Learning Fashion Compatibility with Context Conditioning Embedding

Zhi Lu, Yang Hu, Cong Yu, Yan Chen, *Senior Member, IEEE,* and Bing Zeng, *Fellow, IEEE*

***Abstract*—Fashion compatibility predictions have obtained a lot of attention recently. Mining the compatibility between fashion items in an outfit is different from learning the visual similarity, since this relationship is more delicate. Decomposing the outfit compatibility into pairwise item matching is a popular way to treat the problem. However, in most existing methods, the items are matched without considering the context, i.e, the remaining items in the outfit. Recent efforts have been made to learn the underlying high order relationships among items by treating the outfit as a whole. These models could be sensitive to the properties of different datasets, and the item representations in these models are not as compact as those in the pairwise models. In this paper, we propose a context conditioning embedding approach to learn compact representations that preserve the shared information among items under the existence of contextual items. We use two different spaces, the general and the contextual spaces, to embed items, where the representation in the contextual space contains information from the context. We employ mutual information maximization for model learning, which is shown to be more appropriate for the problem. With extensive experiments, we show that our model achieves superior performance than other state-of-the-art methods.**

***Index Terms*—Fashion analysis, outfit recommendation, compatibility learning, metric learning.**

I. Introduction

THE community has recently seen a surge of interest in fashion compatibility prediction, which has wide applications in fashion oriented social networks and online shopping. Existing methods are mainly different in their ways to tackle the compatibility. A straightforward approach is to decompose the outfit compatibility into matches between pairwise items [1]–[4]. Then the key problem is to learn appropriate item representations that capture the underlying compatibility. However, in most existing works, the item pairs are matched without considering the context, i.e. the presence of the remaining items in the outfit. Although the previous work [5] has considered the contextual information in this task, the context they used are items in other outfits, but not items in the current outfit, which ends up in a very different setting. Besides the item-level approaches, other recent works have also achieved promising results by treating an outfit as a whole [6]–[9] or leveraging graph-based models [5], [10]–[13] to learn higher-order relationships among items. These models need to be carefully learned because their performance may be hurt by an improper negative sampling strategy when other structural differences (such as outfit length and rare item pairs) are easier to learn. Furthermore, the item representations learned are usually not compact and therefore are not convenient to use in downstream tasks. In this work, our task is to learn compact representations of the items for compatibility prediction, which capture the underlying locality of the context.

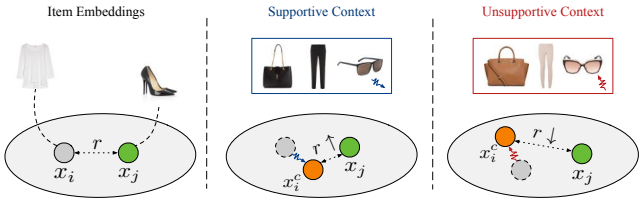


Fig. 1. An example for context conditioning compatibility. x_i (in gray) and x_j (in green) are the top and bottom item embeddings respectively without context injection. The context injection moves x_i (in dashed gray) to x_i^c (in orange). The context conditioning embedding x_i^c can be pushed towards or away from x_j by different context items, making the compatibility increase ($r \uparrow$) or decrease ($r \downarrow$) accordingly.

Fig. 1 illustrates an example where the compatibility between a top and a pair of shoes can change when evaluated with different context items. By injecting different contextual information, the feature representation of one item is pushed towards or away from another item. Intuitively, the compatibility between the item pair becomes higher if the items in the context have a similar style and vice versa. The context introduces a notion of locality for item pairs which involves relationships that can change with the existence of different context. However, the local connectivity cannot be identified by an embedding method [14] that only considers global connectivity, such as the harmony in color or texture of items. Thus, the ignorance of the context would limit the performance and the interpretability of the model.

The context is a widespread concept that exists in different fields [15], [16]. For example, to retrieve the same person in a query image in person re-identification [15], a set of labeled images associated with that person can be considered as the context. In multi-modality learning, features extracted

Zhi Lu, Cong Yu and Bing Zeng are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: zhilu@std.uestc.edu.cn; congyu@std.uestc.edu.cn; eezeng@uestc.edu.cn).

Yang Hu is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: eeyhu@ustc.edu.cn).

Yan Chen is with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: eecyan@ustc.edu.cn).

Corresponding author: Yang Hu. This work was supported by the National Natural Science Foundation of China under Grant 62172381.

from other modalities can be used as context to enhance representations. An outfit consists of a set of items, and each item in the outfit only provides partial information about the outfit. Therefore, in order to determine whether a pair of items are from the same outfit, the rest of items in the outfit play an important role, as they provide the background for the prediction.

On the other hand, in the task of item suggestion [17], [18], different items may have different roles in the set. Some items may be more dominant than the others. And people may want a suggested item that not only is compatible with the item set, but also has a style particularly similar to one dominant item. The model that focuses on the global connectivity ignores the existence of the context, thus the recommended results will solely be based on the query item and may break the overall style consistency. A straightforward solution is to evaluate the overall compatibility of the outfit when the new item is included. This usually treats existing items in the set equally, thus ignoring the role of the dominant item, and evaluating the overall outfit compatibility for all possible items could be less efficient. Learning embeddings and compatibility with context information can provide a good solution to address these limitations. With a proper context encoding approach, it can be simplified as an effective item-biased compatibility measure between the suggested item and the given item set. This motivates us to learn the context conditioning embeddings of items, and maximize the common information retained in the conditional embeddings.

In this paper, we decompose an outfit into item pairs and each pair takes the rest of items as its context. For each pair, we then encode one item conditioned on the context into the contextual space and the other item into a general space without the context. We only inject the context to one of the items because it is an economical way compared to symmetric context injection. For example, to recommend a complementary item, we therefore do not need to recompute the representations for all candidates.

Learning compact embeddings with context injection is challenging. Unlike plain embedding, where the compatibility between two items is inferred only from the representations of the two items, in context conditioning embedding, compatibility can also be deduced by the context. When the contextual information is introduced as in our case, the given item pair can be more distinguishable, which may make the training easier and vulnerable to local optima. The embedding learned is thus less representative because there may not be enough gradient to use. Also, as previous works have noticed [19], [20], a hard sample mining strategy is usually crucial for the performance of metric learning like tasks [21].

To learn the context conditioning embeddings effectively, we propose to use density ratio to model the relationships between items and a normalized score based on the density ratio to represent the compatibility. The item representations are learned under a mutual information maximization framework [22] so that more shared information can be preserved. Unlike the frequently used triplet ranking loss, the mutual information maximization encourages learning from hard samples [23], which is important in model learning. Through extensive experiments, we show this is more suitable for the compatibility prediction task. The experimental results demonstrate that our approach outperforms the state-of-the-art methods. This paper contains three main contributions as follows:

1) We design a context conditioning embedding framework for learning the compatibility between items with context information.
2) We propose to use a normalized score base on density ratio to better measure the compatibility.
3) We present a mutual information maximization framework to learn the compatibility.

II. Related Works

Fashion-related problems have been studied a lot recently in the computer vision community. [24]–[30]. In this section, we discuss existing works on outfit compatibility learning, which can be categorized into item-level and outfit-level approaches based on the ways of treating an outfit. The item-level approaches usually decompose the multiway relationship into pairwise compatibility, and this compatibility is independent of the remaining items in the outfit. The outfit-level approaches model the compatibility by considering the outfit as a whole. Besides, graph-based approaches have also been proposed recently to utilize the co-occurrence relations between fashion items.

A. Item-level approaches

In most existing methods, metric learning is used to learn a latent space for fashion items to measure the compatibility. McAuley et al. [31] used features extracted from CNN to learn compatibilities among fashion items. Veit et al. [32] trained an end-to-end Siamese Network as the distance metric. In [33], a conditional similarity network (CSN) was proposed to learn the embeddings in different subspaces. Vasileva et al. [1] then applied CSN to outfit compatibility prediction and used the categorical information as different conditions. TransNCFM [3] considered the conditional similarity measurement in a translation-based manner. Tan et al. [2] proposed the SCE-Net that learned the compatibility with attention mechanism, which didn't require categorical information. The goal of these methods is to improve the learned metric for item pairs. In contrast, we enhance the feature representation with contextual information, which is more reasonable in the setting of outfit compatibility learning.

B. Outfit-level approaches

To capture higher order relationships among items, outfit-level approaches treat an outfit as a whole to improve the performance. Tangseng et al. [34] learned the outfit representation by concatenating item features in order. Li et al. [9] learned outfit popularity with RNN model. Han et al. [6] treated an outfit as a sequence and learned the compatibility with Bi-LSTM model. However, in these methods the performance is relevant to the order of items. To address this problem, Yang et al. [7] improved the performance by tweaking the

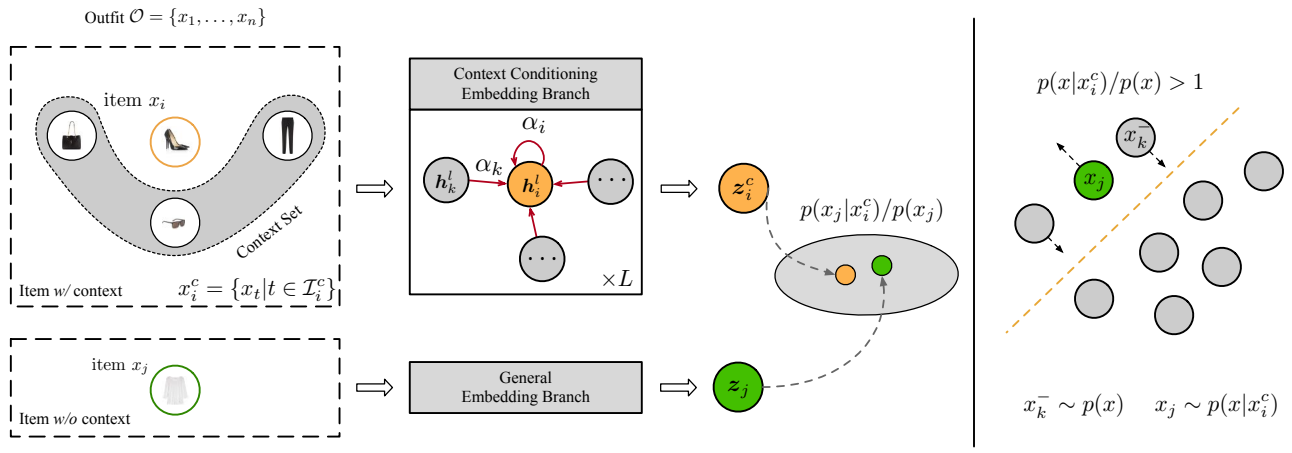


Fig. 2. Overall framework. The model consists of two branches: a context conditioning embedding branch that encodes the item x_i with the context set, and a general embedding branch that encodes the item x_j without the context. $\boldsymbol{z}_i^c$ and $\boldsymbol{z}_j$ are the corresponding embeddings of items x_i and x_j. The density ratio $p(x_j|x_i^c)/p(x_j)$ is estimated based on the similarity between $\boldsymbol{z}_i^c$ and $\boldsymbol{z}_j$. Our model maximizes the density ratio of item x_j and minimizes the density ratio of negative items x_k^- sampled from $p(x)$ given the set x_i^c.

sequence model and used fine-grained categorical information. Lu et al. [8] used a transformer-based [35] aggregation model to learn the outfit embedding which is permutation-invariant to the order of items. Unlike pairwise models, there are no compact representation learned for items in these methods. And these models usually performed differently on different datasets. We conjecture that although the aggregation models used can capture the contextual information, the models may be sensitive to the size of outfit if the dataset is imbalanced. In this work, we use the pairwise decomposition to get a performance robust to the outfit size and capture the contextual information at the same time.

C. Graph-based approaches

To enhance the features of items, graph-based [36], [37] approaches have attracted a lot of attention recently. Cucurull et al. [5] treated the task as a link prediction problem. The edges between items were created if they were from the same positive outfit. And Singhal et al. [38] improved the performance with attentional mechanism. During evaluation, to build the graph it required all testing outfits be available beforehand, which is different from the conventional setting. Cui et al. [10] built a fashion graph of fashion categories with each edge indicting the frequency of co-occurrence in outfit. Singhal et al. [38] used a similar way to build fashion graph and treated the problem as link prediction. Liu et al. [12] used graph-based approaches and focused on the diversity problem. Zhang et at. [39] showed that the color compatibility is an important factor for evaluating the outfit compatibility. They also proposed a graph-based approach for evaluating the compatibility. The graph-based approaches were also used in personalized outfit recommendation [40], [41]. However, these approaches usually require all items be available during training and the computation is intensive when a lot of neighbors are considered.

III. Approach

Let $\mathcal{O} = \{x_1, \ldots, x_n\}$ be an outfit consisting of n items, where x_i is the i-th item. Given an item pair (x_j, x_i) in the outfit, the rest items are treated as the context of x_i. We use $\mathcal{C}_i$ to denote the set of index for the context of item x_i, and $\mathcal{I}_i^c = \mathcal{C}_i \cup \{i\}$ to denote the index set of the i-th item and its context. In addition, let $c_i = \{x_t | t \in \mathcal{C}_i\}$ and $x_i^c = \{x_t | t \in \mathcal{I}_i^c\}$ be the corresponding item sets. Item x_j is considered as positive if it comes from the same outfit of x_i and negative otherwise.

As discussed, it can be insufficient to model the relationships between item pairs alone, as the compatibility between items can be affected by the context. Thus we use a context conditioning compatibility $r(x_j, x_i|c_i)$ to measure the degree that item x_j matches item x_i conditioned on context c_i. However, without any specification, the pairwise compatibility may not be directly comparable between different pairs of items, making compatibility prediction for outfits less effective. In this paper, we propose a density ratio $p(x_j|x_i^c)/p(x_j)$ between items for the computation of compatibility $r(x_j, x_i|c_i)$. Fig. 2 illustrates the overall framework of our approach. The model consists of two branches: a context conditioning embedding branch and a general embedding branch. The context conditioning embedding branch aggregates the features in context c and encode item x_i to $\boldsymbol{z}_i^c \in \mathbb{R}^d$. The general embedding branch encodes the item x_j to $\boldsymbol{z}_j \in \mathbb{R}^d$, which is independent of the context. The general embedding branch consists of multiple nonlinear layers. A density ratio $p(x_j|x_i^c)/p(x_j)$ is then estimated based on the cosine similarity between $\boldsymbol{z}_i^c$ and $\boldsymbol{z}_j$. And the objective of training is to maximize the density ratio of positive item x_j and minimize the density ratio of negative item x_k^-.

In the following sections, we first describe how to inject context c into item x_i to obtain the embedding $\boldsymbol{z}_i^c$. Then we introduce the density ratio to model the relationship between items and define the compatibility score $r(x_j, x_i|c_i)$. After that, we propose a mutual information maximization framework to learn the parameters of the model.

A. Context conditioning embedding

There are many ways to inject the context, such as LSTMs [42], Transformers [35] and MLPs. In this paper, we simply use a stack of attentional layers [43] to compute the context conditioning embedding.

Suppose that in the l-th layer, the features of item x_i and a context item x_k are represented as $\boldsymbol{h}_i^l, \boldsymbol{h}_k^l \in \mathbb{R}^d$. We first apply a linear transformation with parameters $\mathbf{W}^l \in \mathbb{R}^{d\times d}$ to each item as follows:

$$\boldsymbol{g}_t^l = \mathbf{W}^l \boldsymbol{h}_t^l, t \in \mathcal{I}_i^c \tag{1}$$

To capture the different importance of different items, we use the following scores:

$$e_t = \text{LeakyReLU}\left(\boldsymbol{a}^\intercal \left[\boldsymbol{g}_i^l \| \boldsymbol{g}_t^l\right]\right), t \in \mathcal{I}_i^c \tag{2}$$

where "$\|$" is the feature concatenation operation and $\boldsymbol{a} \in \mathbb{R}^{2d}$ is a learnable parameter. The importance scores are then normalized using the Softmax function to get the attentional weights:

$$\alpha_t = \frac{\exp\left(e_t\right)}{\exp(e_i) + \sum_{k\in\mathcal{C}_i}\exp(e_k)}, t \in \mathcal{I}_i^c \tag{3}$$

For the target item x_i, the output is aggregated by a linear combination of all items, while for each item x_k in the context set, we simply use $\boldsymbol{g}_k^l$ as the output, i.e.

$$\boldsymbol{h}_i^{l+1} = \alpha_i \boldsymbol{g}_i^l + \sum_{k\in\mathcal{C}_i} \alpha_k \boldsymbol{g}_k^l \tag{4}$$

$$\boldsymbol{h}_k^{l+1} = \boldsymbol{g}_k^l, k \in \mathcal{C}_i. \tag{5}$$

We do not aggregate the output for items in the context c because aggregating features for all items will make them embed in a single latent space. And the effective number of training pairs is reduced to the number of items for each outfit due to the symmetry. By using the above asymmetric embedding, the number of training pairs could be the square of the item number.

Following the previous work [43], we use a multi-head attention mechanism [35] to create a richer representation. Specifically, for each feature $\boldsymbol{g}_t^l$, we divide it into m partitions $\boldsymbol{g}_t^l = [\boldsymbol{v}_{1t}^l, \ldots, \boldsymbol{v}_{mt}^l]$. The above operation is applied to each partition and the output for each partition is computed as follows:

$$\boldsymbol{v}_{pi}^{l+1} = \alpha_{pi}\boldsymbol{v}_{pi}^l + \sum_{k\in\mathcal{C}_i}\alpha_{pk}\boldsymbol{v}_{pk}^l \tag{6}$$

where α_{pi} is the attentional weight for the p-th partition of x_i. We then concatenate $\boldsymbol{v}_{pi}^{l+1}$ of each partition as the output of item x_i. In this paper, we use $m = 8$ heads and set the dimension d to 128. Only the context conditioning embedding $\boldsymbol{z}_i^c = \boldsymbol{h}_i^L$ for x_i is kept as the final output of the stacked attentional layers.

B. Compatibility modeling

To model the compatibility between items, a conditional probability $p(x_j|x_i^c)$ can be maximized for compatible pairs and minimized for incompatible pairs. However, directly modeling the conditional probability for all pairs is usually challenging. Therefore, we use the density ratio [44] to model the relationships between items as follows:

$$f(x_j, x_i^c) = \frac{p(x_j|x_i^c)}{p(x_j)}. \tag{7}$$

The density ratio can be estimated without explicitly evaluating $p(x_j|x_i^c)$ and $p(x_j)$. When x_i^c carries information about x_j, the value of $p(x_j|x_i^c)$ should be greater than $p(x_j)$ for compatible items, and smaller than $p(x_j)$ for incompatible items. The equality $p(x_j|x_i^c) = p(x_j)$ indicates decision boundary for positive and negative items. And for hard positive and negatives, where $p(x_j|x_i^c) \approx p(x_j)$, they are simply the data points that lie close to the decision boundary. It suggests that we can learn the compatibility by training a classifier with the density ratio $f(x_j, x_i^c)$ as its score function.

For a fixed x_i^c, $f(x_j, x_i^c)$ for different x_j can be used to compare their compatibilities. However, $f(x_j, x_i^c)$ may not be directly comparable across different pairs. For example, given two pairs (u_j, u_i^c) and (v_j, v_i^c), comparing the corresponding density ratios are equivalent to comparing $p(u_j, u_i^c)p(v_j)p(v_i^c)$ with $p(v_j, v_i^c)p(u_j)p(u_i^c)$. We therefore use a normalized compatibility score as the the formal measure, i.e.,

$$r(x_j, x_i|c_i) = \log\frac{f(x_j, x_i^c)}{\sum_{k\in\mathcal{P}} f(x_k, x_i^c)}. \tag{8}$$

where $\mathcal{P}$ is the set of all items. Let $p(t = j|x_i, c_i)$ be the probability that sample x_j is the positive one in $\mathcal{P}$ with "$t = j$" being the indicator of the event, we have:

$$\begin{aligned} p(t = j|x_i, c_i) &= \frac{p(x_j|x_i^c)\prod_{l\neq j}p(x_l)}{\sum_{k\in\mathcal{P}}\left(p(x_k|x_i^c)\prod_{l\neq k}p(x_l)\right)} \\ &= \frac{p(x_j|x_i^c)/p(x_j)}{\sum_{k\in\mathcal{P}}p(x_k|x_i^c)/p(x_k)} \\ &= \frac{f(x_j, x_i^c)}{\sum_{k\in\mathcal{P}} f(x_k, x_i^c)}. \end{aligned} \tag{9}$$

So the normalized compatibility score is the log of this probability. For an outfit $\mathcal{O}$ with n items, the outfit compatibility is defined as the average of the compatibilities between all item pairs in the outfit, i.e.

$$s_{\mathcal{O}} = \frac{1}{n(n-1)}\sum_{i\neq j} r(x_j, x_i|c_i). \tag{10}$$

The formula for density ratio in Eq. (7) defines how to evaluate different notions of connectivity between items. Benefiting from the high capacity of deep neural networks, we simply use a scaled cosine similarity based function [45] between item representations to evaluate the density ratio. In summary, we give three different formulas that focus on different connectivities as follows:

1) *Global connectivity*.

$$f_g(x_j, x_i^c) = \exp\left(\boldsymbol{z}_j \cdot \boldsymbol{z}_i/\tau\right) \tag{11}$$

2) *Local connectivity*.

$$f_l(x_j, x_i^c) = \exp\left(\boldsymbol{z}_j \cdot \boldsymbol{z}_i^c/\tau\right) \tag{12}$$

3) *Hybrid connectivity*.

$$f_h(x_j, x_i^c) = \exp\left(\beta \boldsymbol{z}_j \cdot \boldsymbol{z}_i^c/\tau + (1-\beta)\boldsymbol{z}_j \cdot \boldsymbol{z}_i/\tau\right) \tag{13}$$

The operation "$\cdot$" indicates the cosine similarity and τ is the temperature parameter [45] to control the sharpness of the function. The local connectivity f_l captures the relationship between items with context, and the global connectivity f_g represents the model without context. The hybrid connectivity f_h uses a scaler $\beta \in [0, 1]$ to balance the two different connectivities.

C. Objective function

Since directly maximizing the log-likelihood in Eq. (8) involves computing related to all items, it is computationally intensive and infeasible in practice. Therefore, for each positive item x_i, we sample s negative items $\{x_1^-, \ldots, x_s^-\}$ from the data distribution $p(x)$ to reduce the computation cost. We use $f_\theta(x_j, x_i^c)$ to denote one of the connectivities in Eq. (11)$\sim$(13), and the following objective function is minimized:

$$\mathcal{L}_\theta = -\mathbb{E}\left[\log \frac{f_\theta(x_j, x_i^c)}{f_\theta(x_j, x_i^c) + \sum_k f_\theta(x_k^-, x_i^c)}\right]. \tag{14}$$

This loss function is also known as InfoNCE [46] that has been widely used in contrastive learning [45]. And it is known [47] that the optimal $\mathcal{L}_\theta^*$ is achieved when

$$f_\theta^*(x_j, x_i^c) \propto p(x_j|x_i^c)/p(x_j) \tag{15}$$

Therefore, we can use the learned representations to estimate the density ratio in Eq. (7) and the compatibility score in Eq. (8). Besides, items whose categories are different from the item x_j are easy to be distinguished due to factors such as shape difference. Therefore, to make the learned representation more effective, we sample each negative item x_k^- from the same category of x_j in Eq. (14).

Relation to classification: The equality $p(x_j|x_i^c) = p(x_j)$ indicates decision boundary for positive and negative items. Items sampled from the conditional distribution $p(x|x_i^c)$ are positive items given x_i^c and items sampled from the data distribution $p(x)$ are negative ones. Thus Eq. (14) can also be viewed as classifying items into different distributions and it serves as a proxy task for compatibility prediction.

Relation to mutual information: Learning density ratio for item pairs is maximize the common information shared between x and x^c, i.e. the mutual information:

$$I(x; x^c) = \mathbb{E}\left[\log \frac{p(x_j|x_i^c)}{p(x_j)}\right] = \mathbb{E}\left[\log f(x_j, x_i^c)\right]. \tag{16}$$

The objective function in Eq. (14) is a lower bound for the mutual information $I(x; x^c)$ which satisfies:

$$I(x; x^c) \geq \log(s+1) - \mathcal{L}_\theta. \tag{17}$$

However, our goal is not to accurately estimate the mutual information. We maximize the mutual information to learn compact representations for items for compatibility prediction. For this reason, a large s is not necessary for learning a good representation due to the noise of the sampled negatives. Our experimental results show that there is an optimal s in our tasks.

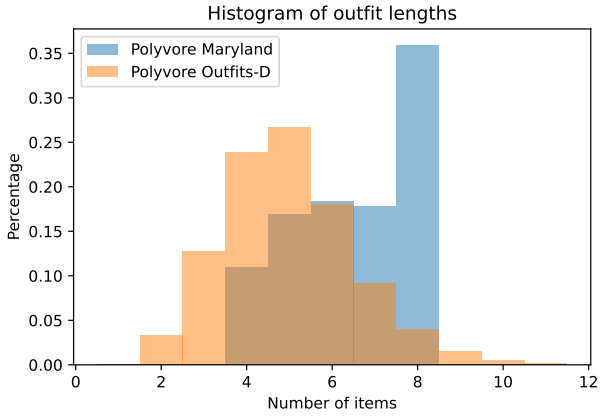


Fig. 3. Histogram of outfit lengths for different datasets.

IV. Outfit Datasets

The datasets crawled from the Polyvore website have been widely used for outfit compatibility learning. In this paper, we use two versions of Polyvore datasets to evaluate our methods: the Polyvore Maryland [6] and the Polyvore Outfits-D [1]. The Polyvore Maryland contains $17,316$ outfits for training and $3,076$ for testing. And the Polyvore Outfits-D contains $16,995$ outfits for training and $15,145$ outfits for testing. The authors of Polyvore Outfits-D also provide a non-disjoint split. In this split, not only the items in the training set appear in the testing set, but also a subset of the training outfits appear in the testing set. The partially overlapped outfits are not suitable to reveal behaviors of different methods, thus we only use the Polyvore Outfits-D dataset. The size of outfits in the Polyvore Outfits-D varies from 2 to 16 and is imbalanced comparing to Polyvore Maryland as shown in Fig. 3. For both datasets, we use their original positive outfits and only generate different negatives. To evaluate the performance, we use the Area Under The Curve (AUC) as the metric.

A. Negative outfits sampling

Sampling proper negative outfits is important for the evaluation of performance. Given the number of items n in an outfit, a direct way is to randomly select n items from the testing set. However, in this strategy, the generated outfits can contain items from two fashion categories that rarely co-occurr in the training set, or include multiple items from the same fashion category which makes the negatives easy to distinguish. A more reasonable strategy is to randomly sample items under the premise of categorical information [1]. We follow this strategy to sample the negatives and give a more general formulation in the following.

For each positive outfit of length n, we randomly select k items and replace each one with an item from the same fashion category to get the negative outfit. For Polyvore Maryland, we follow [1] to use the index of item in the outfit as the category indicator. By controlling the number of items to be replaced, we have different kinds of negative outfits with different difficulties. When k is small, it requires the model to capture more local information to do recommendation.

With the increase of k, the task becomes easier. Capturing more local connectivity or other high-order relationships is better for performance in general. But the model that overfits such relationships may perform worse when the dataset is largely unbalanced. As shown in Fig. 3, the Polyvore Outfits-D contains only a small number of outfits whose sizes are greater than 6, while the Polyvore Maryland dataset has enough outfits in different lengths. This requires the model to capture local connections while being robust to the length of outfits.

We use four different settings, i.e., $k = \{1, 2, 3, \text{all}\}$, to generate the negative outfits. In each setting, we generate the negatives with the same number of positives, and the averaged AUC accuracy is reported over 10 runs. Note that, when $k = 1$, the task is similar to the fill-in-the-blank task [6] but is evaluated under AUC and most works [1] use $k =$ all to generate negatives.

V. Experimental results

A. Baseline methods

We compare our model with the following state-of-the-art methods:

1) *SiameseNet*. We use our implementation of the Siamese Network [32] as a simple baseline. The triplet loss is used for metric learning, and features are normalized before computing the distance.
2) *Bi-LSTM* [6]. The Bi-LSTM treats an outfit as a sequence of items and is trained by maximizing the likelihood of the next item. The hidden dimension of Bi-LSTM layers is set to 128.
3) *CSN* [1]. This method uses the conditional similarity network [33] in outfit recommendation task with conditions being the categories of item pairs. We use the triplet loss to train the model.
4) *SCE-Net* [2]. The SCE-Net maps item embedding into multiple subspaces with attention mechanism. We use 4 subspaces in our experiment.
5) *TransNCFM* [3]. The TransNCFM model uses a translation-based [48] distance for computing the similarities of items pairs. The relation is encoded with categorical information.
6) *NGNN* [10]. The NGNN builds the fashion graph from the co-occurrence of fashion categories. And a graph convolutional network [49] is used to train the compatibility of outfits.
7) *LPAE-Net* [8]. The LPAE-Net uses a transformer-based network to encode the outfit into a single compact embedding and uses the averaged similarity to n learned anchors as the compatibility. We use $n = 64$ in the experiment.

We do not compare to the context-aware method [5], a link prediction approach, because this method needs all testing outfits to build the link graph during evaluation. In general case, i.e., evaluating an outfit once at a time, there will be no link information available, which makes the settings completely different.

B. Implementation details

For a fair comparison, we use the 512-dimensional features extracted from the pre-trained ResNet-34 [50] and one non-linear layer with Dropout [51] to get 128-dimensional features for all methods. The output dimension is also set to 128 for computing the compatibility. We train all models with SGD using initial learning rate $lr = 0.1$. The learning rate is reduced when the performance stops to improve. The general embedding branch consists of two fully-connected layers, and the context conditioning embedding branch consists of two attentional layers. The hidden dimension for both branches is also set to 128. To compute the density ratio, we simply use $\tau = 0.1$ to scale the cosine similarity. And for hybrid connectivity in Eq. (13) we set $\beta = 0.5$ as the default. All methods are implemented with PyTorch [52].

C. Compatibility prediction accuracy

We first show the performance on datasets with different numbers of replacement k in Table I. For LPAE-Net and NGNN, because their loss functions are minimized based on the overall score of outfit, we re-train the model for each k to match the data distribution. We find this improves their performance. For other methods, because their objectives are independent to k, we train the model once and test for all ks. As we can see, our CCE-l and CCE-h models show superior performance over other baselines in all different settings. The CCE-g represents the model without context embedding. It noticeably outperforms the SimaeseNet which show the effectiveness of the objective function used in Eq. (14).

In the Polyvore Maryland dataset, our method shows significant performance gain over other item-level approaches, such as CSN, SCE-Net, etc, especially when k is small. This shows the effectiveness of the proposed context conditioning embedding. With the attentional feature aggregation, the changes of an item in the context can affect the embedding of the target item. This amplifies the influence of the changed item and makes it easier to identify the hard negatives with small k.

In the Polyvore Outfit-D datasets, we find that the item-level approaches outperform most of the outfit-level methods. For example, when k is small, the LPAE-Net performs slight worse than other item-level baselines. When k increases, it starts to show the superiority over other baselines. However, in the Polyvore Maryland dataset, the outfit-level approaches outperform most of the item-level methods. We conjecture that because the outfit length in Polyvore Outfits-D is more imbalanced than that of Polyvore Maryland as shown in Fig. 3, the performance would be compromised when the LPAE-Net overfits the local connectivity. For Bi-LSTM, it also lacks robustness to the imbalance of outfit length. Our model is more robust to the difference of the dataset due to the pairwise decomposition. By combining the pairwise decomposition with the context conditioning embedding, we achieve superior performance over other state-of-the-art methods.

In addition, the comparison between CCE-l and CCE-h shows the difference in negative sampling strategies. The CCE-h only improves the performance when k is large. This is because when k is large, the importance of local information

TABLE I
OUTFIT COMPATIBILITY PREDICTION ON DIFFERENT POLYVORE DATASETS. WE SHOW THE AUCS ON DATASETS WITH DIFFERENT NEGATIVE ITEM REPLACEMENT NUMBER $k = \{1, 2, 3, \text{all}\}$. THE AVERAGED RESULTS OVER 10 RUNS ARE REPORTED.

Methods	Polyvore Maryland				Polyvore Outfits-D			
	1	2	3	all	1	2	3	all
SiameseNet	0.640	0.757	0.839	0.916	0.649	0.747	0.799	0.843
CSN [1]	0.641	0.765	0.847	0.925	0.653	0.755	0.808	0.851
SCE-Net [2]	0.645	0.764	0.843	0.915	0.656	0.756	0.807	0.849
TransNCFM [3]	0.641	0.762	0.845	0.920	0.645	0.743	0.797	0.838
NGNN [10]	0.691	0.786	0.866	0.918	0.590	0.651	0.693	0.734
Bi-LSTM [6]	0.704	0.825	0.892	0.943	0.602	0.676	0.720	0.757
LPAE-Net [8]	0.730	0.841	0.902	0.947	0.646	0.746	0.802	0.854
CCE-g w/ Eq. (11)	0.660	0.782	0.860	0.922	0.668	0.773	0.824	0.863
CCE-l w/ Eq. (12)	**0.741**	**0.861**	**0.917**	0.955	**0.677**	**0.781**	**0.833**	0.873
CCE-h w/ Eq. (13)	0.732	0.856	0.915	**0.958**	0.672	0.778	0.832	**0.874**

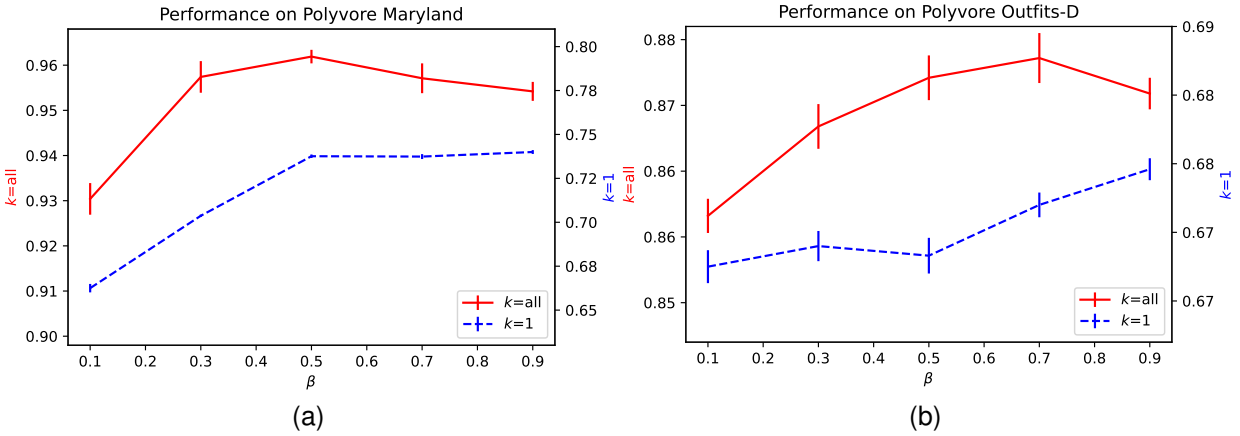


Fig. 4. Performance on the choice of β of the hybrid model. The negative outfits are generated with $k = \{1, \text{all}\}$. (a) Polyvore Maryland. (b) Polyvore Outfits-D.

TABLE II
ABLATION STUDY. WE USE NEGATIVE OUTFITS GENERATED WITH $k = \{1, \text{all}\}$.

Loss	Branch		P. Maryland		P. Outfits-D	
	Context	General	1	all	1	all
Triplet	-	-	0.644	0.890	0.649	0.835
Triplet	-	✓	0.662	0.901	0.662	0.845
Triplet	✓	-	0.706	0.946	0.657	0.858
Triplet	✓	✓	0.684	0.895	0.620	0.782
InfoNCE	-	-	0.651	0.921	0.661	0.857
InfoNCE	-	✓	0.660	0.922	0.668	0.863
InfoNCE	✓	-	0.728	0.953	0.670	0.872
InfoNCE	✓	✓	**0.741**	**0.955**	**0.677**	**0.873**

is weakened. Fig. 4 shows the detailed results of the hybrid model CCE-h trained on different β. With a hybrid of local connectivity and global connectivity, the model can be improved on k =all. However, for $k = 1$, the local connectivity is more important and the performance increases with β. In the following sections, we use CCE-l as the default.

D. Ablation study

1) On different components: To show the role of each component, we train models with different branches and different loss functions. The comparison results are shown in Table II. For model that only uses the context conditioning embedding branch, we use the features before and after that branch to evaluate the corresponding scores. The general branch usually improves performance simply because more layers give the model stronger capacity. However, for model with context conditioning embedding that is trained with triplet loss, the general embedding branch impairs the performance. The reason is that the introduction of context information makes the item pair more distinguishable and thus the training suffers from poor local optima. This shows that learning compact embedding with context is challenging. To better learn the contextual embeddings with triplet loss, a hard sample mining strategy might be necessary. However, the problem is addressed when the InfoNCE loss is used.

Overall, the model with context shows a notable improvement comparing to models without context in most cases, which shows the effectiveness of the proposed framework. And the InfoNCE loss is helpful for learning more effective embedding with additional context. With combination of context and InfoNCE, the performance can be further improved.

2) On the number of negative items: In Eq. (14), we sample a set of negative samples to train the objective function. We show the effect of using different numbers of negatives in Fig. 5. As we can see, using multiple samples improves performance, and there is an optimal s for our task. With the increase of negative items, the performance starts to decrease. Since there are no true negative items, the sampled negatives are noisy and could be harmful for training when s

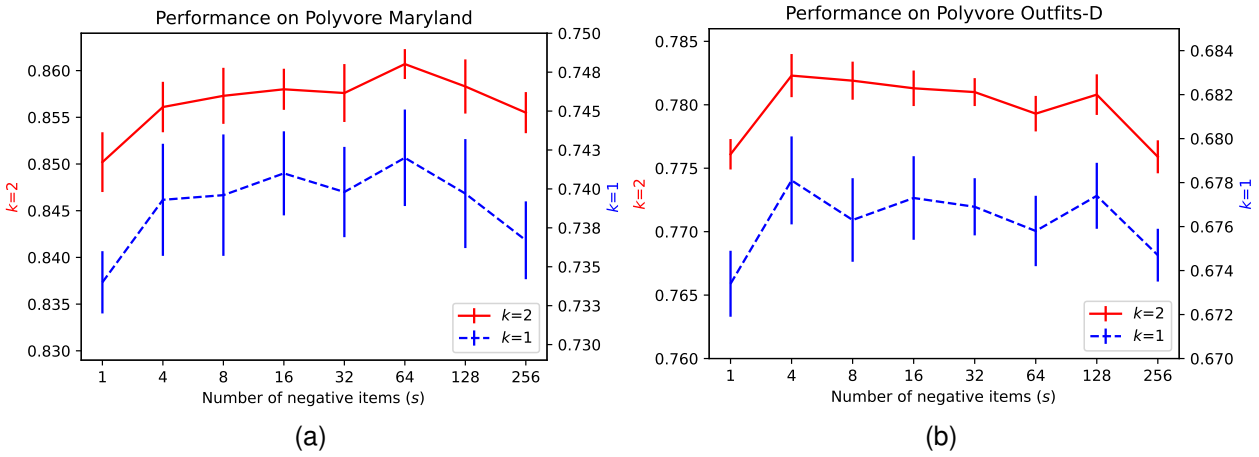


Fig. 5. Performance on different number s of negative items. The negative outfits are generated with $k = \{1, 2\}$. (a) Polyvore Maryland. (b) Polyvore Outfits-D.

TABLE III
PERFORMANCE ON DIFFERENT DATASETS OF WHETHER USING NORMALIZED SCORES AS THE COMPATIBILITY.

Datasets	Normalized	1	2	3	all
P. Maryland	-	0.722	0.840	0.902	0.949
	✓	0.741	0.861	0.917	0.955
P. Outfits-D	-	0.661	0.753	0.802	0.845
	✓	0.677	0.781	0.833	0.873

TABLE IV
ACCURACY FOR ITEM OUTLIER DETECTION TASK.

Methods	P. Maryland	P. Outfits-D
CSN	0.544 ± 0.005	0.506 ± 0.003
SCE-Net	0.523 ± 0.006	0.504 ± 0.003
TransNCFM	0.528 ± 0.007	0.495 ± 0.004
CCE (Ours)	**0.582** ± 0.007	**0.519** ± 0.004

is large [53].

3) On normalized score and density ratio: As discussed, since the density ratio may not be directly comparable across different item pairs, we normalize $f_\theta(\cdot)$ to improve the performance. To show the difference, we evaluate the performance on the unnormalized and normalized $f_\theta(\cdot)$ respectively, and the results are shown in Table III. The model with unnormalized $f_\theta(\cdot)$ achieves acceptable performance because $f_\theta(\cdot)$ is proportional to the density ratio which can capture the correlation between items. However, the performance of using unnormalized $f_\theta(\cdot)$ is less effective comparing to the normalized $f_\theta(\cdot)$, i.e. the log likelihood in Eq. (8), because this is more comparable among different item pairs and thus achieves better performance when predicting the compatibility for outfits.

E. Compatibility interpretability

Given a set of items in an outfit, we use the average compatibility of an item with other items as the contribution of it to the outfit, i.e.

$$w_i = \frac{1}{n-1} \sum_{j \neq i} r(x_i, x_j | c_j) \tag{18}$$

where n is the length of outfit. As a result, the compatibility prediction results can be interpreted using the contribution score. In addition, without contextual information, each w_i is less sensitive to item changes in the outfit, and when an item is replaced with an incompatible one, it's harder for the model to detect the outliers [54]. Fig. 6 illustrates an example, where we randomly replace an item in a positive outfit. Under the images we show the contribution scores before and after the replacement. The randomly replaced item is considered as an outlier to the outfit. As we can see, the models with and without CCE can both give reasonable results on the positive outfit. However, for the model without CCE, outliers have little effect on other items, because pairwise compatibilities are considered independently, and the larger the size of the outfit, the smaller the effect. For the model with CCE, the effects of outliers can flow into the compatibilities of other item pairs through the CCE branch and vice versa, making it more interpretable.

To quantitatively demonstrate the interpretability, we design an outlier detection task. For each positive outfit, we randomly replace an item with an outlier drawn from the same fashion category. This is equivalent to the negative outfit generated with the number of replacements $k = 1$. We report the accuracy that the outlier item has been identified through the contribution score, and the results are shown in Table IV. With context conditioning embedding, our model achieves the best performance on each dataset.

F. Item suggestion task

The model that focuses on the global connectivity will suggest items only based on the query item and may break the overall style consistency. To further show the details, we use our hybrid model trained with different β to balance the results between more global and morel local. And the top-n item suggestion results are shown in Fig. 7. $\beta = 1$ stands for our model with context information and $\beta = 0$ stands for our model without context information. When $\beta = 1$, the suggested results are more consistent on the overall style. When $\beta = 0$, suggesting based on different items yield

Fig. 6. Illustration for interpretability. The contribution scores are shown under each image, and the items in dashed boxes are the replacements.

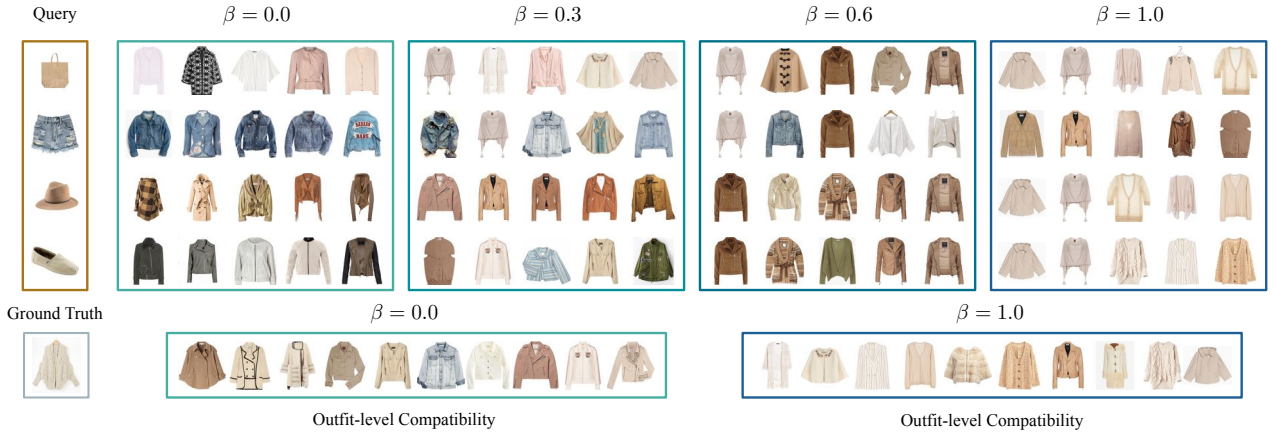


Fig. 7. Top-n item suggestion. The top rows show the tops that suggested based on the corresponding item in the left column using different β and the bottom row shows the suggestion results based on the overall compatibility.

Fig. 8. Top-n item suggestion with different context.

different results. And the results are less relevant to the remaining items. For example, the suggested results for denim shorts are all denim coats. The results for other β show a mixed suggestion with both local and global connectivities.

We further show the item recommendation results based on different context in Fig. 8 and the suggestion changes with different context. Although such kind of suggestion results may also be obtained by evaluating the outfit level compatibility, our approach is more efficient since here each item is embedded compactly.

VI. Conclusion

In this paper, we considered the problem of outfit compatibility prediction and proposed a context conditioning embedding (CCE) approach for it. We used the density ratio to model the correlation between items, based on which we defined an effective compatibility score. We used a classification objective function to learn both the embeddings and the density ratio. Besides, we revealed its relationship to mutual information maximization and provided a probabilistic perspective for compatibility learning. Through extensive experiments, we demonstrated the superiority of our approach over other state-of-the-art methods.

REFERENCES

[1] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. A. Forsyth, "Learning Type-Aware Embeddings for Fashion Compatibility," in *ECCV*, 2018, pp. 390–405.

[2] R. Tan, M. I. Vasileva, K. Saenko, and B. A. Plummer, "Learning Similarity Conditions Without Explicit Supervision," in *ICCV*, 2019, p. 10.

[3] X. Yang, Y. Ma, L. Liao, M. Wang, and T.-S. Chua, "TransNCFM: Translation-Based Neural Fashion Compatibility Modeling," in *AAAI*, 2019, pp. 403–410.

[4] Z. Lu, Y. Hu, Y. Jiang, Y. Chen, and B. Zeng, "Learning Binary Code for Personalized Fashion Recommendation," in *CVPR*, 2019, pp. 10 562–10 570.

[5] G. Cucurull, P. Taslakian, and D. Vazquez, "Context-Aware Visual Compatibility Prediction," in *CVPR*, 2019.

[6] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning Fashion Compatibility with Bidirectional LSTMs," in *ACM MM*, 2017, pp. 1078–1086.

[7] X. Yang, D. Xie, X. Wang, J. Yuan, W. Ding, and P. Yan, "Learning Tuple Compatibility for Conditional Outfit Recommendation," in *ACM MM*, 2020, pp. 2636–2644.

[8] Z. Lu, Y. Hu, Y. Chen, and B. Zeng, "Personalized Outfit Recommendation With Learnable Anchors," in *CVPR*, 2021, pp. 12 722–12 731.

[9] Y. Li, L. Cao, J. Zhu, and J. Luo, "Mining Fashion Outfit Composition Using an End-to-End Deep Learning Approach on Set Data," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1946–1955, 2017.

[10] Z. Cui, Z. Li, S. Wu, X.-Y. Zhang, and L. Wang, "Dressing as a Whole: Outfit Compatibility Learning Based on Node-Wise Graph Neural Networks," in *WWW*, 2019, pp. 307–317.

[11] P. Jing, J. Zhang, L. Nie, S. Ye, J. Liu, and Y. Su, "Tripartite Graph Regularized Latent Low-rank Representation for Fashion Compatibility Prediction," *IEEE Transactions on Multimedia*, 2021.

[12] X. Liu, Y. Sun, Z. Liu, and D. Lin, "Learning Diverse Fashion Collocation by Neural Graph Filtering," *IEEE Transactions on Multimedia*, vol. 23, pp. 2894–2901, 2020.

[13] H. Zhan, J. Lin, K. E. Ak, B. Shi, L.-Y. Duan, and A. C. Kot, "A3-FKG: Attentive Attribute-Aware Fashion Knowledge Graph for Outfit Preference Prediction," *IEEE Transactions on Multimedia*, 2021.

[14] Z. Li, J. Tang, and T. Mei, "Deep Collaborative Embedding for Social Image Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2070–2083, 2019.

[15] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 01, pp. 1–1, 2021.

[16] Z. Li, Y. Sun, L. Zhang, and J. Tang, "CTNet: Context-based Tandem Network for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[17] Y.-L. Lin, S. Tran, and L. S. Davis, "Fashion Outfit Complementary Item Retrieval," in *CVPR*, 2020, pp. 3311–3319.

[18] Z. Chen, Z. Xu, Y. Zhang, and X. Gu, "Query-Free Clothing Retrieval Via Implicit Relevance Feedback," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2126–2137, 2017.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *CVPR*, 2015.

[20] K. Sohn, "Improved Deep Metric Learning with Multi-Class N-Pair Loss Objective," in *NeurIPS*, 2016, pp. 1857–1865.

[21] K. Q. Weinberger, J. Blitzer, and L. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," in *NeurIPS*, vol. 18, 2006.

[22] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, "On Variational Lower Bounds of Mutual Information," in *NeurIPS Workshop*, 2018, p. 9.

[23] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," in *NeurIPS*, 2020.

[24] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion Meets Computer Vision: A Survey," *ACM Computing Surveys*, 2021.

[25] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi, "Mix and Match: Joint Model for Clothing and Attribute Recognition," in *BMVC*, 2015, pp. 51.1–51.12.

[26] R. He and J. McAuley, "Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering," in *WWW*, 2016, pp. 507–517.

[27] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, "Automatic Spatially-Aware Fashion Concept Discovery," in *ICCV*, 2017.

[28] Y. Gao, Z. Kuang, G. Li, P. Luo, Y. Chen, L. Lin, and W. Zhang, "Fashion Retrieval via Graph Reasoning Networks on a Similarity Pyramid," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[29] U. Mall, K. Matzen, B. Hariharan, N. Snavely, and K. Bala, "Geostyle: Discovering Fashion Trends and Events," in *CVPR*, 2019, pp. 411–420.

[30] C. Yu, Y. Hu, Y. Chen, and B. Zeng, "Personalized Fashion Design," in *ICCV*, 2019, p. 10.

[31] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-Based Recommendations on Styles and Substitutes," in *SIGIR*, 2015, pp. 43–52.

[32] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie, "Learning Visual Clothing Style with Heterogeneous Dyadic Co-Occurrences," in *ICCV*, 2015, pp. 4642–4650.

[33] A. Veit, S. Belongie, and T. Karaletsos, "Conditional Similarity Networks," in *CVPR*, 2017, pp. 830–838.

[34] P. Tangseng, K. Yamaguchi, and T. Okatani, "Recommending Outfits from Personal Closet," in *ICCV*, 2017, pp. 2275–2279.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *NeurIPS*, 2017.

[36] J. Tang, X. Shu, Z. Li, Y.-G. Jiang, and Q. Tian, "Social Anchor-Unit Graph Regularized Tensor Completion for Large-Scale Image Retagging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 2027–2034, 2019.

[37] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph Neural Networks: A Review of Methods and Applications," *arXiv*, 2018.

[38] A. Singhal, A. Chopra, K. Ayush, U. P. Govind, and B. Krishnamurthy, "Towards a Unified Framework for Visual Compatibility Prediction," in *WACV*, 2020, pp. 3607–3616.

[39] H. Zhang, X. Yang, J. Tan, C.-H. Wu, J. Wang, and C.-C. J. Kuo, "Learning Color Compatibility in Fashion Outfits," *arXiv*, 2020.

[40] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T.-S. Chua, "Hierarchical Fashion Graph Network for Personalized Outfit Recommendation," in *SIGIR*, 2020.

[41] X. Yang, X. Du, and M. Wang, "Learning to Match on Graph for Fashion Compatibility Modeling," in *AAAI*, 2020.

[42] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[43] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph Attention Networks," in *ICLR*, 2018.

[44] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning*, 2012.

[45] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *ICML*, 2020.

[46] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv*, vol. cs.LG, 2018.

[47] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, "On Variational Bounds of Mutual Information," in *ICML*, 2019.

[48] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-Relational Data," in *NeurIPS*, 2013, pp. 2787–2795.

[49] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated Graph Sequence Neural Networks," in *ICLR*, 2016.

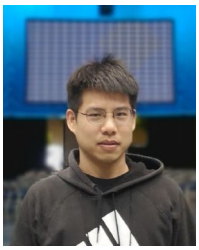
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016, pp. 770–778.

[51] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks From Overfitting," *Journal of Machine Learning Research*, 2014.

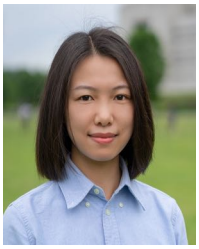
[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "Pytorch: An Imperative Style, High-Performance Deep Learning Library," in *NeurIPS*, 2019, pp. 8024–8035.

[53] C. Wu, F. Wu, and Y. Huang, "Rethinking InfoNCE: How Many Negative Samples Do You Need?" *arXiv*, 2021.

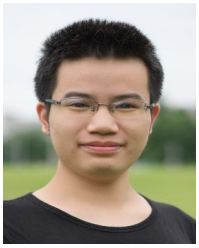
[54] Z. Lu, Y. Hu, Y. Chen, and B. Zeng, "Outlier Item Detection in Fashion Outfit," in *ICIAI*, 2022, pp. 166–171.



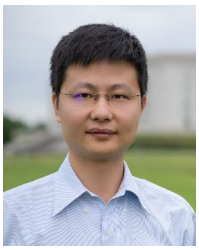
Zhi Lu received the B.S. and M.Phil. degrees from University of Electronic Science and Technology of China, Chengdu, China, in 2015 and 2018 respectively. He is currently pursuing the Ph.D. degree at the School of Information and Communication Engineering, University of Electronic Science and Technology of China. His current research interests include computer vision, machine learning and multimedia signal processing.



Yang Hu received the B.S. and Ph.D. degrees in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2004 and 2009 respectively. She was with the University of Maryland Institute for Advanced Computer Studies as a research associate from 2010 to 2015. She is currently an associate professor with the School of Information Science and Technology at the University of Science and Technology of China. Her current research interests include computer vision, machine learning and multimedia signal processing.

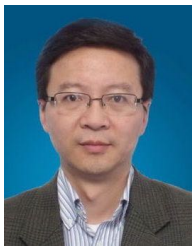


Cong Yu received B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2019. He is currently pursuing the Ph.D. degree at the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include wireless sensing, image synthesis, and adversarial learning.



Yan Chen (SM'14) received the bachelor degree from the University of Science and Technology of China in 2004, the M.Phil. degree from the Hong Kong University of Science and Technology in 2007, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2011. He was with Origin Wireless Inc. as a Founding Principal Technologist. From Sept. 2015 to Feb. 2020, he was a Professor with the School of Information and Communication Engineering at the University of Electronic Science and Technology of China. He is currently a Professor with the School of Cyber Science and Technology at the University of Science and Technology of China.

Dr. Chen's research interests include multimodal sensing and imaging, multimedia signal processing, and wireless multimedia. He is a co-author of "Reciprocity, Evolution, and Decision Games in Network and Data Science" (Cambridge University Press, 2021) and "Behavior and Evolutionary Dynamics in Crowd Networks: An Evolutionary Game Approach" (Springer, 2020), as well as co-author of over 200 technical papers including more than 100 IEEE journal papers. He is the Associate Editor for IEEE Transactions on Network Science and Engineering (TNSE) and IEEE Transactions on Signal and Information Processing over Networks (TSIPN). He is the Chair for APSIPA Signal and Information Processing Theory and Methods (SIPTM) Technical Committee, a Distinguished Lecturer for APSIPA, the Secretary-General for the CES Young Scientist Network Multimedia Technical Committee. He is an Organizing Co-Chair of PCM 2017, a Special Session Co-Chair of APSIPA ASC 2017, the 10K Best Paper Award Committee Member of ICME 2017, the Multimedia Communications Symposium Lead Chair of WCSP 2019, an Area Chair for ACM Multimedia 2021, a TPC Co-Chair of APSIPA ASC 2021. He was the recipient of multiple honors and awards, including an Excellent Editor for IEEE TNSE in 2021, the best paper award at the APSIPA ASC in 2020, the best student paper award at the PCM in 2017, the best student paper award at the IEEE ICASSP in 2016, the best paper award at the IEEE GLOBECOM in 2013.



Bing Zeng received the B.E. and M.Sc. degrees in Electronic Engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1983 and 1986, respectively, and the Ph.D. degree in Electrical Engineering from the Tampere University of Technology, Tampere, Finland, in 1991. He worked as a Postdoctoral Fellow with the University of Toronto from September 1991 to July 1992 and as a Researcher with Concordia University from August 1992 to January 1993. Then he joined the Hong Kong University of Science and Technology (HKUST). After 20 years of service, he returned to the UESTC in the summer of 2013, through Chinas 1000-Talent-Scheme. At UESTC, he leads the Institute of Image Processing to work on image and video processing, 3-D and multiview video technology, and visual big data. During his tenure with the HKUST and UESTC, he graduated more than 30 Master's and Ph.D. students, received about 20 research grants, filed 8 international patents, and published more than 250 papers. He served as an Associate Editor for the IEEE TCSVT for 8 years and received the Best Associate Editor Award in 2011. He was General Co-Chair of the IEEE VCIP-2016, held in Chengdu, China, in November 2016. He is currently on the Editorial Board of Journal of Visual Communication and Image Representation and serves as General Co-Chair of PCM-2017. He was the recipient of a 2nd Class Natural Science Award (the first recipient) from the Ministry of Education of China in 2014 and was elected as an IEEE Fellow in 2016 for contributions to image and video coding.