Generative Attribute Manipulation Scheme for Flexible Fashion Search

Xin Yang[†], Xuemeng Song[†], Xianjing Han[†], Haokun Wen[†], Jie Nie[§], Liqiang Nie[†]
[†]Shandong University, Shandong, China, [§]Ocean University of China, Shandong, China
{joeyangbuer,sxmustc,hanxianjing2018,whenhaokun}@gmail.com,niejie@ouc.edu.cn,nieliqiang@gmail.com

ABSTRACT

In this work, we aim to investigate the practical task of flexible fashion search with attribute manipulation, where users can retrieve the target fashion items by replacing the unwanted attributes of an available query image with the desired ones (*e.g.*, changing the collar attribute from *v-neck* to *round*). Although several pioneer efforts have been dedicated to fulfilling the task, they mainly ignore the potential of generative models in enhancing the visual understanding of target fashion items. To this end, we propose an end-to-end generative attribute manipulation scheme, which consists of a generator and a discriminator. The generator works on producing the prototype image that meets the user's requirement of attribute manipulation over the query image with the regularization of visual-semantic consistency and pixel-wise consistency. Besides, the discriminator aims to jointly fulfill the semantic learning towards correct attribute manipulation and adversarial metric learning for fashion search. Pertaining to the adversarial metric learning, we provide two general paradigms: the pair-based scheme and the triplet-based scheme, where the fake generated prototype images that closely resemble the ground truth images of target items are incorporated as hard negative samples to boost the model performance. Extensive experiments on two real-world datasets verify the effectiveness of our scheme.

CCS CONCEPTS

• **Information systems** → **Retrieval tasks and goals**; *World Wide Web*.

KEYWORDS

Fashion Search; Generative Adversarial Networks; Attribute Manipulation; Deep Metric Learning

ACM Reference Format:
Xin Yang, Xuemeng Song, Xianjing Han, Haokun Wen, Jie Nie, Liqiang Nie. 2020. Generative Attribute Manipulation Scheme for Flexible Fashion Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401150

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401150

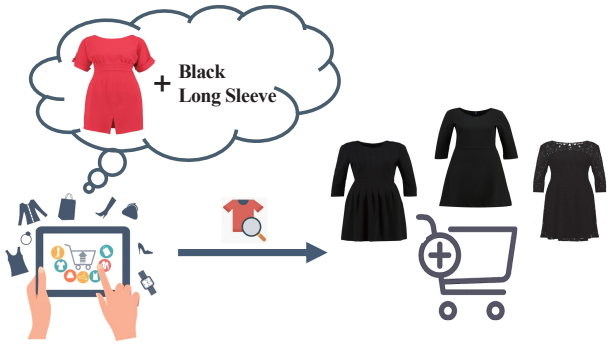


Figure 1: Illustration of the fashion search with attribute manipulation.

1 INTRODUCTION

With the prosperity of online clothing market, the Internet has accumulated numerous clothing data, making people overwhelmed and rather troublesome to search the ideal fashion items. Towards this end, content-based fashion search [18, 20, 28] that allows users to simply upload the query images to retrieve desired fashion items, has been a hot research topic in the multimedia retrieval domain [15, 16, 26, 27]. Nevertheless, in real-world scenarios, the user may not be satisfied with all attributes of the available query image. Namely, to get the desired fashion item, the user needs to modify certain attributes of the query image, *i.e.*, replacing the unwished attribute(s) with the desired one(s). For example, as shown in Figure 1, the user prefers to get a fashion item like the query item but with *black color* and *long sleeve* instead of the original *red color* and *short sleeve*. In this context, the traditional content-based fashion search cannot be directly applied to satisfy the user's needs.

In fact, several pioneer efforts have been made on the practical problem of content-based fashion search with attribute manipulation, which can be broadly classified into two groups, *i.e.*, the *fusion-based* [11, 42] and *substitution-based* [1] methods. The former ones aim to learn the latent representation of the target item by directly fusing the visual features of the query image and the semantic features of wanted attribute(s) with advanced neural networks. Their key limitation is to neglect the long-standing semantic gap [40] between the low-level visual clues and high-level attribute semantics, making the learned latent representation fail to effectively describe the target item. Differently, the substitution-based methods directly characterize the query image with multiple attributes, and the attribute manipulation can be conducted by replacing the unwished attribute features with desired ones. Although these efforts have achieved promising results, they overlook the potential of generative models in enhancing the visual understanding of the target item [24, 25]. Thus, in this work, we

aim to boost the performance of content-based fashion search with attribute manipulation by directly generating the target item image.

As a matter of fact, Generative Adversarial Networks (GANs) [9] have shown compelling success in various image generation tasks, such as the image translation [19, 30] and text-to-image synthesis [31, 39]. Therefore, we adopt the GAN as the model backbone to synthesize a prototype image conditioning on the given query image and the user's attribute manipulation. This can enhance the aesthetic feature learning of the target fashion item and further facilitate the fashion search. However, fulfilling our proposed task by means of the GAN may encounter the following research challenges. 1) Apparently, the optimal search performance cannot be achieved by the naive concatenation of the training processes of the prototype image generation and fashion search. Therefore, how to seamlessly combine them in a unified end-to-end manner, and meanwhile make them mutually reinforce each other to boost the performance constitute a tough challenge. 2) In a sense, as the visual image and the semantic attributes characterize the same fashion item, they should share certain latent features of the item. Accordingly, how to model this visual-semantic consistency in the context of fashion search with attribute manipulation poses another challenge for us. And 3) the ideal generated prototype image should satisfy all the user's requirements of attribute manipulation, and closely resemble the ground truth image of the target item. Thus, how to take into account the generated prototype images to learn a robust distance metric for the fashion search is also a crucial challenge.

To address the aforementioned challenges, we present a generative attribute manipulation scheme, dubbed as AMGAN, for flexible fashion search. As shown in Figure 2, AMGAN seamlessly integrates the prototype image generation and the target fashion item search within a unified model. In particular, it consists of two key components: *a generator G* and *a discriminator D*. The generator works on producing a prototype image that meets the user's requirement of attribute manipulation over the query image, which is regularized by the visual-semantic consistency and pixel-wise consistency. Instead of the simple real/fake image judging like traditional GANs, the discriminator is devised for the metric learning for fashion search from two perspectives: the semantic discriminative learning and the adversarial metric learning. Specifically, towards the former one, we introduce a set of attribute learners to ensure the generated prototype image to present the desired attribute semantics. Regarding the adversarial metric learning, we introduce two adversarial metric learning paradigms: *the pair-based scheme* and *the triplet-based scheme*. Motivated by the fact that the generated prototype images are fake but highly similar to the ground truth images of target items, we incorporate them as the hard negative samples [7, 32, 33] to learn more robust distance metrics. Ultimately, the prototype image generation and metric learning for fashion search can be jointly trained by playing an adversarial game. In this game, the discriminator attempts to identify the correct attribute manipulation, and meanwhile distinguish the positive and negative examples (including the generated hard negative ones) by learning a robust distance metric. Nevertheless, the generator makes great efforts to synthesize the prototype image that imitates the ground truth image and fool the learned distance metric.

The main contributions of this paper are summarized as follows:

- To our best knowledge, we are the first to adopt GANs to enhance the visual understanding in fashion search with attribute manipulation, where the generated prototype image is employed to guide the metric learning for fashion search.
- We seamlessly integrate the prototype image generation and metric learning for fashion search in an end-to-end network. In particular, the discriminator is devised to jointly fulfil the semantic discriminative learning towards the correct attribute manipulation and adversarial metric learning for fashion search.
- We incorporate the generated prototype images as hard negative examples to boost the performance, and accordingly present two novel adversarial metric learning paradigms, which adopt the pair-based and triplet-based training policy, respectively. Extensive experiments on two real-world datasets validate the superiority of our model. As a byproduct, we release the codes to benefit other researchers[1].

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. Section 3 details the proposed AMGAN. Experimental results and comprehensive analyses are presented in Section 4, followed by the conclusion and future work in Section 5.

2 RELATED WORK

2.1 Attribute Manipulation for Fashion Search

In recent years, there has been growing interest in the fashion search with attribute manipulation due to the practical demands and huge potential benefits. For example, WhittleSearch [21] allows the user to upload a query image and a text description representing the relative attribute (*e.g.,* more "brighter" color) to tune the specific attribute of the item to meet the user's demand. To achieve more direct attribute manipulation, Zhou et al. [45] employed a hybrid topic model [3, 37] to capture the intricate attribute semantics, and fulfill the demand-adaptive retrieval according to the user's specific requirement. AMNet [42] resorts to the memory block to record the template representations of various attributes. Then the wished attribute template representation can be integrated into the query image representation to search the desired fashion item. In addition, FashionSearchNet [1] characterizes a fashion item by the concatenation of multiple attribute features extracted from associated regions. Hence, it can directly replace the specific attribute features according to the user's needs to conduct the flexible search. Overall, the existing efforts dedicated to conducting attribute manipulation are mainly from the feature-level, ignoring the intuitive visual signals of the target item. In this work, we lean upon the GAN to synthesize a prototype image conditioning on the given query image and user's attribute manipulation intention to enhance the visual understanding of the desired item and facilitate the fashion search.

2.2 Generative Adversarial Networks

GANs have achieved remarkable success in various generation tasks of the fashion domain, ranging from clothing try-on [12, 43] to fashion design [5, 22]. For example, Zheng et al. [43] proposed a pose-guided virtual try-on scheme with the cycle-based GAN [46],

[1] https://joeyangbuer.wixsite.com/amgan.

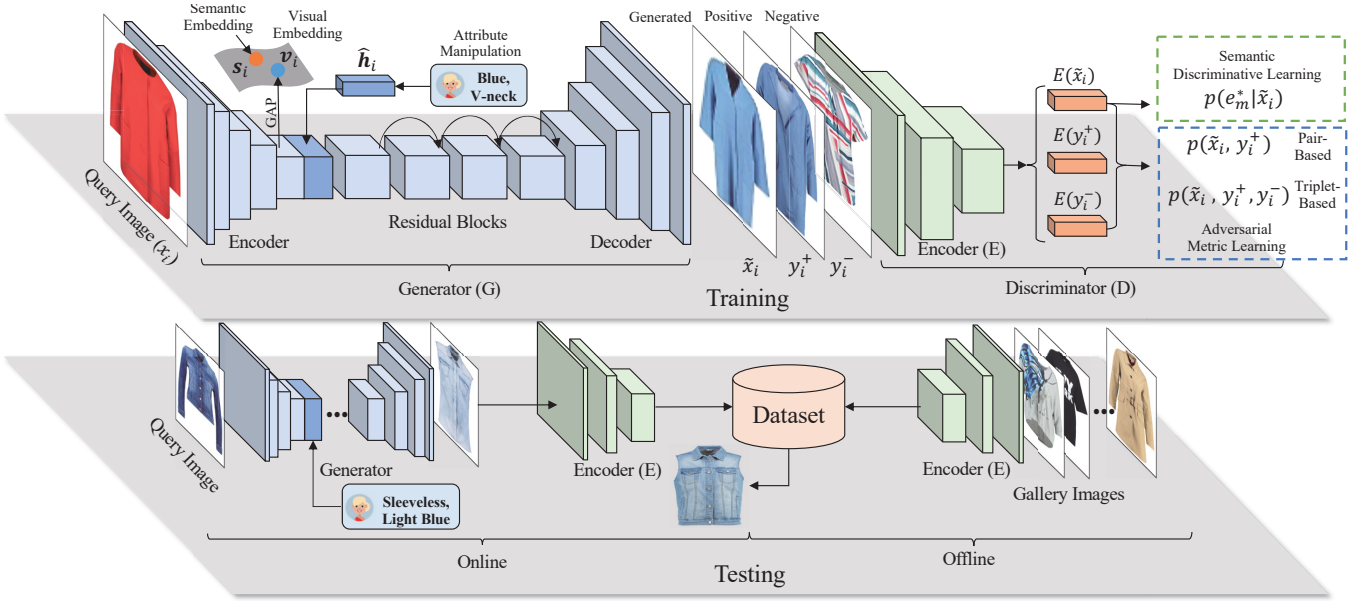


Figure 2: Illustration of the proposed AMGAN, consisting of two core parts: a generator for synthesizing prototype image and a discriminator for the semantic discriminative learning and the adversarial metric learning.

which is able to render a try-on image given a clothing item and an arbitrary pose. Cui et al. [5] allowed users to input a desired fashion sketch and a specified fabric image to generate impressive garment images for fashion design. Moreover, GANs have also harvested striking success in dealing with the information retrieval task [41]. Particularly, Liang [23] proposed an unsupervised semantic minimax two-player game by extending the GAN to address the expert retrieval task. Wang et al. [36] resorted to the GAN to unify the generative retrieval model and the discriminative retrieval model for improving the web search and recommendation applications. Distinguished from these studies that solely focus on the image generation or the information retrieval, we integrate the image generation and retrieval process in a consolidated GAN to figure out the fashion search with attribute manipulation task. In our model, the generator aims at the prototype image synthesis, while the discriminator dedicates to justifying the correct attribute manipulation as well as enhancing the metric learning for fashion search. Moreover, we make them mutually reinforce each other to boost the model performance.

3 METHODOLOGY

In this section, we first formally define the research task, and then detail our proposed model.

3.1 Problem Formulation

In the real-world content-based fashion search scenario, the exact query image that meets all the user's requirements can be hard to obtain. It is more likely that the user has an almost ideal query image, which can turn to be the ideal one with certain attribute manipulation to help retrieve the ideal fashion item. In light of this, in this work, we focus on studying the problem of essential attribute manipulation for flexible fashion search.

Formally, suppose we have a predefined set of attributes (*e.g.*, category and color) $\mathcal{A} = \{a_m\}_{m=1}^{M}$, where a_m is the m-th attribute and M is the total number of attributes. Each attribute a_m is associated with a set of possible values $\mathcal{E}_m = \{e_m^1, e_m^2, \cdots, e_m^{J_m}\}$, where e_m^j denotes the j-th value of the attribute a_m, and J_m is the total number of possible values regarding a_m. For simplicity, we compile all $\mathcal{E}_m$ in order and hence acquire a unified set of attribute values $\mathcal{E} = \bigcup_{m=1}^{M} \mathcal{E}_m = \{e^1, e^2, \cdots, e^J\}$, where $J = \sum_{m=1}^{M} J_m$. Accordingly, each image can be characterized by J attribute value labels. Meanwhile, we have a set of fashion query pairs $\mathcal{Q} = \{(x_i, \mathbf{h}_i)\}_{i=1}^{N_q}$, where N_q is the total number of query pairs, x_i is the i-th query image, and $\mathbf{h}_i = [h_i^1, h_i^2, \cdots, h_i^J]^T \in \{0,1\}^J$ is a binary vector, indicating the user's attribute manipulation over query image x_i. For example, if a user merely wants to modify the color attribute of x_i into *red* while maintaining all the rest attributes of x_i, then he/she can set $\mathbf{h}_i$ as a vector with all zeros except that $h_i^j = 1$, where j is the index of the entry e^j that represents the attribute value *red*. In addition, each query image x_i is associated with a binary vector of attribute labels, *i.e.*, $\mathbf{f}_i = [f_i^1, f_i^2, \cdots, f_i^J]^T \in \{0,1\}^J$, where $f_i^j = 1$ means that the query image x_i possesses the attribute value e^j, and 0 otherwise.

Besides, we have a set of gallery images $\mathcal{Y}$, and for each query pair $(x_i, \mathbf{h}_i)$, we sample one ground truth image $y_i \in \mathcal{Y}$ which exactly meets all the user's requirements. In a sense, we can construct the triplet training set $\mathcal{T} = \{(x_i, \mathbf{h}_i, y_i)\}_{i=1}^{N_q}$. In this work, we aim to deal with the fashion search with attribute manipulation on the basis of the GAN. The generator (G) works on synthesizing the ideal prototype fashion image $\tilde{x}_i$ for each input pair $(x_i, \mathbf{h}_i)$. The discriminator (D) is devised to learn a distance metric $d(\cdot,\cdot)$, based on which we can generate a ranking list of gallery images from $\mathcal{Y}$ for each query pair $(x_i, \mathbf{h}_i)$.

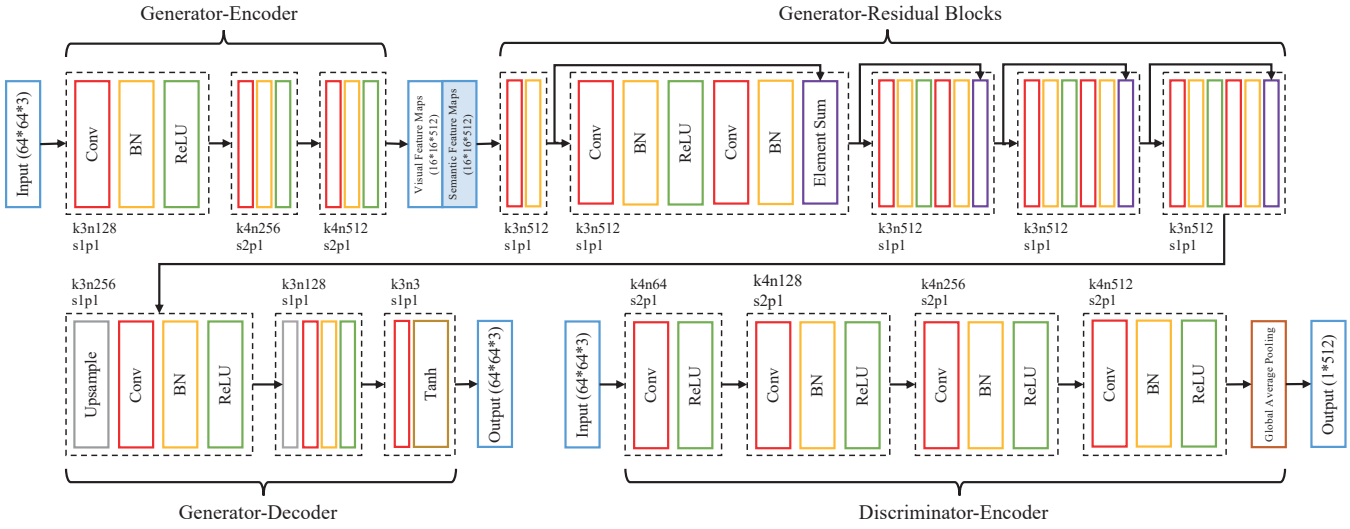


Figure 3: Details of the generator and the discriminator in AMGAN, where k represents the kernel size, n represents the number of channels, s denotes the stride, and p refers to the padding.

3.2 Prototype Image Generation

In our context, to synthesize the prototype fashion image that satisfies the user's preference of attribute manipulation over the query image, we naturally adopt the conditional GAN framework [13]. As shown in Figure 3, we generate a prototype image conditioning on both the given query image and the attribute manipulation indicator with an encoder-decoder architecture. In particular, as for the original visual signal of the query image x_i, we encode it into visual feature maps $\mathbf{F}_{\mathbf{x_i}} \in \mathbb{R}^{W \times H \times L}$ with several convolution layers followed by the Batch-Normalized (BN) and ReLU layer. Symbols W and H are the width and height of the output feature maps, respectively, and L is the number of channels.

Pertaining to the attribute manipulation indicator $\mathbf{h}_i$, to fully explore its latent semantic cues, we introduce an attribute semantic embedding matrix $\mathbf{W}_a \in \mathbb{R}^{L \times J}$, where J is the total number of attribute values, and L denotes the embedding dimension. Each column of $\mathbf{W}_a$ represents the embedding of an attribute value. Accordingly, we can obtain the attribute semantic representation of $\mathbf{h}_i$ as follows,

$$\begin{cases} \overline{\mathbf{h}}_i = \dfrac{\mathbf{h}_i}{\sum_{j=1}^{J} h_i^j}, \\ \widehat{\mathbf{h}}_i = \mathbf{W}_a \overline{\mathbf{h}}_i, \end{cases} \tag{1}$$

where $\overline{\mathbf{h}}_i$ is the normalized attribute manipulation indicator and $\widehat{\mathbf{h}}_i$ is the corresponding attribute semantic representation.

To seamlessly fuse both visual feature maps and the attribute semantic representation, similar to [6, 39], we replicate $\widehat{\mathbf{h}}_i$ to form the semantic feature maps $\mathbf{F}_{\mathbf{h_i}} \in \mathbb{R}^{W \times H \times L}$ with the same shape of visual feature maps $\mathbf{F}_{\mathbf{x_i}}$. Thereafter, we concatenate $\mathbf{F}_{\mathbf{x_i}}$ and $\mathbf{F}_{\mathbf{h_i}}$ into the fusion feature maps $\mathbf{F_i} = [\mathbf{F}_{\mathbf{x_i}}; \mathbf{F}_{\mathbf{h_i}}]$, and then transform $\mathbf{F_i}$ by several residual blocks, which have been proven to be effective in retaining indispensable visual features of the query image by learning the identify function, and hence be conducive to improving the quality of the generated images [6].

Ultimately, we use several upsampling layers and convolutional layers as the decoder, which converts the output of the residual blocks into the prototype image $\tilde{x}_i$. In a sense, the whole prototype image generation process can be briefly summarized as follows,

$$\tilde{x}_i = G(x_i, \mathbf{h}_i | \mathbf{\Theta}), \tag{2}$$

where $\mathbf{\Theta}$ refers to the parameters in the generator.

3.2.1 Visual-semantic Consistency. In a sense, the image and attributes characterize a fashion item from two different levels, *i.e.*, the visual-level and semantic-level, and thus should share certain latent consistency. For this purpose, inspired by [8, 38], we introduce a joint visual-semantic embedding space to model the inter-modal consistency by regularizing the aforementioned attribute semantic embedding matrix learning and prototype image generation. On the one hand, we map the visual feature maps $\mathbf{F}_{\mathbf{x_i}}$ of the query image x_i into the joint space with the Global Average Pooling (GAP) [44] layer, which is able to effectively translate the spatial visual features into low dimension features [2]. Let $\mathbf{v}_i \in \mathbb{R}^L$ denote the latent visual embedding of x_i. On the other hand, we encode the vector of attribute value label $\mathbf{f}_i$ by the attribute semantic embedding matrix $\mathbf{W}_a$ to derive the latent semantic embedding $\mathbf{s}_i \in \mathbb{R}^L$ of the query image x_i, which can be formulated as,

$$\mathbf{s}_i = \frac{1}{M} \mathbf{W}_a \mathbf{f}_i, \tag{3}$$

where M, the total number of attributes, is used for normalization. Propelled by the superior performance in characterizing the inter-modal consistency, we utilize the contrastive loss as the regularizer, which can be expressed as follows,

$$\begin{aligned} \min_G \mathcal{L}_G^{vse} = \mathbb{E}_{i \neq j} \big(& max\{0, g - cos(\mathbf{s}_i, \mathbf{v}_i) + cos(\mathbf{s}_i, \mathbf{v}_j)\} \\ & + max\{0, g - cos(\mathbf{v}_i, \mathbf{s}_i) + cos(\mathbf{v}_i, \mathbf{s}_j)\} \big), \end{aligned} \tag{4}$$

where g is a margin and $cos(\cdot, \cdot)$ refers to the cosine similarity operator. $\mathbf{v}_j$ and $\mathbf{s}_j$ are the latent visual embedding and attribute semantic embedding of the query image $x_j (i \neq j)$, respectively.

3.2.2 Pixel-wise Consistency. As we expect that the generator can synthesize the ideal prototype image, which satisfies all the user's requirements and hence facilitates the following metric learning for fashion search. We adopt the L1-norm regularization over the generated prototype image $\tilde{x}_i$ and the ground truth image y_i for each query pair to ensure their pixel-wise consistency and promote the generation quality of prototype images. Formally, we have,

$$\min_G \mathcal{L}_G^{L1} = \mathbb{E}_{\tilde{x}_i, y_i}[\|\tilde{x}_i - y_i\|_1]. \tag{5}$$

3.3 Metric Learning for Fashion Search

In this part, we shift to the discriminator introduction which is devised to work on the metric learning for fashion search. In our task, it is essential to learn an effective metric to measure the similarity between the generated prototype image and the gallery image. Like most metric learning methods [14, 35], we first adopt a common encoder E in the discriminator to learn the visual representation of each prototype/gallery image, denoted as $E(\tilde{x}_i)/E(y_j)$. Then we introduce the *semantic discriminative learning* and the *adversarial metric learning with two paradigms* to regularize the discriminative distance metric learning.

3.3.1 Semantic Discriminative Learning. In a sense, in order to fully meet the user's requirements, the ideal generated prototype image $\tilde{x}_i$ needs to manipulate certain attributes of the query image according to $\mathbf{h}_i$ and keeps others untouched. To this end, we introduce a set of attribute learners to encourage the generated prototype image $\tilde{x}_i$ to be discriminative towards the semantic classification. In particular, we align each attribute a_m with a separate attribute classification network C_m, $m = 1, 2, \cdots, M$, which takes the encoded representation $E(\tilde{x}_i)$ of $\tilde{x}_i$ as the input and outputs its probability distribution regarding the attribute a_m, *i.e.*, $\mathbf{p}(\mathbf{a_m}|\tilde{\mathbf{x}}_\mathbf{i}) = [p(e_m^1|\tilde{x}_i), p(e_m^2|\tilde{x}_i), \cdots, p(e_m^{J_m}|\tilde{x}_i)]$. It is worth noting that each fashion item can take only one value (label) on each attribute, namely the values (labels) of each attribute are mutually exclusive. Let e_m^* denote the ground truth label of generated prototype image $\tilde{x}_i$ regarding the attribute a_m, which can be derived by manipulating $\mathbf{f}_i$ with $\mathbf{h}_i$. To retrain the essential attribute semantics in the generated prototype image, a natural way is to maximize the corresponding classification probability.

However, due to the poor quality of images generated in the early stage of the GAN, it is inevitable to undermine the generator if we directly train these attribute classifiers via these images. Consequently, we resort to the iterative strategy [34]. Specifically, the attribute classifiers would be trained with the ground truth images y_i's to learn how to accurately classify the attributes in the discrimination stage. Nevertheless, their duty in the generation stage is to compel the generator to synthesize the prototype image with correct attribute manipulation. This idea can be expressed as,

$$\begin{cases} \max\limits_D \mathcal{L}_D^{cls} = \mathbb{E}_{y_i, a_m}[p(e_m^*|y_i)], \\ \min\limits_G \mathcal{L}_G^{cls} = \mathbb{E}_{\tilde{x}_i, a_m}[-p(e_m^*|\tilde{x}_i)]. \end{cases} \tag{6}$$

3.3.2 Adversarial Metric Learning. In the context of fashion search, we expect the learned metric can minimize the distance between similar (*i.e.*, positive) images while maximizing that between dissimilar (*i.e.*, negative) ones. Towards this end, the common strategy is to lean upon either the contrastive loss [4, 10] with paired data samples (*i.e.*, an anchor sample with a positive or negative sample) or the triplet loss [35] with triplet data samples (*i.e.,* an anchor sample, a positive sample and a negative sample). Accordingly, we propose two metric learning paradigms, *i.e.*, *the pair-based scheme* and *the triplet-based scheme*.

Algorithm 1 The Training Procedure of Our Proposed AMGAN.

Input: $\mathcal{T}$, $\mathcal{Y}$, g, b, γ, μ, λ;
Output: Parameters Θ in the generator G, parameters Φ in the discriminator D.

```
Initialize parameters in the networks G and D.
repeat
  for d-steps do
    Sample minibatch from T.
    Update the discriminator according to Eqn. (13).
  end for
  for g-steps do
    Sample minibatch from T.
    Update the generator according to Eqn. (14).
  end for
until Converge
```

Pair-based Scheme. In our context, the goal of the pair-based scheme is to minimize the distance among positive gallery image pairs and maximize the distance among the negative ones. For simplicity, similar to [29], we convert the distance estimation between two gallery images, y_m and y_n, into the similarity probability as follows,

$$p(y_m, y_n) = \sigma(b - d(y_m, y_n)) = \frac{1}{1 + e^{d(y_m, y_n) - b}}, \tag{7}$$

where $d(\cdot, \cdot)$ is the Euclidean distance and b is a shifted factor to ensure the probability close to 1 when a pair distance is near 0.

Traditionally, the paired-based scheme would only make use of the positive/negative pairs in $\mathcal{Y}$. Nevertheless, in our context, it is inappropriate to neglect the generated prototype image $\tilde{x}_i$, which is highly similar to the ground truth image y_i but still the fake image. In a sense, it can be treated as the hard negative example [7], which is a negative sample but highly resembles the positive sample and thus difficult to be distinguished. We thus argue that taking into account these hard negative examples would strengthen the capability of the learned distance metric. Particularly, we first sample a positive image y_i^+ from $\mathcal{Y}$, which shares the same attribute values with y_i. Then we impose the distance metric to maximize the probability between the real similar pair (y_i, y_i^+) but minimize the similarity probability between the fake similar pair $(\tilde{x}_i, y_i^+)$ in the discrimination stage, which can be formulated as,

$$\max_D \mathcal{L}_D^{metric} = \mathbb{E}_{y_i, y_i^+}[log\ p(y_i, y_i^+)] + \mathbb{E}_{\tilde{x}_i, y_i^+}[log(1 - p(\tilde{x}_i, y_i^+))]. \tag{8}$$

On the contrary, as an opponent, the generator should work on producing the prototype image that imitates the ground truth image as much as possible and fools the learned distance metric by minimizing the following objective function,

$$\min_G \mathcal{L}_G^{metric} = \mathbb{E}_{\tilde{x}_i, y_i^+}[log(1 - p(\tilde{x}_i, y_i^+))]. \tag{9}$$

Triplet-based Scheme. Different from the pair-based scheme, the triplet-based one focuses on modeling the relative similarity. Towards this end, we enforces the distance between the positive pair (y_i, y_i^+) to be smaller than that between the negative one (y_i, y_i^-) in a given triplet (y_i, y_i^+, y_i^-), where y_i^- is randomly sampled from the gallery image set $\mathcal{Y}$. Here we define the relative similarity probability as,

$$p(y_i, y_i^+, y_i^-) = \sigma(d(y_i, y_i^-) - d(y_i, y_i^+)) = \frac{1}{1 + e^{d(y_i, y_i^+) - d(y_i, y_i^-)}}. \tag{10}$$

Similar to the pair-based setting, we take the hard negative example $\tilde{x}_i$ into consideration, and enforce the distance metric to distinguish the positive sample y_i^+ and the negative samples (including y_i^- and $\tilde{x}_i$). Specifically, we maximize the probability that the ground truth image y_i resembles with the positive image y_i^+ more closely than the negative sample y_i^- and the hard negative sample $\tilde{x}_i$ at the same time. Therefore, we have,

$$\begin{aligned} \max_D \mathcal{L}_D^{metric} = \mathbb{E}_{y_i, y_i^+, y_i^-}[log\ p(y_i, y_i^+, y_i^-)] + \\ \mathbb{E}_{y_i, y_i^+, \tilde{x}_i}[log\ p(y_i, y_i^+, \tilde{x}_i)]. \end{aligned} \tag{11}$$

Yet the generator plays as a rival to synthesize the prototype image $\tilde{x}_i$ to imitate the ground truth image and further fool the learned distance metric by maximizing the probability that $\tilde{x}_i$ is more similar with the positive image y_i^+ than the negative one y_i^-. Accordingly, we have the following objective function,

$$\min_G \mathcal{L}_G^{metric} = \mathbb{E}_{\tilde{x}_i, y_i^+, y_i^-}[log(1 - p(\tilde{x}_i, y_i^+, y_i^-))], \tag{12}$$

where we convert the maximization to the minimization to unify the overall optimization towards the generator.

3.4 Joint Optimization

Integrating the two key components of the prototype image generation and the metric learning for fashion search, we parameterize the final objective function as below,

$$\mathbf{\Phi}^* = \arg\max_D (\mathcal{L}_D^{metric} + \lambda \mathcal{L}_D^{cls}), \tag{13}$$

$$\mathbf{\Theta}^* = \arg\min_G (\mathcal{L}_G^{metric} + \gamma \mathcal{L}_G^{vse} + \mu \mathcal{L}_G^{L_1} + \lambda \mathcal{L}_G^{cls}), \tag{14}$$

where γ, μ and λ are non-negative trade-off hyper-parameters. $\mathbf{\Phi}$ denotes the parameters in the discriminator. Overall, we optimize the above two components by an adversarial strategy. In the discrimination stage, the generated image is expected to not only conduct the correct attribute manipulation but also as a complement to learn the robust distance metric. In the generation stage, we enforce the generator to produce prototype images that imitate the ground truth images and meanwhile fool the learned distance metric. With the competition in this game, we can ultimately derive a robust distance metric for flexible fashion search. The overall procedure of the joint optimization is briefly summarized in Algorithm 1.

4 EXPERIMENT

In this section, we thoroughly evaluated the two paradigms of our model: the pair-based scheme (AMGAN-P) and the triplet-based scheme (AMGAN-T).

Table 1: Attributes and value examples of Shopping100K.

Attributes	Attribute Values	Total
Category	Shirt, Dress, Trousers, Coat, ⋯	16
Color	Black, Pink, White, Green, ⋯	19
Pattern	Animal, Plain, Photo, Print, ⋯	16
Fit	Skinny, Regular, Loose, Oversize, ⋯	15
Sleeve	Long, Short, Sleeveless, Strapless, ⋯	9
Pocket	Side, Sleeve, Zip, Flap, ⋯	7
Neckline	Boat, Backless, Round, Square, ⋯	11
Fastening	Zip, Belt, Covered, Button, ⋯	10
Collar	High, Round, Hood, Lapel, ⋯	17
Fabric	Denim, Canvas, Lace, Leather, ⋯	14
Sport	Basketball, Hiking, Swim, Tennis, ⋯	15
Gender	Male, Female	2

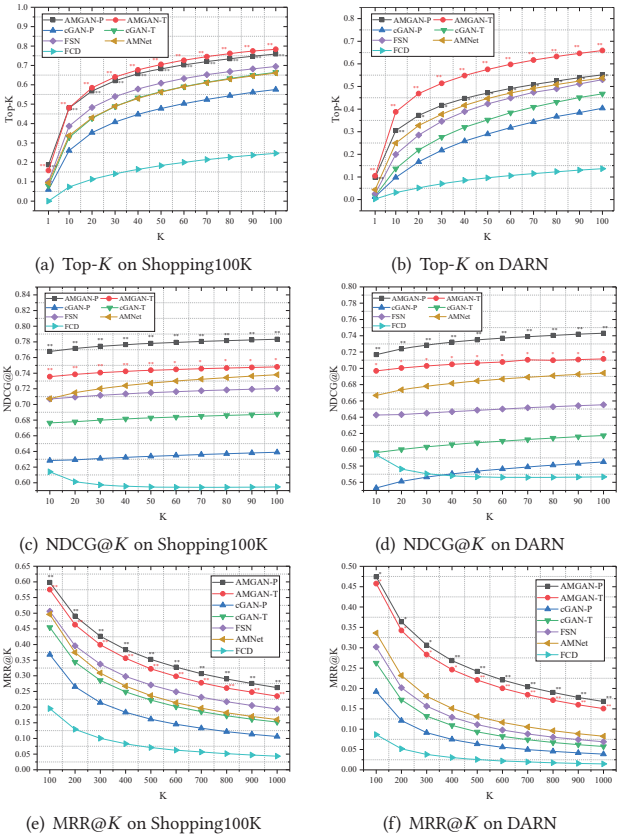
Table 2: Attributes and value examples of DARN.

Attributes	Attribute Values	Total
Clothes Category	T-shirt, Skirt, Leather ⋯	20
Clothes Color	Black, White, Red, Blue, ⋯	56
Clothes Button	Zipper, Pullover, ⋯	12
Clothes Pattern	Pure, Stripe, Dot, Lattice, ⋯	27
Clothes Length	Normal, Long, Short, ⋯	6
Clothes Shape	Slim, Straight, Cloak, ⋯	10
Sleeve Length	Short, Long, Sleeveless, ⋯	7
Sleeve Shape	Puff, Raglan, Petal, Pile, ⋯	16
Collar Shape	Round, Lapel, V-Neck, ⋯	25

4.1 Experimental Settings

Datasets. In this work, we chose the two fashion datasets annotated with rich attributes, *i.e.*, **Shopping100K** [1] and **DARN** [17]. Shopping100K consists of 101, 021 fashion items characterized by 12 attributes with 151 possible attribute values. By contrast, DARN is comprised of 253, 983 fashion items annotated with 9 attributes and 179 possible values. Tables 1 and 2 show their detailed attributes and corresponding value examples. Without loss of generality, we focused on the attribute manipulation of tops, where we utilized all tops of Shopping100K (57, 834 tops in total) and randomly sampled 50, 000 tops from DARN to balance the sizes of the two datasets. Moreover, similar to [1] and [42], we particularly studied the fashion search with manipulation over one or two attributes of the fashion items. In particular, for each given query image x_i, we fetched a target image y_i, if possible, that differs from x_i with respect to one or two attributes. Hence, we can get the attribute manipulation indicator $\mathbf{h}_i$ for x_i, whose j-th entry $\mathbf{h}_i^j = 1$ if the attribute value e^j is presented in y_i but not x_i, and 0 otherwise. In this manner, we obtained 39, 764 and 38, 291 triplets, *i.e.*, $(x_i, \mathbf{h}_i, y_i)$'s, for Shopping100K and DARN, respectively. Each triplet set is then split into three chunks: training set (80%), validation set (10%), and testing set (10%). Notably, as each query image corresponds to only one attribute manipulation indicator, the query item images in these three chunks have no overlap. As for building the gallery sets, we used all the tops of Shopping100K and the randomly selected 50, 000 tops in DARN for these two datasets, respectively.

Baselines. We chose the following state-of-the-art methods regarding fashion search with attribute manipulation for comparison.



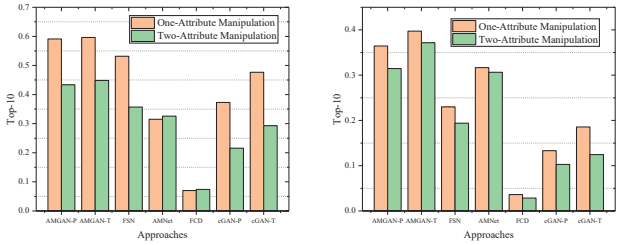
(a) Top-K on Shopping100K (b) Top-K on DARN

(c) NDCG@K on Shopping100K (d) NDCG@K on DARN

(e) MRR@K on Shopping100K (f) MRR@K on DARN

Figure 4: Overall performance comparison with baseline methods. Symbols $*$ and $**$ denote the statistical significance for $p_{value} < 0.05$ and $p_{value} < 0.01$, respectively, compared to the best baseline.

- **FCD:** FCD [11] learns attribute representations with the multi-modal data of fashion items, where the attribute manipulation can be conducted by integrating the desired attribute representations to the query image while removing the unwanted ones.
- **AMNet:** AMNet [42] learns a memory block to store the attribute template representation, and the attribute manipulation is fulfilled by directly fusing the desired attribute representations into the query image with a fully-connected layer.
- **FSN:** FashionSearchNet [1] represents each fashion item as the concatenation of multiple region-aware attribute representations. Consequently, the unwished attribute representations can be substituted with the user's desired one(s) to achieve the flexible search. For simplicity, we denoted FashionSearchNet as FSN.
- **cGAN-P**: To demonstrate the necessity of coupling the prototype image generation and the metric learning for fashion search in an end-to-end manner, we separated our model into a cGAN module and a metric learning module. The cGAN part is derived from our model by replacing the adversarial metric learning with the traditional real/fake discriminator [9]. Besides, the metric learning part follows the pair-based scheme (*i.e.*, Eqn. (9)) but keeps the generated prototype image $\tilde{x}_i$ fixed.
- **cGAN-T**: Similar to cGAN-P, we derived cGAN-T by using triplet-based scheme (*i.e.*, Eqn. (12)) in the metric learning part.



(a) Shopping100K (b) DARN

Figure 5: Performance of fashion search with one-attribute manipulation and two-attribute manipulation.

Training Setup. The detailed network structure of AMGAN is shown in Figure 3. In addition, we empirically found that the proposed model achieves the optimal performance with $g = 0.2$, $b = 1$ and $\gamma = \mu = \lambda = 1$. We iteratively trained the discriminator and the generator by Adam optimizer with the learning rate of 0.0002 and bath size of 64. For the sake of fairness, we adopted same evaluation metrics used in [1] and [42], *i.e.*, *Top-K* and Normalized Discounted Cumulative Gain (*NDCG@K*), to assess the search performance. Meanwhile, we adopted Mean Reciprocal Rank (*MRR@K*), which measures the average position of the ground truth image in the retrieved ranking list, as one metric.

4.2 On Model Comparison

Figure 4 shows the performance comparison among different methods on both Shopping100K and DARN, where we launched 5-fold cross validation and reported the average performance. From this figure, we obtained the following observations: 1) AMGAN-P(T) consistently outperforms all baseline methods over the two datasets. This confirms the advantage of our model that utilizes the generated prototype image to capture visual clues of the target item and hence boost the performance of fashion search. 2) AMGAN-P(T) shows significant superiority over cGAN-P(T) in all testing scenarios. It suggests that the adversarial learning framework can benefit the robust distance metric learning by incorporating the hard negative samples, *i.e.*, the generated prototype images. Meanwhile, this observation also verifies the necessity of jointly modeling the prototype image generation and metric learning in a unified end-to-end manner. 3) FCD shows the worst performance compared to other methods. It can be attributed to the fact that FCD focuses on the latent representation learning of the target item while overlooks the similar/dissimilar relation among fashion items. Thus, this leads to the limited discriminative capability of the learned representations. And 4) the performance of all methods on Shopping100K is better than that on DARN. One possible explanation is that the clothing item images of DARN suffer from both complicated backgrounds and various deformations and occlusions, which raises the difficulties of the fashion search task.

To gain more deep insights, we split the testing set into two parts according to the number of to-be-manipulated attributes. Ultimately, we derived 944 samples for one-attribute manipulation and 3,033 samples for two-attribute manipulation. Figure 5 shows the performance comparison in different testing scenarios. As we can see, most methods exhibit better performance on one-attribute manipulation compared with two-attribute manipulation, which is

Table 3: The ablation experiments of AMGAN on two metric learning paradigms.

Approaches	Shopping100K				DARN			
	Top-1	Top-10	NDCG@10	MRR@1000	Top-1	Top-10	NDCG@10	MRR@1000
AMGAN-P-NoVSE	0.1609	0.4717	0.7567	0.2417	0.0857	0.2967	0.7078	0.1486
AMGAN-P-NoL1	0.1652	0.472	0.7551	0.2324	0.0875	0.2998	0.7067	0.1603
AMGAN-P-NoCls	0.0649	0.2824	0.6693	0.1008	0.0144	0.1175	0.5856	0.0324
AMGAN-P	**0.1873**	**0.4805**	**0.7715**	**0.2624**	**0.0978**	**0.3051**	**0.7167**	**0.1678**
AMGAN-T-NoVSE	0.1390	0.4476	0.7222	0.2015	0.0681	0.3141	0.6681	0.1230
AMGAN-T-NoL1	0.1368	0.4317	0.7154	0.2001	0.0690	0.3172	0.6692	0.1231
AMGAN-T-NoCls	0.0714	0.3053	0.6565	0.1275	0.0164	0.1219	0.5805	0.0559
AMGAN-T	**0.1582**	**0.4811**	**0.7384**	**0.2351**	**0.1051**	**0.3876**	**0.7004**	**0.1505**

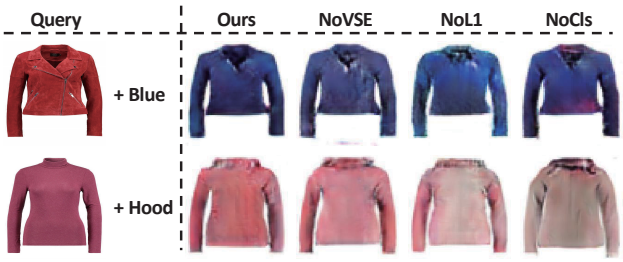


Figure 6: Intuitive examples of ablation study. The first case comes from AMGAN-P, while the second is from AMGAN-T.

Figure 7: Examples of prototype image generation. Each case is in the form: "query image + attribute manipulation = prototype image".

reasonable as the latter scenario is more challenging. Besides, we found that AMGAN-P(T) surpasses all baseline methods in both settings, indicating the effectiveness of our model in all scenarios.

4.3 On Ablation Study

To verify the importance of each component in our model, we also compared AMGAN-P(T) with the following three derivatives.

- **AMGAN-P(T)-NoVSE**: To check the impact of the visual-semantic consistency, we removed the $\mathcal{L}_G^{vse}$ by setting $\gamma = 0$.
- **AMGAN-P(T)-NoL1**: To exploit the facility of the pixel-wise consistency, we removed the $\mathcal{L}_G^{L_1}$ by setting $\mu = 0$.
- **AMGAN-P(T)-NoCls**: To study the effect of the semantic discriminative learning, we removed both $\mathcal{L}_G^{cls}$ and $\mathcal{L}_D^{cls}$ by setting $\lambda = 0$.

Table 3 shows the results of different ablation methods. As can be seen, AMGAN-P(T) shows superiority over AMGAN-P(T)-NoVSE. This verifies that the visual-semantic consistency loss can bridge the semantic gap between the low-level visual clues and high-level attribute semantics to boost the model performance. Besides, AMGAN-P(T) exceeds AMGAN-P(T)-NoL1 across all evaluation metrics, implying the importance of regularizing the visual detail preservation of the generated prototype image in our context. Furthermore, we found that AMGAN-P(T)-NoCls presents the worst performance, suggesting the pivotal role of the attribute semantic discriminative learning for regularizing the correct attribute manipulation. In a sense, the underlying philosophy is that in our task, whether the generated prototype image meets the desired attribute manipulations directly affects the downstream search performance. To intuitively show the impact of each component, we further visualized some generated prototype images in Figure 6. As shown in the first case, AMGAN-P-NoL1 synthesizes the prototype image with more fuzzy neckline than AMGAN-P, which confirms the contribution of the L1 loss in the visual detail preservation. As to the second example of Figure 6, we noticed that the generated result of AMGAN-T-NoCls fails to maintain the unmanipulated attributes (*e.g.*, color), reflecting the significance of the semantic discriminative regularizer. Overall, these three modules are all pivotal in our model for achieving superior performance.

4.4 On Case Study

To gain the thorough understanding of our model, apart form the quantitative evaluation, we also conducted the case study on both the prototype image generation and the flexible fashion search.

4.4.1 Prototype Image Generation. We provided several visualized cases regarding the prototype image generation in Figure 7. We observed that the generated prototype images for target fashion items generally meet the requirements of desired attribute manipulation over the given query images. As can be seen from the second example in the first row, according to the attribute manipulation requirement, only the color of the generated image is modified. As for other attributes, such as category and neckline, are remained the same with the query image. In addition, as shown in the third row, the generated results for cases with two-attribute manipulation are also satisfactory. Moreover, we found that our model can generate prototype images properly even for the query images that involve complicated backgrounds and fashion model noise. For example, in the fourth row, the indicated attributes are correctly manipulated, while the model poses and backgrounds are well retained.

4.4.2 Flexible Fashion Search. As shown in Figure 8, we illustrated several intuitive fashion search results obtained by our AMGAN-P and AMGAN-T. It can be seen that our proposed AMGAN-P(T)

(a) AMGAN-P. (b) AMGAN-T.

Figure 8: The top-4 retrieval results of AMGAN-P and AMGAN-T, and the ground truth images are marked with green boxes.

is capable of capturing users' attribute manipulation intents and generating proper prototype images to facilitate the precise fashion search. As shown in the first example of the left column, the retrieved items are quite similar to the given query item except the attribute of "sleeve length". In addition to the one-attribute manipulation, we observed that fashion search with two-attribute manipulation also achieved promising performance (see the third row). Besides, as shown in the second row of the left column, the third and fourth retrieved items cannot perfectly meet the attribute manipulation requirements over the query image due to the undesired color attribute. Even so, they seem to highly resemble the target item, indicating the practical value of our model.

4.5 On Running Time

In this part, we compared the time cost of our model with existing methods over a server equipped with Intel(R) Xeon(R) CPU E5 – 2620 v4 (@2.10GHz), 128 GB RAM memory, and four NVIDIA TITAN X GPUs. For each method, we carried out 1,000 fashion search queries with attribute manipulation, where the gallery set consists of 50,000 fashion items. Table 4 shows the average time cost for each query in different phases, where the item representation dimension adopted by different methods are also provided. In a sense, the time cost of the representation learning mainly depends on the model complexity and the resolution of the input image, while the search time is subject to the image representation dimension. Notably, as AMGAN-P(T) and cGAN-P(T) have the same computation complexity, we only provided the time cost of AMGAN-P(T). We can see that generating the auxiliary prototype image increases the time cost slightly in the item representation learning stage. Nevertheless, benefit from the generated prototype image, we can learn the low-dimensional and discriminative item representation, which greatly reduces the distance calculation burden. As a consequence, our model is the most efficient one compared to the other methods, which costs 0.191s in total for each query.

Table 4: The comparison of system time cost. Dim: the dimension of image representations, $T_{repr}(s)$: the time cost of the representation learning, $T_{sea}(s)$: the time cost of fashion search, $T_{sum}(s)$: the total time cost.

Approaches	Dim	T_{repr}	T_{sea}	T_{sum}
FCD	512	0.180	0.153	0.333
AMNet	4,096	**0.019**	0.928	0.947
FSN	4,096	0.139	0.940	1.079
AMGAN-P(T)	512	0.042	**0.149**	**0.191**

5 CONCLUSION AND FUTURE WORK

In this work, we present a novel generative attribute manipulation scheme for flexible fashion search. Particularly, the generator directly synthesizes a prototype image that meets the user's requirements of attribute manipulation over the query image to promote the metric learning for fashion search. Towards the correct attribute manipulation and robust distance metric learning, the discriminator is devised to simultaneously tackle the semantic discriminative learning and adversarial metric learning. Additionally, the generated prototype images are incorporated as the hard negative samples to boost the model performance. Extensive experiments have been conducted on two real-world datasets, and the encouraging empirical results prove the effectiveness and efficiency of our proposed model. This also confirms the advantage of GANs in enhancing the visual understanding for flexible fashion search. Besides, our model can be easily extended to the attribute manipulation tasks in other domain, like facial editing and birds search. One limitation of our work is that currently we only focus on the supervised prototype image generation, and thus we plan to explore the unsupervised setting for handling cases without the ground truth images of target items in the future.

ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Project of New Generation Artificial Intelligence, No.:2018AAA0102502; the National Natural Science Foundation of China, No.:61772310, No.:61702300, and No.:U1936203; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; the Shandong Provincial Key Research and Development Program, No.:2019JZZY010118; the Innovation Teams in Colleges and Universities in Jinan, No.:2018GXRC014.

REFERENCES

[1] Kenan E. Ak, Ashraf A. Kassim, Joo-Hwee Lim, and Jo Yew Tham. 2018. Learning Attribute Representations With Localization for Flexible Fashion Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7708–7717.

[2] Xavier Alameda-Pineda, Andrea Pilzer, Dan Xu, Nicu Sebe, and Elisa Ricci. 2017. Viraliency: Pooling Local Virality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 484–492.

[3] Georgios Balikas, Massih-Reza Amini, and Marianne Clausel. 2016. On a Topic Model for Sentences. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 921–924.

[4] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 539–546.

[5] Y. R. Cui, Q. Liu, C. Y. Gao, and Z. Su. 2018. FashionGAN: Display your fashion design using Conditional Generative Adversarial Nets. *Computer Graphics Forum* 37, 7 (2018), 109–119.

[6] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. 2017. Semantic Image Synthesis via Adversarial Learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 5707–5715.

[7] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. 2018. Deep Adversarial Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2780–2789.

[8] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 2121–2129.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 2672–2680.

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1735–1742.

[11] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. 2017. Automatic Spatially-Aware Fashion Concept Discovery. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1472–1480.

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. 2018. VITON: An Image-Based Virtual Try-On Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7543–7552.

[13] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing* 28, 11 (2019), 5464–5478.

[14] Elad Hoffer and Nir Ailon. 2015. Deep Metric Learning Using Triplet Network. In *Similarity-Based Pattern Recognition*. Springer, 84–92.

[15] Richang Hong, Zhenzhen Hu, Ruxin Wang, Meng Wang, and Dacheng Tao. 2016. Multi-View Object Retrieval via Multi-Scale Topic Models. *IEEE Transactions on Image Processing* 25, 12 (2016), 5814–5827.

[16] Richang Hong, Lei Li, Junjie Cai, Dapeng Tao, Meng Wang, and Qi Tian. 2017. Coherent Semantic-Visual Indexing for Large-Scale Image Retrieval in the Cloud. *IEEE Transactions on Image Processing* 26, 9 (2017), 4128–4138.

[17] Junshi Huang, Rogério Schmidt Feris, Qiang Chen, and Shuicheng Yan. 2015. Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1062–1070.

[18] Junshi Huang, Wei Xia, and Shuicheng Yan. 2014. Deep Search with Attribute-aware Deep Network. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 731–732.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5967–5976.

[20] Shuhui Jiang, Yue Wu, and Yun Fu. 2016. Deep Bi-directional Cross-triplet Embedding for Cross-Domain Clothing Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 52–56.

[21] Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2012. WhittleSearch: Image search with relative attribute feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2973–2980.

[22] Hanbit Lee and Sang-goo Lee. 2019. Fashion Attributes-to-Image Synthesis Using Attention-Based Generative Adversarial Network. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 462–470.

[23] Shangsong Liang. 2019. Unsupervised Semantic Generative Adversarial Networks for Expert Retrieval. In *Proceedings of the ACM International Conference on World Wide Web*. ACM, 1039–1050.

[24] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019. Improving Outfit Recommendation with Co-supervision of Fashion Generation. In *Proceedings of the ACM International Conference on World Wide Web*. ACM, 1095–1105.

[25] Linlin Liu, Haijun Zhang, Yuzhu Ji, and Q. M. Jonathan Wu. 2019. Toward AI fashion design: An Attribute-GAN model for clothing match. *Neurocomputing* 341 (2019), 156–167.

[26] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 15–24.

[27] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 843–851.

[28] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1096–1104.

[29] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.

[30] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. 2018. Unsupervised Attention-guided Image-to-Image Translation. In *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 3697–3707.

[31] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. In *Proceedings of the International Conference on Machine Learning*. 1060–1069.

[32] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. 2016. Training Region-Based Object Detectors with Online Hard Example Mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 761–769.

[33] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. 2019. Stochastic Class-Based Hard Example Mining for Deep Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7251–7259.

[34] Changchang Sun, Xuemeng Song, Fuli Feng, Wayne Xin Zhao, Hao Zhang, and Liqiang Nie. 2019. Supervised Hierarchical Cross-Modal Hashing. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 725–734.

[35] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning Fine-Grained Image Similarity with Deep Ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1386–1393.

[36] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 515–524.

[37] ChengXiang Zhai. 2017. Probabilistic Topic Models for Text Data Retrieval and Analysis. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1399–1401.

[38] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3107–3115.

[39] Han Zhang, Tao Xu, and Hongsheng Li. 2017. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5908–5916.

[40] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 33–42.

[41] Weinan Zhang. 2018. Generative Adversarial Nets for Information Retrieval: Fundamentals and Advances. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1375–1378.

[42] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6156–6164.

[43] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. 2019. Virtually Trying on New Clothing with Arbitrary Poses. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 266–274.

[44] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2921–2929.

[45] Zhengzhong Zhou, Jingjin Zhou, and Liqing Zhang. 2016. Demand-adaptive Clothing Image Retrieval Using Hybrid Topic Model. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 496–500.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2242–2251.