be used to learn an embedding space where the compatible items are brought closer compared to non-compatible ones. However, creating such labeled outfit pairs is expensive, laborious and sometimes require expert knowledge. This also becomes cumbersome and infeasible

even for a reasonably large fashion catalog as one needs to compose several possible outfit combinations. In this work, we aim to learn powerful representation for compatibility task with very limited labeled outfit data. We propose a semi-supervised learning approach to leverage large amount of unlabeled fashion images that can be easily obtained. The approach is based on data augmentation technique [2, 49] where the goal is to enhance the training set through techniques namely - *pseudo-labeling* and *consistency regularization*. Based on the idea that new combinations or associations can be formed from the current associations of the outfit items, we aim to generate *pseudo positive* and *negative* outfit pairs on-the-fly during each training iteration. For example, consider that if an item A is compatible with another item B and if item C is visually similar to item A, it is possible to then create a new outfit with item pairs B and C. For each image in the positive/negative training pairs, we find the most visually similar example in the unlabeled set and create a new pseudo-outfit pairs as shown in Fig. 1C. Thus even with few training outfit collections, it is possible to generate a large stream of pseudo-outfits pairs.

As image attributes such as colour, shape and texture play a big role for compatibility, we additionally introduce self-supervised consistency regularization [2] on unlabeled images to explicitly learn those attributes. For instance, we need to disentangle shape information from our representation as items in an outfit are usually of different shape. Similarly, colour can be very informative. We achieve this by applying random transformation on the images and direct the network to produce (dis)similar representations. Note that this is different from explicit attention mechanism employed by current schemes where they train conditional masks to learn a subspace for different attributes such as color and patterns [24, 40].

We conduct extensive experiments on Polyvore and Polyvore-Disjoint [43] datasets and show that our proposed approach can achieve on-par performance with fully supervised methods even with 5% of labeled outfits. In a fully supervised setting, our approach can achieve state-of-the-art performance on compatibility prediction task. Finally, we create a large scale Fashion Outfits dataset consisting of around 700K outfits with more than 3M images and demonstrate consistent improvement in performance over fully-supervised baseline.

To summarize, we make the following contributions in this paper.

- We propose a semi-supervised approach for fashion compatibility prediction by learning powerful representation with limited outfit labels.
- We introduce consistency regularization and pseudo-labeling based data augmentation techniques to learn different attributes and generate pseudo-outfits, completely from the unlabeled fashion images.
- We demonstrate on-par performance as fully supervised approaches with only a fraction of labeled outfit on Polyvore, Polyvore-D and our newly created datasets.

2 RELATED WORK

In this section, we discuss previous works on compatibility prediction and other related areas.

Fashion Compatibility introduced by Han *et al.*[15] was formulated as a sequential problem and trained a bi-directional recurrent network model that predicts the next compatible item conditioned on previously observed items in the outfit sequence. They also introduced Polyvore dataset in this work. Vasileva *et al.*[43] train pairwise embedding spaces and employ metric learning to learn representations for fashion compatibility. Representations are learnt for different pairs of categories which is not feasible when the catalog of large category types. Further, they enriched the Polyvore dataset by introducing more challenging evaluation sets by filtering out common items across train and test splits. Tan *et al.*[40] learn embeddings by relying on a conditional weight branch on two image representations as an attention mechanism for selecting the subspace. Unlike [43], this work does not require access to the type information during evaluation.

Another additional ingredient of these works [15, 40, 43] is the usage of additional meta-data such as text description. They train a text module using word2vec features [29] and enforce a visual-semantic embedding (VSE) loss to align text and image representations. In our work, we focus on only visual information.

More recently, some works [7, 9, 48] have also exploited item-item relationships and trained a graph convolutional network (GCN) by exploiting didactic co-occurrences of items. Further, Lin *et al.*[24] introduce a fashion retrieval problem for images and learn a type-dependent attention mechanism. Most of the existing literature in fashion compatibility focus on the fully supervised paradigm. In contrast, we address a more practical and challenging semi-supervised paradigm without using additional text data.

Semi-supervised learning has seen lot of progress [2, 46, 49] in the last few years where ever there is a scarcity of labeled data. In the context of deep representation learning, following techniques are widely employed. In consistency regularization [6, 34], output class predictions are forced to remain unchanged for different augmentations of the input data. This provides regularization to the model to achieve good generalization. In our work, we incorporate consistency regularization to explicitly capture appearance and shape attributes for fashion items. However unlike previous approaches, we minimize the distance between original image and its shape transformation and simultaneously maximize the distance between original image and its colour transformation. This ensures that the model learns shape-invariant features and color variant features.

On the other hand, entropy minimization [13, 21] aims to obey cluster assumption by forming low density regions between different classes. Pseudo-labeling [22, 23, 45] is a type of self-training algorithm that assign hard labels to unlabeled examples based on the maximum prediction probability. In our work, we create pseudo-outfits and augment our training dataset to exploit transitivity of items across different outfits. Generative adversarial network (GAN) [11] based approaches have also been proposed for semi-supervised learning, however these are challenging to train [38]. In a related area of few-shot learning, Prototypical networks [39] exploit the simple inductive bias of a classifier by defining a prototype embedding as the mean of the support set in the latent space. We make use of a simple inductive bias that the nearest neighbour in embedding space should have similar attributes to construct pseudo-outfits.

Self-supervised learning literature is broadly based on solving a pre-defined task that exploits the structure present in the data [3, 4, 12, 16], or equivariance [20, 33] and invariance [30] properties of image transformations. For image classification, different pretext tasks such as jigsaw puzzle [32], colorization task [8], rotation prediction [10] have been proposed. It has been shown recently that even simple data augmentation methods [4, 5, 16] such as colour jittering and gray scale transformations can provide good supervision. These approaches employ contrastive loss to learn consistent representation for an image and its augmentation while treating other images in the batch as negative samples. In our work, we explicitly define positive and negative data augmentations and impose a self-supervised consistency loss on them. However, our regularizer does not require access to labels as required by [4, 16] during evaluation.

3 OUR APPROACH

We are interested in learning powerful visual representation for the task of fashion compatibility. This is achieved through metric learning by bringing the embedding of items (from different categories) that go well together in a outfit closer compared to non-compatible items. While the previous works [24, 40, 43] relied mostly on large labeled outfits, in this work, we instead consider only a fraction of labeled outfits and leverage unlabeled fashion items to learn such representations.

Our proposed approach is a data augmentation technique where *pseudo-positive* and *negative* triplet pairs (Fig 2(b)) are generated on-the-fly during each training iteration based on visual similarity of individual examples in the triplet using the images in the unlabeled set. Since color and texture play a key role in determining matching outfits [18], we incorporate an additional self-supervised loss on the individual fashion items (Fig 2(c)) to implicitly capture those attributes. Note that, this is different from explicit attention mechanism employed by current methods where they learn conditional masks that implicitly learns colour and pattern attributes [24, 40].

We next formulate the problem of semi-supervised fashion compatibility and then describe our proposed data augmentation techniques in greater detail.

3.1 Formulation

We are given a dataset $\mathcal{D} = \{\mathcal{D}_l \cup \mathcal{D}_u\}$ for training our model. $\mathcal{D}_l = \{X^1, X^2 \dots X^l\}$ corresponds to the labeled set where each $X^i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_k^i\}$ is an outfit containing more than one fashion item from different categories. Note that, a fashion item $\mathbf{x}_j^i$ can be a part of multiple outfits. Similarly, $\mathcal{D}_u = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^u\}$ is the large unlabeled corpus of individual fashion items that are not part of any outfit. We also denote $\mathcal{X}$ as the set of all fashion images. Our goal is then to learn a mapping function $f_\theta : \mathcal{X} \rightarrow \Phi$ that transforms the fashion images into an embedding space where compatible items are closer to each other. In our case, we use deep convolutional neural network (CNN) as f_θ where θ denotes the parameters of our network.

3.2 Architecture

Our network architecture is similar to previous approaches [43] that are based on siamese network [37] with ResNet18 backbone. To train our model, we need a batch of labeled triplets and unlabeled images. Each triplet consists of anchor (A), positive (P) and negative (N) image pairs as shown in the Figure 2. Anchor and positive images are complementary items that are part of the same outfit but from different categories (e.g. *shirt* and *trouser*) while negative image is chosen randomly from the positive item category. The embeddings ϕ are obtained after the fully-connected layer, similar to [43]. We train our network with triplet based margin loss defined as,

$$\mathbf{max}(0, d(\phi_A, \phi_P) - d(\phi_A, \phi_N) + m) \tag{1}$$

where, A, P and N denote anchor, positive and negative images of the triplet, respectively. $d(\cdot, \cdot)$ is the Euclidean distance function and m is the margin.

Labeled dataset: We directly minimize the above margin loss ($\mathcal{L}_l$) for the triplets sampled from labeled outfits to bring compatible items closer to each other pushing non-compatible items farther away as shown in Fig. 2A. However, in the absence of large labeled dataset, we employ consistency regularization and pseudo-labeling losses on the unlabeled images that are described next.

3.3 Consistency regularization

Based on the hypothesis that learning discriminative representation for fashion compatibility prediction requires reasoning about important attributes such as colour and texture since matched outfits often have similar attributes. At the same time, it is necessary to disentangle shape information as items from different categories usually have different shapes. This becomes especially crucial when training the ImageNet pre-trained networks that learn discriminative shape information [4, 16] for generic image classification. We propose a self-supervised consistency regularization on individual fashion items that enforce these observations on the network. The main difference to existing approaches is that we leverage unlabeled fashion images to explicitly learn such attributes without the need for an attention mechanism [40] or a large labeled outfit dataset [24, 40].

We consider entire fashion item set $\mathcal{X}$ that includes both labeled and unlabeled samples without their outfit association. Given each image in the batch, we apply random transformations for shape and appearance, and then measure the discrepancy between the representations of the original and perturbed images. Unlike [2, 49] that ensures consistent output class distribution for a sample and its transformation, here we enforce the consistency for the representation as network is optimized for distance based loss function. We apply following transformations on the images and visualize a few examples in Fig. 3.

- *Shape transformation ($\mathcal{T}_s$)*: For each image x^i in the mini-batch, we apply random affine transformations such as rotation and shearing. We rotate the image randomly in the range of ±5 degrees and shear the image by utmost 30 pixels along x and y coordinates. We finally scale the images by a maximum of 1.2 times and take the center crop of the image.

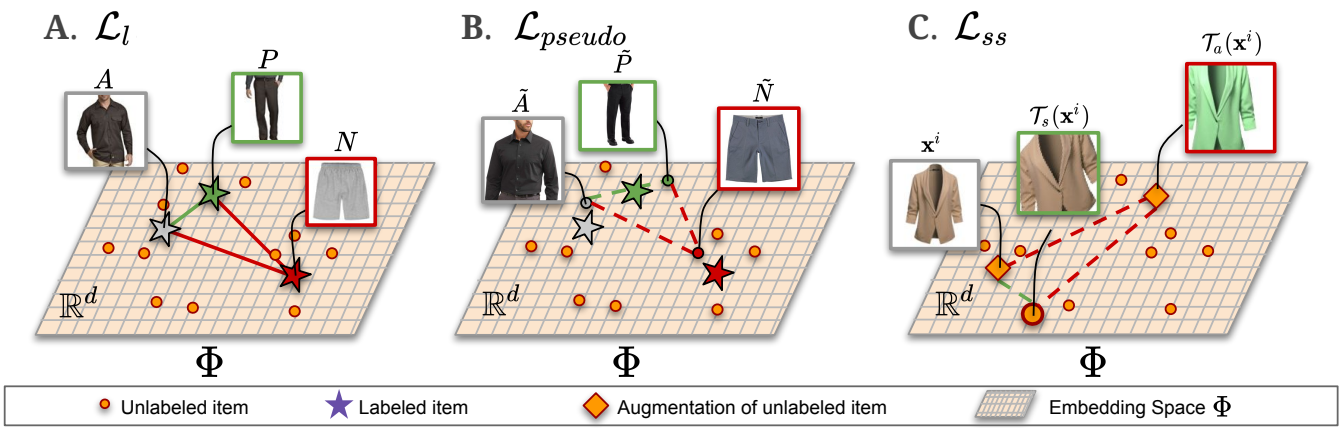

Figure 2: Overview of our proposed approach. A. Triplet loss $\mathcal{L}_l$ on labeled items: Anchor item A, Compatible positive item P and Non-compatible negative item N. B Triplet loss $\mathcal{L}_{pseudo}$ on pseudo-labelled items in the unlabeled item batch $\mathbf{b}_U$. The figure shows visually rich low-dimensional embedding space Φ where we compose nearest pseudo triplets for training. See Sec. 3.4. C. Triplet losses $\mathcal{L}_{ss}$ on *shape* and *appearance* transformed images. See Sec. 3.3

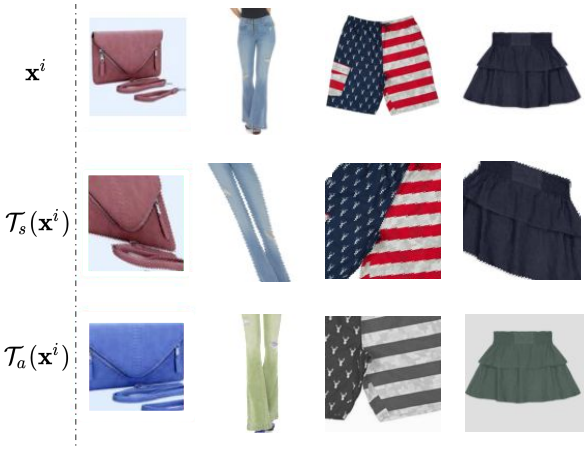

Figure 3: Examples showing different *shape* ($\mathcal{T}_s$) and *appearance* transformations ($\mathcal{T}_a$) of the original images (first column). Refer Sec. 3.3 for details.

Colour and texture remain unaltered. We represent shape perturbed images as $\mathbf{x}^i_{[s]} = \mathcal{T}_s(\mathbf{x}^i)$. See Fig. 3 for examples.

- *Appearance transformation ($\mathcal{T}_a$)*: While it is challenging to modify the colour/texture of the items, we resort to simple transformations such as random gray scale, colour jitter and random cropping to obtain the perturbations $\mathbf{x}^i_{[a]} = \mathcal{T}_a(\mathbf{x}^i)$. Such techniques are previously employed in self-supervised works [4, 16] for representation learning. See Fig. 3 for examples.

Finally, we construct a triplet with original fashion item x^i as anchor, and its shape $\mathbf{x}^i_{[s]}$ and color $\mathbf{x}^i_{[a]}$ perturbed images as positive and negative instances. We then impose a self-supervised loss ($\mathcal{L}_{ss}$) on the augmented triplet $(\mathbf{x}^i, \mathbf{x}^i_{[s]}, \mathbf{x}^i_{[a]})$ using margin triplet loss as shown in Fig. 2C.

3.4 Data Augmentation with Pseudo-Labels

While consistency regularization applied on individual fashion items directs the network what to focus on, we present a labeling strategy that generates *pseudo-outfits* by exploiting the knowledge of fashion items in their vicinity. Pseudo-outfits are the synthetically created outfit pairs created on-the-fly during each training iteration. We again leverage unlabeled images to achieve this.

Algorithm 1 Algorithm of our proposed approach

require: Labeled $\mathcal{D}_l$ and unlabeled $\mathcal{D}_u$ data, margin m, labeled fraction α, λ_{pseudo} and λ_{ss} // $[\cdot]$ is indexing operation
Optimize for model parameters θ

for iteration i in n_{iters} **do**
  // Obtain a batch of fashion items
  sample $\mathbf{b}_l \sim \mathcal{D}_l$ and $\mathbf{b}_u \sim \mathcal{D}_u$.
  // Labeled dataset
  create $(A, P, N) \sim \mathbf{b}_l$, type$(P)$ =type(N)
  $\phi_A, \phi_P, \phi_N \leftarrow f_\theta(A), f_\theta(P), f_\theta(N)$
  compute $\mathcal{L}_l(\phi_A, \phi_P, \phi_N)$
  // Consistency regularization
  $\mathbf{b}_{[s]} \leftarrow \mathcal{T}_s(\mathbf{b}_u), \mathbf{b}_{[a]} \leftarrow \mathcal{T}_a(\mathbf{b}_u)$
  $\tilde{\phi_u}, \tilde{\phi_{[s]}}, \tilde{\phi_{[a]}} \leftarrow f_\theta(\mathbf{b}_u), f_\theta(\mathbf{b}_{[s]}), f_\theta(\mathbf{b}_{[a]})$
  compute $\mathcal{L}_{ss}(\tilde{\phi_u}, \tilde{\phi_{[s]}}, \tilde{\phi_{[a]}})$
  // Find nearest pseudo-triplet in $\mathbf{b}_u$
  pairwise distance $\text{idx}_A \leftarrow \textbf{argmin}(d(\phi_A, \tilde{\phi_u}))$. Similarly, compute idx_P, idx_N
  $\tilde{\phi}_A, \tilde{\phi}_P, \tilde{\phi}_N \leftarrow \tilde{\phi}_u[idx_A], \tilde{\phi}_u[idx_P], \tilde{\phi}_u[idx_N]$
  compute $\mathcal{L}_{pseudo}(\tilde{\phi}_A, \tilde{\phi}_P, \tilde{\phi}_N)$
  // Minimize the final objective
  update θ by minimizing $\mathcal{L}$ (Eq. 2)
end for

We draw our motivation from few-shot learning [39] and argue that nearest neighbour should have similar attributes at a *thematic*

level (e.g. different office-wear items to be closer to each other compared to travel-wear items) due to the inductive bias of the network instilled by $\mathcal{L}_l$ and $\mathcal{L}_{ss}$. Thus given a triplet with anchor, positive and negative items from the labeled outfit, we create a pseudo-triplet $(\tilde{A}, \tilde{P}, \tilde{N})$ by finding nearest element for each of these items in the embedding space. As it is computationally infeasible to perform nearest neighbor search on entire unlabeled dataset $\mathcal{D}_U$, we randomly sample sufficiently large mini-batch of unlabeled images, $\mathbf{b}_u$, to generate pseudo-outfits and finally impose margin loss ($\mathcal{L}_{pseudo}$) on the pseudo-triplets $(\tilde{A}, \tilde{P}, \tilde{N})$ as shown in Fig. 2B.

3.5 Loss function

We formalize our algorithm in Algo. 1. We minimize the following objective function to train our model

$$\mathcal{L} = \mathcal{L}_l + \lambda_{ss}\mathcal{L}_{ss} + \lambda_{pseudo}\mathcal{L}_{pseudo} \tag{2}$$

where $\mathcal{L}_l$, $\mathcal{L}_{pseudo}$ and $\mathcal{L}_{ss}$ are the triplet margin losses on the labeled, pseudo-outfits and individual instances, respectively. λ_{ss} and λ_{pseudo} are the hyper-parameters.

4 DATASETS

Polyvore outfits [43] and Polyvore disjoint [43] are the two primary datasets used in the literature for evaluating fashion compatibility. We conduct our ablations and comparisons with previous state-of-the-art approaches on these datasets. In addition to these, we baseline our results on a newly created fashion dataset to provide large scale evaluation. We provide the statistics of these datasets in Table 1.

Polyvore outfits is a crowd-sourced dataset collected from Polyvore website where users upload fashion photos and organize them into outfits by associating fashion items that go well together. Each outfit consists of product images along with their metadata such as item IDs, product name, fine-grained item type, title and semantic category. There are 12 semantic categories such as tops, bottoms, outerwear, shoes *etc.*. There are about a total of 68,306 outfits split into train, valid and test splits as shown in Table 1. In our work, we demonstrate visual representation learning using only α% of train split as labeled dataset $\mathcal{D}_L$ and consider rest of the outfit images as unlabeled items $\mathcal{D}_U$. We report our final results on the entire validation and test set as done in previous works.

Polyvore-Disjoint dataset is a subset of Polyvore dataset consisting of around 32,140 outfits. It is created by filtering out training outfits that have common items with validation and test outfits. This ensured that train and test outfits have mutually exclusive fashion items for realistic evaluation. Similar to Polyvore dataset, we use only α% of the training split as $\mathcal{D}_L$ and use all remaining items as $\mathcal{D}_U$.

Fashion outfits is a proprietary dataset created based on the user purchase transactions on an e-commerce platform. Our collection process makes a reasonable assumption that multiple fashion items from different categories that are purchased by an user in a single session are complementary and go well together. Only those fashion categories that are similar to high-level categories defined in the Polyvore dataset are considered. Based on the user purchase history over a period of time, we retained user sessions (a) with purchases from more than one category, (b) with uniquely purchased items and (c) that do not have multiple items from the same category. We finally apply association mining algorithm [1] on these subset to retain highly frequent co-occurring items and remove any noisy transactions.

Table 1: Statistics of Polyvore, Polyvore-D and newly created fashion dataset in terms of number of outfits in train, validation and test splits. We also mention overall fashion items in these datasets. Our dataset has ~10 times more outfits than existing datasets.

Dataset	Train	Validation	Test	#items
Polyvore	53K	10K	5K	365K
Polyvore-D	17K	-	15K	175K
Fashion Outfits	675K	10K	20K	3M

Overall, the dataset consists of 705K outfits with more than 3M images. We randomly split the dataset into 675K train, 10K validation, and 20K test outfits. In our experiments, we primarily intend to use the train set as an unlabeled set to demonstrate the efficacy of our proposed approach with unlabeled examples.

5 EXPERIMENTS

5.1 Implementation Details

We use the same implementation procedure as [40, 43] and modify ImageNet-pretrained ResNet-18 [17] as our backbone. We consider a batch size of 256 for labeled set where each sample within a batch consists of anchor, positive and negative images. For the unlabeled set, we consider 1024 individual fashion items. The triplet loss margin m is set to 0.4. We determine the values of λ_{ss} and λ_{pseudo} empirically and set them to 0.1 and 1, respectively. The network is optimized with Adam [19] optimizer with a learning rate of 5e-5. The network is trained for 10 epochs and the best results are reported based on a validation set. Our implementation is done in PyTorch [35] and trained on Nvidia-V100 GPU machine with 16GB memory.

5.2 Evaluation Tasks

Fill-In-The-Blank (FITB) is a question and answering task in which model is presented with an incomplete outfit along with four candidate items as possible answers. The task is then to choose a candidate that is most compatible with the given outfit. As done in [43], we measure the pairwise cosine similarity between the candidate embedding and the average embedding of the outfit and choose the one with highest similarity. The performance is reported as overall accuracy on this task.

Compatibility prediction is a binary prediction task where model has to predict whether all the items in a given outfit are compatible or not. We calculate the average pairwise distance between items in the outfit and report the performance as area under the receiver operating curve (AUC).

Table 2: Comparison of our approach against previous state-of-the-art methods. Red color denotes that the configuration is less suitable in low-data regime. We compare our method against certain fully supervised methods that use additional supervision such as text embeddings and type-specific supervision. Underlined values indicate best reported fully supervised results on the dataset. We report FITB accuracy and Compatibility AUC. *Higher is better*. See Sec. 5.3 for more analysis. $\mathcal{D}$ indicate additional unlabeled data from Fashion outfits dataset used for training. Best viewed in color.

Method	Labels α%	Uses text labels	Explicit type conditioning	Polyvore-D dataset		Polyvore dataset	
				FITB Acc.	Comp. AUC	FITB Acc.	Comp. AUC
Baselines							
Color attribute	-	×	×	39.2	0.68	41.0	0.71
Siamese Network [43]	5%	×	×	47.0	0.77	50.3	0.78
Fully Supervised Methods							
Siamese Network [43]	100%	×	×	51.8	0.81	52.9	0.81
Bi-LSTM +VSE [15]	100%	✓	×	39.4	0.62	39.7	0.65
CSN T1:1 [44]	100%	×	✓	52.5	0.82	54.0	0.83
CSN T1:1 + VSE [44]	100%	✓	✓	53.0	0.82	54.5	0.84
Type-Aware [43]	100%	✓	✓	55.2	0.84	56.2	0.86
SCE-Net (avg) [40]	100%	✓	×	53.6	0.82	59.1	0.88
CSA-Net [24]	100%	×	✓	59.3	0.87	63.7	0.91
Ours	100%	×	×	54.6	0.84	57.9	0.89
Ours - Semi Supervised Approach							
Ours	5%	×	×	51.4	0.81	54.7	0.86
Ours + $\mathcal{D}$	5%	×	×	**51.5**	**0.82**	**54.9**	**0.86**

5.3 Results

Baseline methods. To validate our hypothesis that color plays an important role for compatibility learning, we report our baseline result with color histogram features. As shown in Table 2, these simple features perform remarkably well and achieve results on par with some deep architectures (Bi-LSTM approach [15]). This motivates us to explicitly encode colour information in our representation from unlabeled images. Another baseline is a siamese network with triplet loss as reported in [43]. Note that these baselines do not have include any other meta data such as text or label information.

Comparison with state-of-the-art. We make comparisons with previously reported approaches on Polyvore datasets in Table 2. All these approaches rely on a fully supervised outfit dataset for representation learning. We also specify whether these approaches rely on any additional metadata information such as category and title. It is clear from the table that our approach achieves on par performance compared to fully supervised methods with only a fraction of labeled outfits and does not require any additional metadata information.

Result on Fashion outfits. We report our large-scale tests on Fashion outfits in Fig. 4 C. Our method obtains scores of 0.87/57.6 on Compat AUC/FITB tasks while [43] obtains scores of 0.83/55.0. This demonstrates that our method works well even on large benchmarks. Due to non-availability of code for some methods, we report the result only for [43].

5.4 Ablation studies

Consistency regularization and pseudo-labels. We first conduct our ablation study to understand the effectiveness of $\mathcal{L}_{ss}$ and $\mathcal{L}_{pseudo}$. We include different regularization terms to the baseline siamese model and report the results in Table 3. It is evident that both these forms of regularization on unlabeled data have complementary benefits and improve the overall performance significantly achieving results on par with supervised methods.

Table 3: Ablation studies on Polyvore and Polyvore-D datasets indicating the performance of different components of our model.

Model	Polyvore-D		Polyvore	
	FITB	Comp.	FITB	Comp.
Siamese $\mathcal{L}_l$	47.0	0.77	50.3	0.78
$\mathcal{L}_l + \lambda_{ss}\mathcal{L}_{ss}$	49.2	0.79	53.7	0.82
$\mathcal{L}$ (Eqn 2)	51.4	0.81	54.7	0.86

Batch size. To create good quality pseudo-outfits, we need to have reasonably large unlabeled image batches to ensure better quality matches that share similar attributes for items in the labeled triplet [4, 23]. To evaluate this, we conduct an experiment with different unlabeled bath sizes and report the results in Fig 4B. Results indicate that increasing the batch size consistently improves the performance of our model. However, we could not go beyond a batch size of 1024 due to GPU memory limitations.

How many labeled outfits are enough? Since it is challenging to annotate outfit pairs, an important question we seek to answer is the number of labeled outfits that are sufficient for learning good representations. For this study, we consider a proportion of the Polyvore dataset as labeled and consider remaining outfits as unlabeled in each experiment. As expected, performance improves by increasing the labeled set as shown in Fig 4A. As depicted in the figure, our semi-supervised method (with α = 5%) outperforms

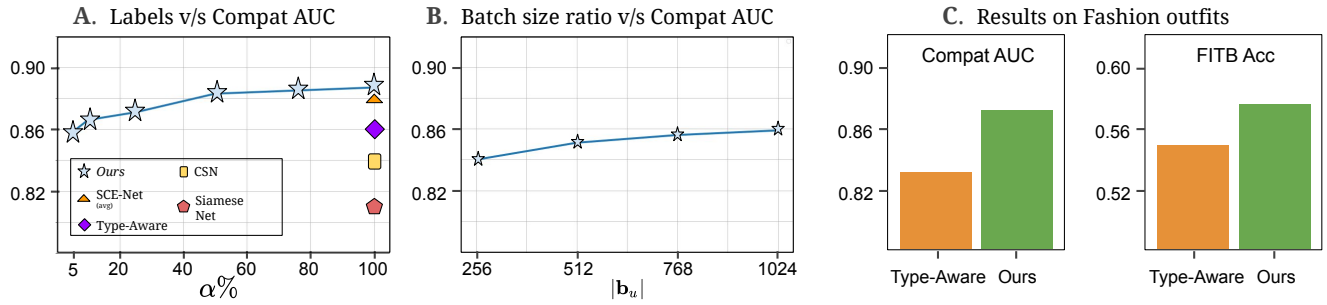

Figure 4: A. Performance of our method with different proportion of training labels (α%) measured by Compatibility AUC on Polyvore dataset. B. Performance of our method with different unlabeled batch size $|b_u|$ measured by Compatibility AUC on Polyvore dataset. See Sec. 5.4 C. Results of our method and Vasileva *et al.* [43] on Fashion Outfits dataset. See Sec. 5.4.

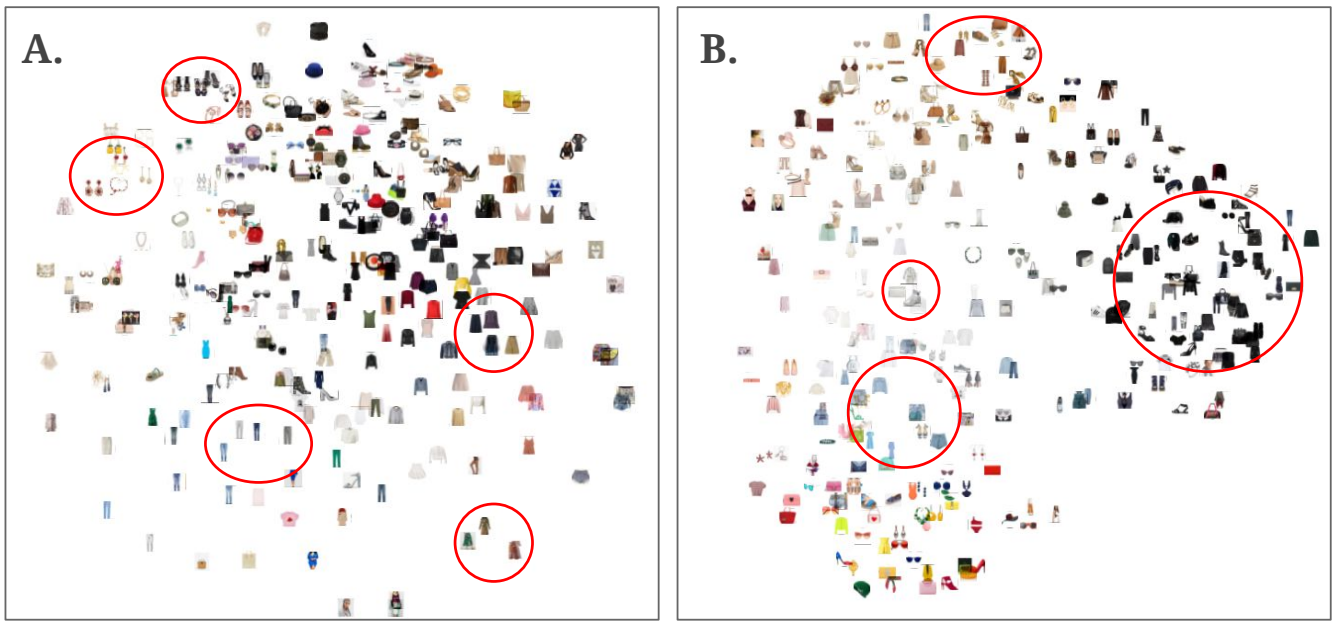

Figure 5: Visualization of embeddings from (left) ImageNet pre-trained model. The representations are generally shape-variant as it helps to distinguish different object classes indicating pre-training task bias [4, 16]. (right) Ours (α = 5%). The representations show that our embedding space can effectively capture appearance information (such as color). At the same time, items from different categories are closer to each other compared to pre-trained ImageNet model.

fully-supervised CSN [44] and Siamese network. Further, as we increase α to 50%, our approach starts to outperform even the fully-supervised approaches that use additional data such as text description [43, 44].

Another interesting point to note is that, at full supervision (i.e.α = 100%), our approach achieves 0.89 AUC outperforming many supervised approaches on the compatibility task compared in Table 2 due to the added consistency regularization supervision.

5.5 Visualization

Embedding Space. We plot t-SNE [28] of the visual representation space Φ for ImageNet and our model as shown in Fig 5. As shown in Fig. 5A, ImageNet pre-trained models trained for classification depicts stronger bias for learning shape discriminative features that bringing different category items farther away. This bias hampers the learning of fashion compatibility especially in a semi-supervised setting like ours as discussed in Sec. 3.3. In Fig. 5B, we portray the t-SNE plot of the representation space learned by our semi-supervised approach (α = 5%). The t-SNE plot demonstrates that the embedding space has learnt strong representations for visual-appearance characterized by the color and texture information of fashion items. Hence, by explicitly disentangling the shape information of the fashion items, our approach overcomes the pre-training task bias.

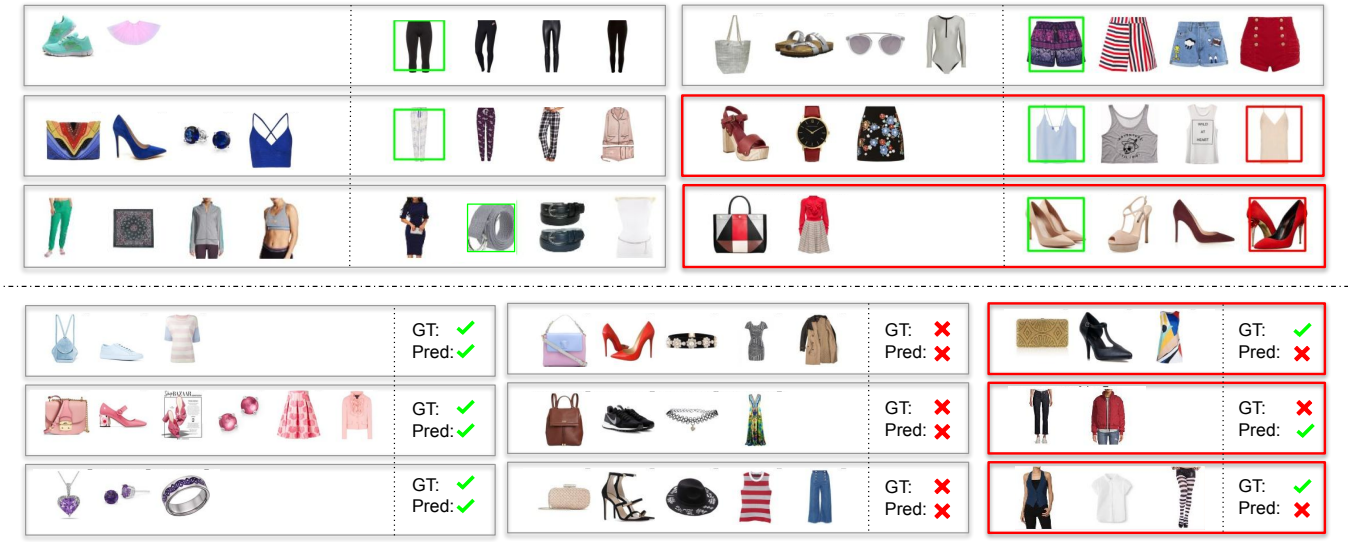

Figure 6: Qualitative results on Polyvore and Fashion Outfits datasets. Top three rows show the results on FITB task. Each box contains a query outfit and four candidate choices. Green and red boxes indicates correct and incorrect predictions of our model, respectively. Bottom three rows show the results on compatibility tasks. Some of the failure cases of our model are highlighted with a red box.

Qualitative results: In Fig. 6, we present qualitative results of our approach on the Polyvore and Fashion outfits dataset. The results show that our approach is able to model the concepts of color and texture well. In the Fig. 6, we also present some of the failure cases where our approach produces suboptimal predictions for FITB and Compatibility tasks. Our model performs suboptimally on outfits that contain items with significantly varying appearance. For example, in the last box of Fig. 6, the texture of the *trouser* is very different from that of the *shirt* and the *outerwear*.

6 CONCLUSION

In this work, we have presented an approach for learning strong visual representations for fashion compatibility using limited labeled training outfit data. We proposed two techniques for leveraging unlabeled data. First, to learn important attributes such as color, we introduced a self-supervision scheme that enforces consistency in the representation of input and its random transformations. While this acts like a data augmentation at instance level, we also proposed pseudo-labeling technique that creates pseudo-labels based on the visual similarity of labeled and unlabeled images. We conducted our experiments on Polyvore, Polyvore-D and newly created Fashion outfits dataset and achieved results on-par with supervised methods. This, however, is achieved with only a fraction of labeled data and without using any meta data such as text description.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules. In *International Conference on Very Large Data Bases (VLDB)*.
[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Neural Information Processing Systems (NeurIPS)*.
[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *European Conference of Computer Vision (ECCV)*.
[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *ICML* (2020).
[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).
[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501* (2018).
[7] Guillem Cucurull, Perouz Taslakian, and David Vazquez. 2019. Context-aware visual compatibility prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
[8] Aditya Deshpande, Jason Rock, and David Forsyth. 2015. Learning large-scale automatic image colorization. In *International Conference of Computer Vision (ICCV)*.
[9] Jiali Duan, Xiaoyuan Guo, Son Tran, and Jay Kuo. 2019. Fashion Compatibility Recommendation via Unsupervised Metric Graph Learning. In *Neural Information Processing Systems workshop*.
[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* (2018).
[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Neural Information Processing Systems (NIPS)*.
[12] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. 2019. Scaling and benchmarking self-supervised visual representation learning. In *International Conference of Computer Vision (ICCV)*.
[13] Yves Grandvalet and Yoshua Bengio. 2005. Semi-supervised learning by entropy minimization. In *Neural Information Processing Systems (NIPS)*.
[14] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *International Conference of Computer Vision (ICCV)*.
[15] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. 2017. Learning fashion compatibility with bidirectional lstms. In *ACMMM*.
[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A Plummer. 2020. Self-supervised Visual Attribute Learning for Fashion Compatibility. *arXiv* (2020).

[19] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)* (2015).

[20] Jogendra Nath Kundu, Ambareesh Revanur, Govind Vitthal Waghmare, Rahul Mysore Venkatesh, and R Venkatesh Babu. 2020. Unsupervised Cross-Modal Alignment for Multi-Person 3D Pose Estimation. *European Conference of Computer Vision (ECCV)* (2020).

[21] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. 2020. Towards Inheritable Models for Open-Set Domain Adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Jogendra Nath Kundu, Rahul Mysore Venkatesh, Naveen Venkat, Ambareesh Revanur, and R. Venkatesh Babu. 2020. Class-Incremental Domain Adaptation. In *European Conference of Computer Vision (ECCV)*.

[23] Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLw*.

[24] Yen-Liang Lin, Son Tran, and Larry S Davis. 2020. Fashion Outfit Complementary Item Retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Zhi Lu, Yang Hu, Yunchao Jiang, Yan Chen, and Bing Zeng. 2019. Learning Binary Code for Personalized Fashion Recommendation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. 2020. Learning to dress 3d people in generative clothing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning (JMLR)* (2008).

[29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems (NIPS)*.

[30] Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. 2020. Image Based Virtual Try-On Network From Unpaired Data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference of Computer Vision (ECCV)*.

[33] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. 2017. Representation learning by learning to count. In *International Conference of Computer Vision (ICCV)*.

[34] Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. Semi-Supervised Semantic Segmentation with Cross-Consistency Training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems (NeurIPS)*.

[36] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. 2020. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[37] MV Rahul, Revanur Ambareesh, and G Shobha. 2017. Siamese network for underwater multiple object tracking. In *International Conference on Machine Learning and Computing (ICMLC)*.

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Neural Information Processing Systems (NIPS)*.

[39] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Neural Information Processing Systems (NeurIPS)*.

[40] Reuben Tan, Mariya I Vasileva, Kate Saenko, and Bryan A Plummer. 2019. Learning similarity conditions without explicit supervision. In *International Conference of Computer Vision (ICCV)*.

[41] Pongsate Tangseng and Takayuki Okatani. 2020. Toward explainable fashion recommendation. In *Winter Conference on Applications of Computer Vision (WACV)*.

[42] P. Tangseng, K. Yamaguchi, and T. Okatani. 2018. Recommending Outfits from Personal Closet. In *Winter Conference on Applications of Computer Vision (WACV)*.

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

[44] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. 2017. Conditional similarity networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[45] Naveen Venkat, Jogendra Nath Kundu, Durgesh Kumar Singh, Ambareesh Revanur, and R. Venkatesh Babu. 2020. Your Classifier can Secretly Suffice Multi-Source Domain Adaptation. In *Neural Information Processing Systems (NeurIPS)*.

[46] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. 2019. Interpolation consistency training for semi-supervised learning. *International Joint Conference on Artificial Intelligence (IJCAI)* (2019).

[47] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 2020. Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Image Content. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Xun Yang, Xiaoyu Du, and Meng Wang. 2020. Learning to Match on Graph for Fashion Compatibility Modeling. In *AAAI conference on Artificial Intelligence (AAAI)*.

[49] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference of Computer Vision (ICCV)*.