DiffFashion: Reference-Based Fashion Design With Structure-Aware Transfer by Diffusion Models

Shidong Cao, Wenhao Chai, Shengyu Hao, *Graduate Student Member, IEEE*, Yanting Zhang, Hangyue Chen, and Gaoang Wang, *Member, IEEE*

***Abstract*— Image-based fashion design with AI techniques has attracted increasing attention in recent years. We focus on the reference-based fashion design task, where we aim to combine a reference appearance image and a clothing image to generate a new fashion clothing image. Although existing diffusion-based image translation methods have enabled flexible style transfer, it is often difficult to transfer the appearance of the image realistically during reverse diffusion. When the referenced appearance domain greatly differs from the source domain, it often leads to the collapse in the translation. To tackle this issue, we present a novel diffusion model-based unsupervised structure-aware transfer method, namely *DiffFashion*. Our method is free of model tuning and structure-preserving and has high flexibility in transferring from images with large domain gaps. Specifically, based on the optimal transport properties, we keep a shared latent across the clothing image and reference appearance image to bridge the gap between the two domains in the denoising process, and the latent of the reference image is gradually adapted to the clothing domain. Simultaneously, the structure is transferred from the source clothing to the output fashion image with mixed guidance, including pre-trained Vision Transformer (ViT) guidance and a foreground mask guidance, to further preserve the structure and appearance semantics from source and reference images. Our experimental results show that the proposed method outperforms state-of-the-art baseline models, generating more realistic images in the fashion design task.**

***Index Terms*—Fashion design, diffusion models, structure-aware.**

Manuscript received 15 May 2023; revised 21 July 2023; accepted 15 September 2023. Date of publication 22 September 2023; date of current version 23 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0162000, in part by the National Natural Science Foundation of China under Grant 62106219, and in part by the Natural Science Foundation of Zhejiang Province under Grant QY19E050003. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Chong Luo. *(Shidong Cao and Wenhao Chai contributed equally to this work.) (Corresponding authors: Hangyue Chen; Gaoang Wang.)*

Shidong Cao, Wenhao Chai, and Shengyu Hao are with the Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University, Haining 314400, China (e-mail: 22271126@zju.edu.cn; wenhaochai.19@intl.zju.edu.cn; shengyuhao@zju.edu.cn).

Yanting Zhang is with the Donghua University, Shanghai 201620, China (e-mail: ytzhang@dhu.edu.cn).

Hangyue Chen is with the Hangzhou Dianzi University, Hangzhou 310005, China (e-mail: chy@hdu.edu.cn).

Gaoang Wang is with the Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University, Haining 314400, China, also with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: gaoangwang@intl.zju.edu.cn).

Code and demo can be found at https://github.com/Rem105-210/DiffFashion.

Digital Object Identifier 10.1109/TMM.2023.3318297

I. Introduction

IMAGE-BASED fashion design with artificial intelligence (AI) techniques [1], [2], [3], [4], [5], [6] has attracted increasing attention in recent years. There is a growing expectation that AI can provide inspiration for human designers to create new fashion designs. One of the emerging tasks in fashion design is to add specific texture elements from non-fashion domain images into clothing images to create new fashions. For example, given a source clothing image, a designer may want to generate a new fashion design with the appearance of another domain object as a reference, as shown in Fig. 1.

Generative adversarial network (GAN)-based methods [2], [7], [8] are commonly adopted in the fashion design task. However, GAN-based methods have several limitations. They usually require a large number of training samples to ensure the realism of the new fashion. In addition, it is difficult to generalize the trained model to novel and unseen styles. Furthermore, they can hardly have good control over the appearance and shape of clothes when transferring from non-fashion domain images. Image transfer methods, like neural style transfer (NST), have shown great success in transferring artistic styles [9], [10], [11], [12]. They have more flexibility than GAN-based methods, which can be easily extended to novel-style translation. Recently, diffusion models [13], [14], [15] have been applied in various generative tasks, such as text image generation [16], [17] and image translation [18], due to the realism of their generated results. Some approaches [19], [20] consider both structure and appearance in image transfer. For example, Kwon et al. [19] use a diffusion model and a special structural appearance loss for appearance transfer, which performs well in transforming the appearance between similar objects, such as from zebras to horses and from cats to dogs.

However, there are two main challenges when applying the commonly used image transfer methods to the reference-based fashion design task. First, common image transfer methods have a big degradation when **the source and reference images have large domain gaps**. As shown in the examples of Fig. 1, the source is the handbag image, while the reference is the ocean animal image. The semantic features of reference appearance images are usually far different from the source clothing images. As a result, commonly used image transfer methods usually generate unrealistic fashions in this task and difficult to transfer the appearance. Besides, these methods only transfer the style or appearance, which can hardly convert the appearance to a suitable

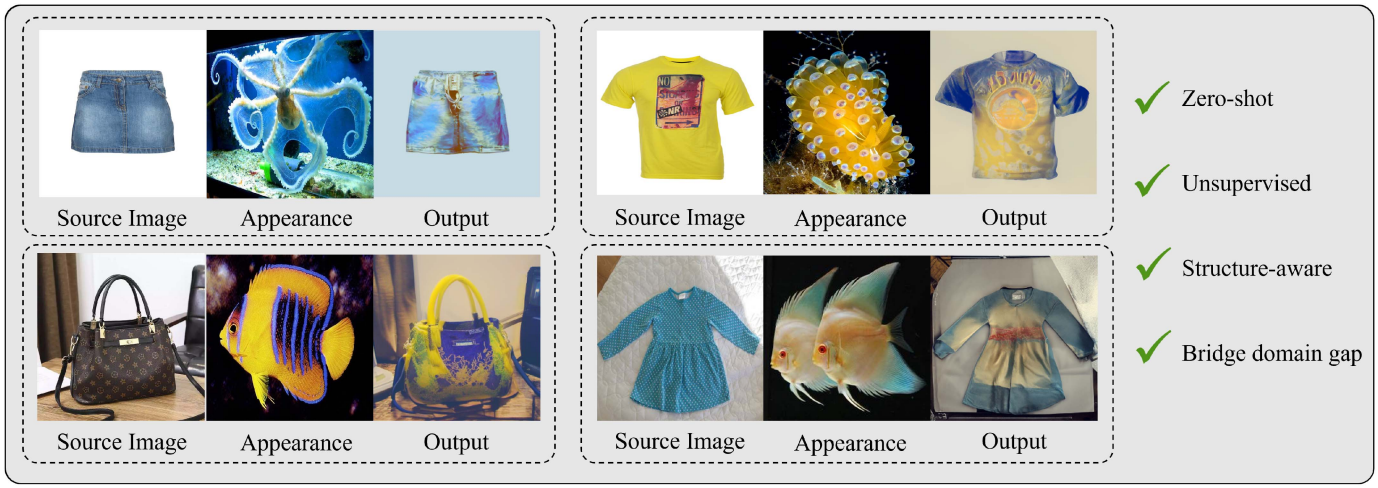


Fig. 1. Examples of the reference-based fashion design task. Given a source clothing image and a referenced appearance image, our method can generate a new fashion clothing image.

texture material by using a non-clothing image. For example, the recently proposed diffusion-based methods [19], is based on the similarity of the semantically related objects in vision transformer (ViT) [21] features, which generates poor results in the reference-based fashion design task. Second, to ensure the realism of the generated results, image transfer methods usually **require a large number of training samples from both source and target domains** [22]. However, there are no samples available for newly designed output domains, resulting in a lack of guidance during the transfer process. Thus, the generated new fashion images are likely to lose the structural information of the input clothing images. To address the aforementioned issues, we propose an unsupervised transfer framework based on diffusion models, namely *DiffFashion*, which semantically generates new fashion clothes from a source clothing image and a reference appearance image. Our method is **free of model tuning in adapting to novel styles and objects**, **structure-preserving**, and has **high flexibility in transferring from images with large domain gaps**. In contrast to existing diffusion-based image translation methods, where images from different domains employ distinct latent embeddings in the denoising process [19], our approach maintains a shared latent representation across clothing images and reference appearance images. This shared latent representation, guided by optimal transport properties, serves to bridge the gap between the two domains, significantly enhancing the transferability capability. In the diffusion reverse process, the latent representation of the reference image is gradually adapted to the clothing domain. Simultaneously, the structure is transferred from the source clothing to the output fashion image with mixed guidance, including the pre-trained Vision Transformer (ViT) guidance and a foreground mask guidance, to further preserve the structure and appearance semantics from source and reference images. As a result, the proposed DiffFashion can learn the patterns from the reference image and the output fashion remains realistic even if there is a big domain gap between source and reference images. The process is illustrated in Fig. 2. Our contributions are summarized as follows:

- We propose a novel unsupervised diffusion model-based fashion design method. As far as we know, this is the first work that extends the diffusion model into the fashion design field.
- We introduce a novel transfer strategy based on the optimal transport properties to gradually adapt the latent from the reference appearance image to the output clothing image, which addresses the issue of large domain gaps and differs from existing diffusion-based image translation methods.
- We adopt the mixed guidance, including the mask guidance and ViT guidance, to further transfer both structure and appearance to the output fashion in the denoising process.
- Extensive experimental zero-shot fashion design results, including *handbags*, *slippers*, *pants* and *hats* fashions, verify that our method achieves state-of-the-art (SOTA) performance in generating new fashions.

The outline of the article is as follows: In Section II, we review state-of-the-art (SOTA) fashion design and image translation methods. Section III introduces the preliminary background of DDPM. We introduce our proposed method in Section IV. The experiments of our proposed method are provided in Section V, followed by the conclusion and future work in Section VII.

II. Related Work

A. Fashion Design

Fashion design models aim to design new clothing from a given clothing collection. Sbai et al. [3] first use GAN to learn the encoding of clothes, and then use the latent vector to perform the stylistic transformation. Recently, there has been an increased focus on intelligent fashion design, with more methods aiming to design new fashion items based on disentangled textures and shapes. Inspired by Conditional GAN [23], Sketched-based methods [5], [8] use the sketch image of the clothes to control the generated structure. However, these methods require additional information input, and using texture propagation will decrease the realism of the generated images.

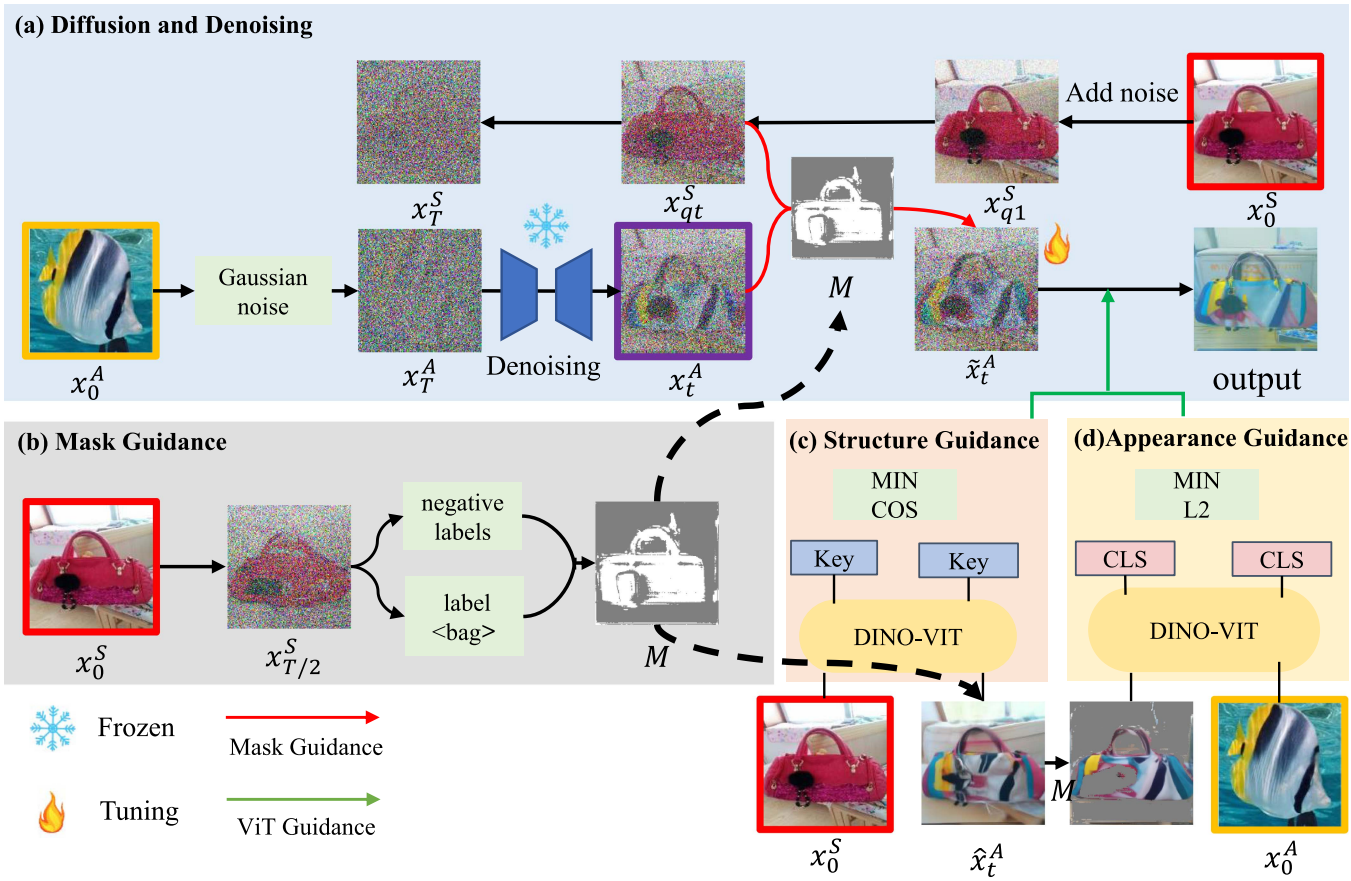


Fig. 2. Pipeline of our approach. (a) Diffusion and Denoising: We denoise the reference appearance image $\mathbf{x}_0^A$. In the denoising process, we use the mask in (b) to replace the background with pixel values obtained from the encoding process at the same timestamp. Mixed guidance is used in the denoising process. (b) Mask Guidance: We add noise to clothing image $\mathbf{x}_0^S$, and then use different label conditions to estimate the noise. The semantic mask of the $\mathbf{x}_0^S$ can be obtained from the noise difference. (c) Structure Guidance and (d) Appearance Guidance: We use DINO-VIT features to compute structure loss between $\mathbf{x}_t^A$ and $\mathbf{x}_0^S$, appearance loss between $\mathbf{x}_t^A$ and $\mathbf{x}_0^A$, to guide the denoising process. The pre-trained model is frozen and the gradient flow is updated directly on the output image.

Jiang et al. [24] first combine the methods of patch-based GAN and NST to extract texture information more comprehensively. Yan et al. [25] combined patch-based GAN and a learnable semantic disentanglement attention-based encoder to better disentangle the structure and texture information, which enhances the authenticity of the generated results. However, training a patch-based GAN requires a large amount of data. In addition, because the transformation is based on the patch similarity, the generated results mainly involve the transfer of overall color and small local textures, rather than the patterns of the reference image. Besides, all these GAN-based methods involve adversarial training, which can be challenging, and these methods are limited to the dataset on which they are trained.

B. Image-to-Image Translation

The image-to-image translation often uses a GAN network to learn the mapping between the source and the target domain. Paired data methods like [26], [27] use the target image corresponding to each input for the condition in the discriminator. Unpaired data methods like [28], [29], [30] decouple the common content space and the specific style space in an unsupervised way. But both these methods require amounts of data from both domains. Besides, the encoding structure of GANs makes it difficult to decouple appearance and structural information. When the gap between the two domains is too large, the result may not be transformed [30], [31], [32] or have lost information from the original domain [33].

Recently, denoising diffusion probabilistic models (DDPMs) have emerged as a promising alternative to GANs in image-to-image translation tasks. Palette [22] firstly applies the diffusion model in image translation and achieves good results in colorization, inpainting, and other tasks. However, this approach requires the target image as a condition for diffusion, making it infeasible for unsupervised tasks. For appearance transfer, DiffuseIT [19] uses the same DINO-ViT guidance as [20], which greatly improves the realism of the transformation. However, it still cannot solve the problem of lacking matching objects in the clothing design task.

C. Neural Style Transfer (NST)

Neural style transfer (NST) has shown great success in transferring artistic styles. There are mainly two types of approaches

to modeling the style or visual texture in NST. One is based on statistical methods [9], [10], in which the style is characterized as a set of spatial summary statistics. The other is based on non-parametric methods, such as using Markov Random Field [11], [12], in which they swap the content neural patches with the most similar ones to transfer the style. After texture modeling, a pre-trained convolutional neural network (CNN) network is used to complete the style transfer. Although NST-based methods work well for global artistic style transfer, their content/style decoupling process is not suitable for fashion design. In addition, NST-based methods assume the transfer is between similar objects or domains. Tumanyan et al. [20] propose a new NST loss from DINO-ViT, which succeeds in transferring appearance between two semantically related objects, such as "cat and dog" or "orange and ball". However, in our task, there are no specific related objects between the clothing image and the appearance image.

III. Preliminary of Denoising Diffusion Probabilistic Models

Diffusion probabilistic models [13], [14], [15] are a type of latent variable model that consists of a forward diffusion process and a reverse diffusion process. In the forward process, we gradually add noise to the data, and then sample the latent $\mathbf{x}_t$ for $t = 1, \ldots, T$ as a sequence. Noise added to data in each step is sampled from a Gaussian distribution, and the transmission can be represented as $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I})$, where the Gaussian variance $\{\beta_t\}_{t=0}^{T}$ can either be learned or scheduled. Importantly, the final latent encoding by the forward process can be directly obtained by,

$$\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{(1-\overline{\alpha}_t)}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Then in the reverse process, the diffusion model learns to reconstruct the data by denoising gradually. A neural network is applied to learn the parameter $\boldsymbol{\theta}$ to reverse the Gaussian transitions by predicting $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ as follow,

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t), \sigma^2 \boldsymbol{I}). \tag{2}$$

To achieve a better image quality, the neural network takes the sample $\mathbf{x}_t$ and timestamp t as input, and predicts the noise added to $\mathbf{x}_{t-1}$ in the forward process instead of directly predicting the mean of $\mathbf{x}_{t-1}$. The denoising process can be defined as,

$$\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right), \tag{3}$$

where $\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ is the diffusion model trained by minimizing the objective, i.e.,

$$\mathcal{L}(\boldsymbol{\theta}) = E_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}[(\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, t))^2]. \tag{4}$$

In the image translation task, there are two mainstream methods to complete the translation. One uses the conditional diffusion model, which takes extra conditions, such as text and labels as input in the denoising process. Then the diffusion model $\epsilon_{\boldsymbol{\theta}}$ in (3) and (4) can be replaced with $\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t, y)$, where y is the condition. The other type of method [34] uses pre-trained classifiers to guide the diffusion model in the denoising process and freezes the weights of the diffusion model. With the diffusion model and a pre-trained classifier $p_{\boldsymbol{\phi}}(y|\mathbf{x}_t)$, the denoising process $\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ in (3) can be supplemented with the gradient of the classifier, i.e., $\hat{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + \sigma_t \nabla log p_{\boldsymbol{\phi}}(y|\mathbf{x}_t)$.

TABLE I
Detailed Description of the Notations Mentioned in This Paper

Notation	Description
$\mathbf{x}^S$	structure image
$\mathbf{x}^A$	appearance image
$\hat{\mathbf{x}}$	predicted initial image from Tweedy's formula
$\mathbf{x}_t$	image at timestamp t in denoising process
$\mathbf{x}_{qt}$	image at timestamp t in forward process
$\boldsymbol{M}$	object mask in the image
y_p	positive class label
y_n	negative class label
$\boldsymbol{M}_p$	positive noise map
$\boldsymbol{M}_n$	negative noise map
$\tilde{\mathbf{x}}$	image after mask guidance

IV. Proposed Method

A. Overview of Fashion Design With DiffFashion

Given a source clothing image $\mathbf{x}_0^S$ and a reference appearance image $\mathbf{x}_0^A$, our proposed *DiffFashion* aims to design a new clothing fashion $\hat{\mathbf{x}}_0$ that can preserve the structure from $\mathbf{x}_0^S$ and the appearance from $\mathbf{x}_0^A$ while keeping it realistic, as shown in Fig. 1. The detailed description of the notation is provided in Table I. Two main challenges need to be addressed when simply applying image translation methods in this setting. Firstly, common image translation methods have a big performance drop when the source and reference images have large domain gaps. Secondly, to ensure the realism of the generated results, image translation methods usually require a large number of training samples from both source and target domains. However, there is no samples available for newly designed output domains, resulting in a lack of guidance during the transfer process.

The proposed DiffFashion addresses these two issues with a novel transfer strategy based on the optimal transport properties and mixed guidance to generate fashion output in the denoising process. The framework is shown in Fig. 2 and details are illustrated in the following sections.

B. Structure-Aware Transfer Diffusion

For diffusion-based translation methods [19], [35], they usually assume the source and target images are within the same or similar domains. In the denoising process, such methods initialize the target image by the input structure or content image to keep the structure similarity high enough, and then transfer the appearance or style based on the loss guidance. However, when there is a large domain gap between the source and reference images, it is difficult to transfer the appearance of the reference image to a new fashion clothing image with the loss of guidance. Besides, when using a natural non-clothing image for appearance reference, the generated texture becomes unrealistic and

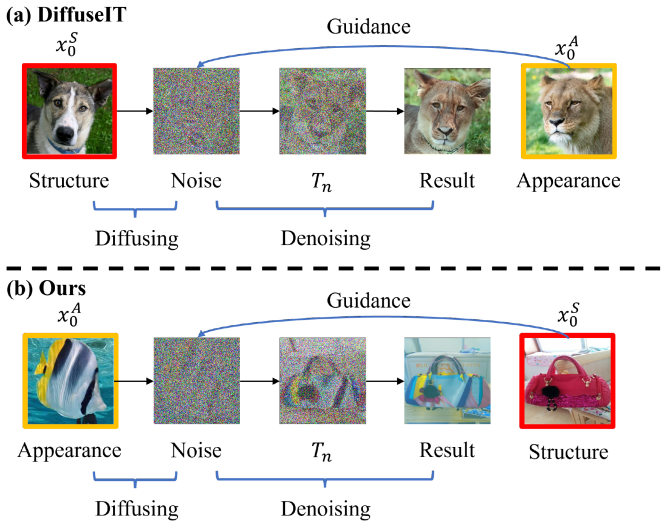


Fig. 3. Transfer difference between DiffuseIT and our method.

not suitable for clothing since these methods cannot well cover the large domain gaps.

To be specific, our proposed DiffFashion addresses the above challenges with a novel structure-aware transfer strategy. Unlike existing diffusion models [19] that take source image $\mathbf{x}_0^S$ as input and transfer the reference appearance image $\mathbf{x}_0^A$ to generate output target image $\hat{\mathbf{x}}_0$, we generate fashion clothes $\hat{\mathbf{x}}_0$ from $\mathbf{x}_0^A$ directly, as shown in Fig. 3. This transfer is guaranteed by the optimal transport properties [36] of the forward procedural encoder of DDPM. Given an original data distribution φ_0, DDPM encoder performs a Monge optimal transport between φ_0 and $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$,

$$\mathbf{x}_T^A = \mathbf{E}_{\varphi_0} = \Sigma(0)^{-0.5}(\mathbf{x}_0^A - \mathbf{a}(0)), \tag{5}$$

with the assumption that φ_0 is an arbitrary multivariate normal distribution $\mathcal{N}(\mathbf{a}(0), \Sigma(0))$, and a quadratic cost function $c(x, y)$ as

$$c(x, y) = c(\mathbf{x}_0^A, \mathbf{x}_T^A) = \left\|\mathbf{x}_0^A - \mathbf{x}_T^A\right\|^2. \tag{6}$$

By solving the Fokker-Planck equation [37] with the low-rank tensor approximations, khrulkov et al. [36] prove that the original data distribution can be generalized to arbitrary distributions with extremely low error.

With the optimal transport properties, it is demonstrated that the same DDPM encoding latent can be transferred into different domains under the conditions with different given labels. In DiffFashion, the same latent is shared for both appearance image $\mathbf{x}_0^A$ and fashion clothes $\hat{\mathbf{x}}_0$. The fashion clothing image has much semantic similarity to the reference appearance image $\mathbf{x}_0^A$. As result, we cover the domain gaps between reference appearance $\mathbf{x}_0^A$ and the fashion clothes $\hat{\mathbf{x}}_0$. Specifically, in each time step t of the denoising process, we use the latent $\mathbf{x}_t^A$ of the reference appearance image to keep more appearance information to the output fashion. The latent of the reference image is gradually adapted to the clothing domain with T time steps, as shown in Fig. 2. Simultaneously, mixed guidance is conducted in the denoising process, which is demonstrated in the next subsection.

C. Mixed Guidance for Denoising

To further adapt the structure and appearance semantics to the output fashion image, we employ mixed guidance in the denoising process, including the pre-trained Vision Transformer (ViT) guidance and a foreground mask guidance. The details are illustrated as follows.

1) ViT Guidance: As mentioned in [19], [20], the structure features and appearance features can be separated from DINO-ViT [21]. We use both appearance and structure guidance in the denoising process to keep the output image realistic and faithful.

Following [19], [20], we employ the global token($[CLS]$ token) in the last layer of ViT to guide the semantic appearance to be similar between the reference image $\mathbf{x}_0^A$ and the generated image $\hat{\mathbf{x}}_t^A$ at t-th step,

$$\mathcal{L}_{\text{app}}(\mathbf{x}_0^A, \hat{\mathbf{x}}_t^A) = \\ ||\boldsymbol{e}_{[CLS]}^L(\mathbf{x}_0^A \boldsymbol{M}) - \boldsymbol{e}_{[CLS]}^L(\hat{\mathbf{x}}_t^A \boldsymbol{M})||_2 + \lambda_{\text{MSE}}||\mathbf{x}_0^A - \hat{\mathbf{x}}_t^A||_2, \tag{7}$$

where $\boldsymbol{e}_{[CLS]}^L$ is the last layer $[CLS]$ token, and λ_{MSE} is the coefficient of global statistic loss between images. To better leverage the appearance between the object and the appearance image, we use the object semantic mask $\boldsymbol{M}$ to remove the background pixel of $\hat{\mathbf{x}}_t^A$ in (7), and only compute the appearance loss of the object within the mask. The generation of the mask is demonstrated in Section IV-C2 later.

In addition, we adopt a patch-wise method in the structural loss design to better leverage the local features. We adopt the i-th key vector in the l-th attention layer of the ViT model, denoted as $\boldsymbol{k}_i^l$, to preserve the structural information of the i-th patch of the source clothing image $\mathbf{x}_0^S$ as follows,

$$\mathcal{L}_{\text{struct}}(\mathbf{x}_0^S, \hat{\mathbf{x}}_t^A) = \\ -\sum_i \log\left(\frac{\text{sim}(\boldsymbol{k}_i^{l,S}, \boldsymbol{k}_i^{l,A})}{\text{sim}(\boldsymbol{k}_i^{l,S}, \boldsymbol{k}_i^{l,A}) + \sum_{j\neq i}\text{sim}(\boldsymbol{k}_i^{l,S}, \boldsymbol{k}_j^{l,A})}\right), \tag{8}$$

where $\text{sim}(\cdot,\cdot)$ is the exponential value of normalized cosine similarity, i.e.,

$$\text{sim}(\boldsymbol{k}_i^{l,S}, \boldsymbol{k}_j^{l,A}) = \exp\left(\cos\left(\boldsymbol{k}_i^l(\mathbf{x}_0^S), \boldsymbol{k}_j^l(\hat{\mathbf{x}}_t^A)\right)/\tau\right), \tag{9}$$

and τ is the temperature parameter. By using the loss in (8), we minimize the loss between keys at the same position of two images while maximizing the loss between keys of different positions. Then our total loss for guidance is as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{struct}}\mathcal{L}_{\text{struct}} + \lambda_{\text{app}}\mathcal{L}_{\text{app}}, \tag{10}$$

where $\lambda_{\text{struct}}, \lambda_{\text{app}}$ are the coefficient of structure loss and appearance loss.

2) Mask Guidance: Rather than only using ViT loss as semantic guidance in the denoising steps, we also generate the foreground mask of the source clothing image as the guidance to provide structural prior to the output fashion. Existing methods commonly use additional inputs (i.e. masks, boxes) to obtain the foreground region. However, this leads to increased annotation expenses. Instead, we propose a mask generation approach that can obtain the foreground mask without additional information or depending on any segmentation models.

In the denoising process of the label-conditional diffusion model, there can be different noise estimates for the same latent given negative label conditions like *phone* and the positive label *bag*. For these different noise estimates, the regions of the foreground objects that are denoised tend to vary little in background regions but greatly in foreground regions. By taking the difference in the noise area, we can obtain the mask of the object to be edited, as shown in Fig. 2(a).

Unlike the editing task that generates the mask with the latent in the forward process like [38], we generate the mask in the denoising process. This is based on the observation that $\mathbf{x}_t^S$ in the reverse process has less perceptual appearance information than $\mathbf{x}_{qt}^S$ (the image in the forward process with timestamp t). Although the structure of $\mathbf{x}_t^S$ may have some slight variations, it still provides a better representation of the overall structure information of the foreground object.

Specifically, we input the source clothing image $\mathbf{x}_0^S$ into the diffusion model. After DDPM encoding in the forward process to achieve $\mathbf{x}_T^S$, we obtain the image latent $\mathbf{x}_{T/2}^S$ in the half of the reverse process. Denote the foreground label as y_p, representing the foreground clothing object. Then the noise map for the foreground clothing can be obtained by

$$\boldsymbol{M}_p = \epsilon_\theta(\hat{\mathbf{x}}_{T/2}^S, T/2, y_p), \tag{11}$$

where $\hat{\mathbf{x}}_{T/2}^S$ is the estimated source image predicted from $\mathbf{x}_{T/2}^S$ by the method of Tweedie [39], i.e.,

$$\hat{\mathbf{x}}_{T/2}^S = \frac{\mathbf{x}_{T/2}}{\sqrt{\bar{\alpha}_{T/2}}} - \frac{\sqrt{1-\bar{\alpha}_{T/2}}}{\sqrt{\bar{\alpha}_{T/2}}}\epsilon_\theta(\mathbf{x}_{T/2}^S, T/2, y_p). \tag{12}$$

Denote non-foreground labels as y_n, representing negative objects. We use N different non-foreground label conditions to get an averaged noise map, i.e.,

$$\boldsymbol{M}_n = \frac{1}{N}\sum_{i=1}^{N}\epsilon_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_{T/2}^S, T/2, y_i), \tag{13}$$

where $i \in \{1, \ldots, N\}$. Then the difference between the two noise maps $\boldsymbol{M}_p$ and $\boldsymbol{M}_n$ can be obtained. A threshold is set for binarization, which returns an editable semantic mask $\boldsymbol{M}$ for the foreground clothing region.

As shown in Fig. 2(b), the appearance image $\mathbf{x}_0^A$ is first used to be encoded by the forward process of DDPM. Then the mask-guided denoising process is employed. Specifically, at each step in the denoising process, we estimate the new prediction $\mathbf{x}_t^A$ from the diffusion model as follows,

$$\mathbf{x}_t^A = \frac{1}{\sqrt{\alpha_{t+1}}}\left(\mathbf{x}_{t+1}^A - \frac{1-\alpha_{t+1}}{\sqrt{1-\bar{\alpha}_{t+1}}}\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_{t+1}^A, t+1, y_p)\right). \tag{14}$$

Then we combine the transferred foreground appearance $\mathbf{x}_t^A$ and the clothing image of corresponding timestamp $\mathbf{x}_{qt}^S$ with the generated mask $\boldsymbol{M}$ as guidance, i.e.,

$$\tilde{\mathbf{x}}_t^A = \boldsymbol{M}\cdot\mathbf{x}_t^A + (1-\boldsymbol{M})\cdot\mathbf{x}_{qt}^S, \tag{15}$$

Algorithm 1: Fashion Generation

Input structure image $\mathbf{x}_0^S$ and appearance image $\mathbf{x}_0^A$, After the inference, the diffusion model generate the result image.
Forward process, add noise to structure image
$t \sim Uniform(\{1, \ldots, T\})$
$\boldsymbol{\sigma} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
$\mathbf{x}_T^S \leftarrow \sqrt{\bar{\alpha}_T}\mathbf{x}_0^S + \sqrt{(1-\bar{\alpha}_T)}\boldsymbol{\epsilon}$
Denoising process of structure image
for $t = 1, \ldots, T$
  $\mathbf{x}_{t-1}^S \leftarrow \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t^S - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t^S, y_p, t))$
end for
Get mask from noise estimation
$\hat{\mathbf{x}}_{T/2}^S \leftarrow$ predicted from $\mathbf{x}_{T/2}^S$ with (12)
$\boldsymbol{M} \leftarrow$
$\epsilon_\theta(\hat{\mathbf{x}}_{T/2}^S, T/2, y_p) - \frac{1}{N}\sum_{i=1}^{N}\epsilon_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_{T/2}^S, T/2, y_i)$
Generation with guidance in denoising process
for $t = T, \ldots, 1$ **do**
  $\mathbf{x}_{t-1}^A \leftarrow \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t^S - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t^S, y_p, t))$
  $\boldsymbol{L} \leftarrow$ ViT Loss Guidance with (7), (8), (9), (10)
  $\mathbf{x}_{t-1}^A \leftarrow \mathbf{x}_{t-1}^A + \nabla_x \boldsymbol{L}$
  Perform ViT guidance
  if $t < kT$ **then**
    $x_{qt}^S \leftarrow \sqrt{\bar{\alpha}_t}\mathbf{x}_0^S + \sqrt{(1-\bar{\alpha}_t)}\boldsymbol{\epsilon}$
    $\mathbf{x}_{t-1}^A \leftarrow \boldsymbol{M}\cdot\mathbf{x}_{t-1}^A + (1-\boldsymbol{M})\cdot\mathbf{x}_{qt}^S$
    Perform Mask guidance
  end if
end for

where ω_{mix} is the mix ratio of the appearance image and the clothing image. This change ensures that the appearance information in the mask is transferred, while other structural information keeps consistent with the clothing image. After T denoising steps, we obtain the final output fashion image $\hat{\mathbf{x}}_0 = \tilde{\mathbf{x}}_0^A$. The whole inference procedure is shown as Algorithm 1.

V. Experiments

In this section, we describe our fashion design dataset and experiment settings. We also demonstrate the qualitative and quantitative results to show the effectiveness of our proposed method.

A. Dataset

1) OceanBag: To our best knowledge, there is no specific reference-based fashion design dataset currently. Thus, we collect a new dataset, namely *OceanBag*, with real handbag images and ocean animal images as reference appearances for generating new fashion designs. *OceanBag* has 6,000 images of handbags in various scenes and 2,400 images of various marine lives in the real world. The 2,400 marine scene images contain more than 80 kinds of marine organisms, including fish, starfish, crabs, algae, and other sea creatures, as shown in Fig. 4. We randomly select 40 pairs of images to conduct the experiment.

Fig. 4. Samples from our proposed dataset of *OceanBag*. The left part shows some examples of marine life images, and the right part shows some samples of handbag images.

2) Other Clothing Datasets: To validate our method on a wider variety of clothing, we also conduct experiments using images from other clothing datasets [40], [41] as our source structure images.

B. Experimental Setup

We conduct all experiments using a label-conditional diffusion model [44] pre-trained on the ImageNet dataset [45] with 256×256 resolution. In all experiments, we use a diffusion step of $T = 120$ and re-sampling repetitions of $N = 10$. In a single RTX 3090 unit, it takes 20 seconds to generate each mask and 60 seconds to generate each image. For fairness of comparison, other parameters in the diffusion model are kept the same as [19].

In the mask generation part, we set the binarization threshold to -0.2. Due to the stochastic nature of the diffusion model, we generate masks using three different sets of negative labels, including "cellphone, forklift, pillow", "waffle iron, washer, Guinea pig" and "brambling, echidna, custard apple". Then we choose the best one among them for guidance. To ensure a fair comparison, we run the baseline DiffuseIT [19] three times as ours.

In the guidance part, to mitigate the uncontrollable effect of the mask and avoid information loss when the structural gap between the two objects is too large, we use mask guidance in the first 30% steps of the denoising stage. In the ViT guidance part, we set the coefficient of the appearance loss λ_{app} to 0.1 and the structure loss λ_{struct} to 1. We keep other parameters and further use color matcher, which is a sinple color transformation method, following [19].

C. Evaluation Methods and Metrics

There is currently no existing automatic metric suitable for evaluating fashion design across two natural images. To keep the fashion image realistic, the migration degree of the appearance and the similarity of the structure sometimes are mutually contradictory when measured. To compare among different methods, we follow existing appearance transfer/fashion design works [19], [20], [46], [47], [48], [49], which rely on human perceptual evaluation to validate the results. Inspired by Tumanyan et al. [20], we also use other pre-trained models to measure the structural similarity of the results. Here, we use Mask-RCNN [50] trained on COCO [51] to detect the generated results and calculate the recall rate and precision of Mask-RCNN. We also utilize perceptual image patch similarity (LPIPS) [52], which is commonly used to measure the similarity between the generated images and the input structure and appearance images. To measure the authenticity of the generated images, we use the Frechet inception distance score (FID) [53] metric to assess the realism and diversity of the generated dataset.

D. Experimental Results

We perform both quantitative and qualitative evaluations on the *OceanBag* dataset and other fashion datasets [40], [41]. We compare our model with state-of-the-art (SOTA) methods, including Splice [20], DiffuseIT [19], WCT2 [42] and SANet [43]. Figs. 5 and 6 show qualitative results on *OceanBag* and other fashions for all methods. In all examples, it can be seen that in terms of fashion design, our method has achieved better performances in terms of realism and structure, while completing appearance transfer. As for the DINO-ViT-based image-to-image translation methods, DiffuseIT successfully keeps the structure for most images, but it shows less appearance similarity to the reference image. Though Splice transfers the appearance well, the generated results are far away from realistic fashion images. Parametric NST methods like WCT2 effectively retain the structure of the source image, but WCT2 outputs exhibit limited changes apart from color adjustments. Non-parametric NST methods like SANet successfully transfer the appearance. However, the results often suffer from color bleeding artifacts and thus show poor authenticity.

We also conduct a user study to evaluate the samples and obtain subjective evaluations from participants. Specifically, we ask 30 users to score all the output fashion images from all methods for each input pair. Detailed questions we have asked are as follows: 1) Is the picture realistic? 2) Does the structure of the image resemble that of the input image? 3) Is the output appearance similar to the input appearance image? The scores range from 0 to 100. The overall score is the average of the three scores. We show the averaged subjective evaluation results in Tables II and IV. Our model obtains the best score in the overall performance and appearance correlation, and the second place in structure similarity and realism. WCT2 shows the best in

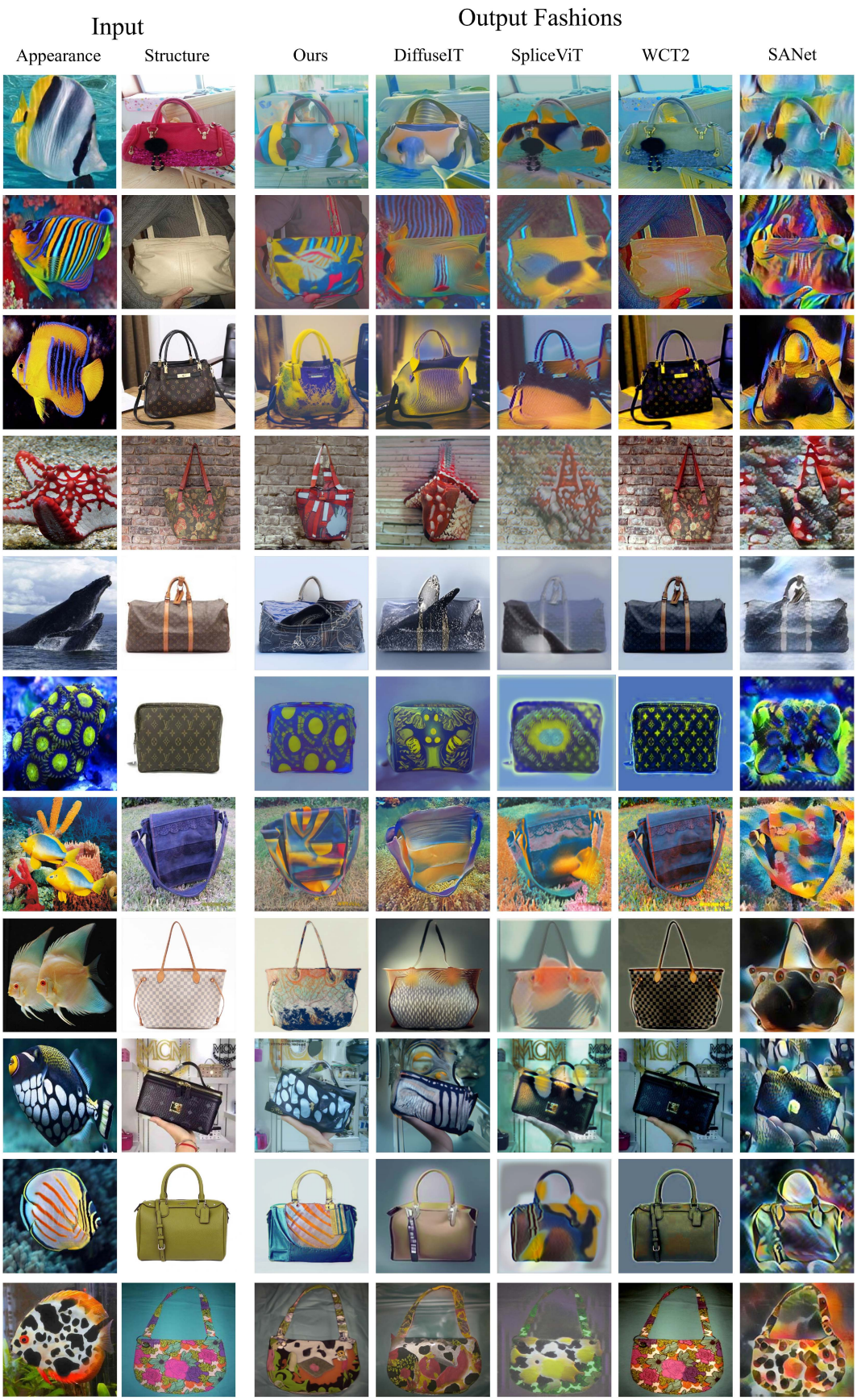


Fig. 5. Comparison with state-of-the-art (SOTA) methods on *OceanBag* dataset.

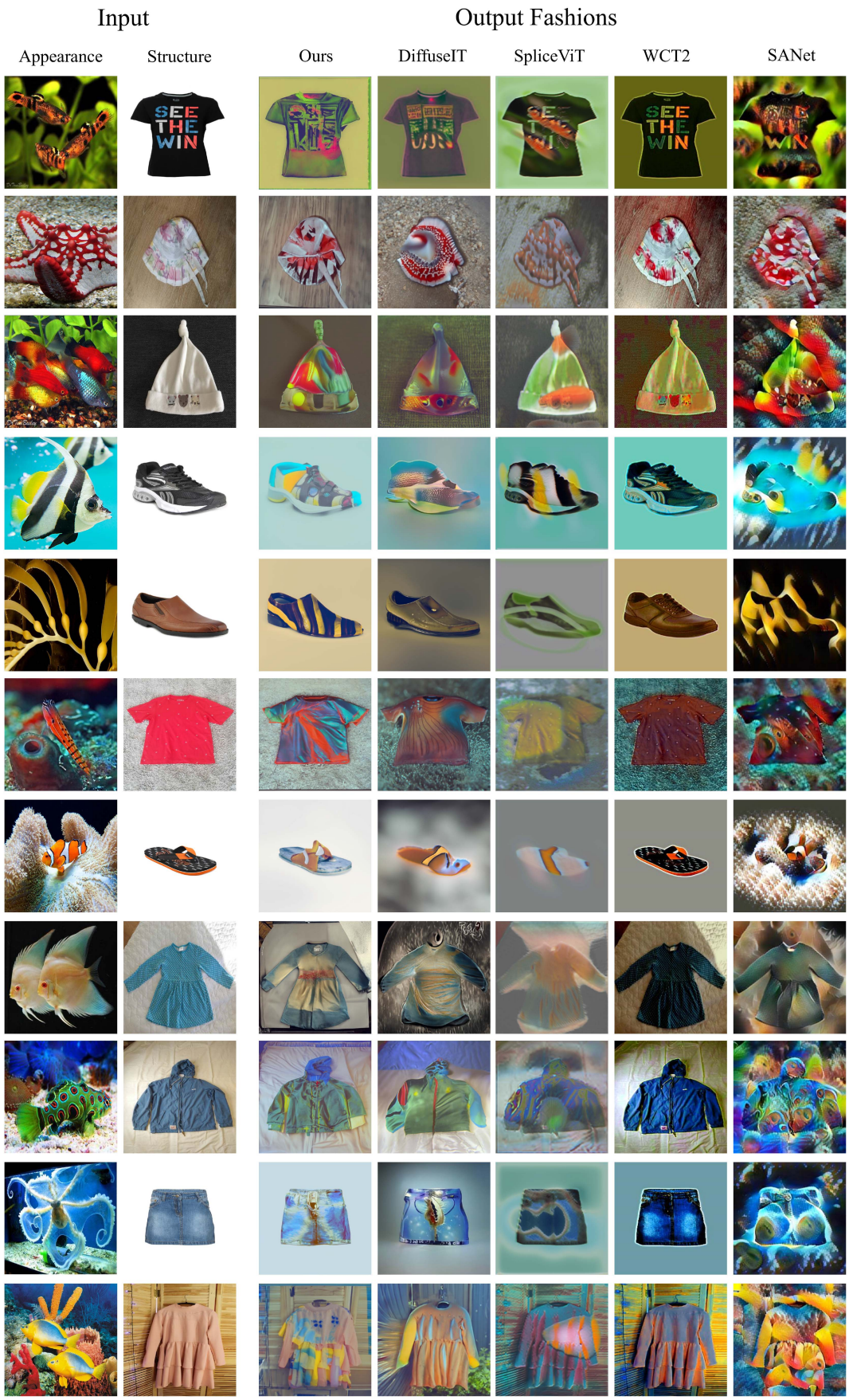


Fig. 6. Comparison with other state-of-the-art (SOTA) methods on other fashions, including clothes, hats, shoes, t-shirts, pants, and dresses.

Fig. 7. Illustration of mask generation by label condition.

TABLE II
RESULTS OF THE USER STUDY ON DATASET *OCEANBAG*

Method	**Overall**↑	Realism↑	Structure↑	Appearance↑
DiffuseIT [19]	67.2	73.4 ± 5.5	88.3 ± 4.8	40.0 ± 6.1
Splice [20]	59.6	65.6 ± 5.9	81.5 ± 5.3	31.9 ± 6.3
WCT2 [42]	66.0	$\mathbf{85.5} \pm 2.4$	$\mathbf{95.9} \pm 1.4$	16.6 ± 3.9
SANet [43]	56.1	47.7 ± 3.3	74.9 ± 4.5	45.7 ± 6.5
Ours	**73.6**	79.8 ± 3.8	91.6 ± 3.2	$\mathbf{49.4} \pm 5.9$

The output fashion images are evaluated based on their realism, structure, and appearance scores, ranging from 0 to 100. The overall performance is the average of the three scores. The best performance is shown in bold and the second best is shown in light blue.

TABLE III
EVALUATION RESULTS BASED ON *OCEANBAG*

Method	L.app↓	L.struct↓	M.rec↑	M.prec↑	FID ↓
DiffuseIT	0.762	0.571	0.15	0.10	200.3
Splice	0.782	0.579	0.03	0.02	258.2
WCT2	0.787	**0.378**	**0.43**	**0.28**	**86.2**
SANet	0.759	0.659	0.03	0.02	318.0
Ours	**0.754**	0.488	0.28	0.26	131.8

The best performance is shown in bold and the second best is shown in light blue. "L.app", "L.struct","M.rec" and "M.prec." represent LPIPS with input appearance image, LPIPS with input clothing image, Mask-RCNN recall, and Mask-RCNN precision, respectively.

TABLE IV
EVALUATION RESULTS BASED ON OTHER FASHION DATASETS

Method	L.app↓	L.struct↓	FID ↓
DiffuseIT	0.771	0.572	237.7
Splice	0.789	0.624	294.7
WCT2	0.784	**0.411**	**93.4**
SANet	0.766	0.708	321.6
Ours	**0.764**	0.523	152.0

The best performance is shown in bold and the second best is shown in light blue. "L.app", "L.struct" represent lpips with input appearance image, LPIPS with input clothing image, respectively.

realism and structure similarity scores, but it shows the worst score in appearance correlation because the outputs are almost unchanged from the inputs except for the overall color. Both the qualitative and subjective evaluations show the effectiveness of our proposed method.

Following [20], we also adopt other pre-trained models to evaluate the result. We use the learned LPIPS score between input and output to verify the content preservation performance and the appearance similarity. We also apply the Mask-RCNN model pre-trained on the COCO dataset to detect the mask of the object of each method. The results are shown in Table III. At the same time, since Mask-RCNN is trained on out-of-distribution (OOD) data, the overall recall rate is quite low. Our model demonstrates the second-best structure preservation performance. Though WCT2 achieves the best in the structure-aware evaluation, it only transforms the color for the whole image. Besides the structure preservation, our method achieves better appearance similarity than other methods. Some NST methods like SANet have a good appearance similarity, but they tend to transfer the whole image with color transformation and lose the authenticity, as shown in Figs. 5 and 6.

TABLE V
EVALUATION ABOUT METHOD WITH OR WITHOUT MASK ON *OCEANBAG*

Method	L.struct↓	L.app↓	M.rec↑	M.prec↑	FID↓
Ours	**0.488**	0.754	0.28	0.26	131.8
w/o. mask	0.668	**0.731**	0.13	0.08	224.9
w/o. ViT-struct	0.556	0.749	0.18	0.15	188.1
w/o. ViT-app	0.493	0.766	**0.30**	**0.29**	**129.2**

"L.app", "L.struct", "M.rec" and "M.prec" represent lpips with appearance image, LPIPS with clothing image, Mask-RCNN recall, and Mask-RCNN precision.

E. Ablation Study

In order to verify the effectiveness of the method, we study the individual components of our method through several ablation studies.

1) Mask Generation: We show some examples of mask generation with label conditions in Fig. 7. Due to the randomness of the diffusion, the generated masks are not perfect. However, these masks can still preserve some of the overall structure prior of the foreground bags. In Fig. 8, we compare the results using generated messy masks, the results with no mask guidance, and the results of DiffuseIT. According to the figure, we can see that using messy masks still makes some improvement in most cases.

2) Mixed Guidance: We conduct experiments to validate the effectiveness of individual guidance in the denoising process. The results are shown in Fig. 9 and Table V. From the results, it can be observed that using mask guidance leads to a significant increase in structure similarity while only causing a slight

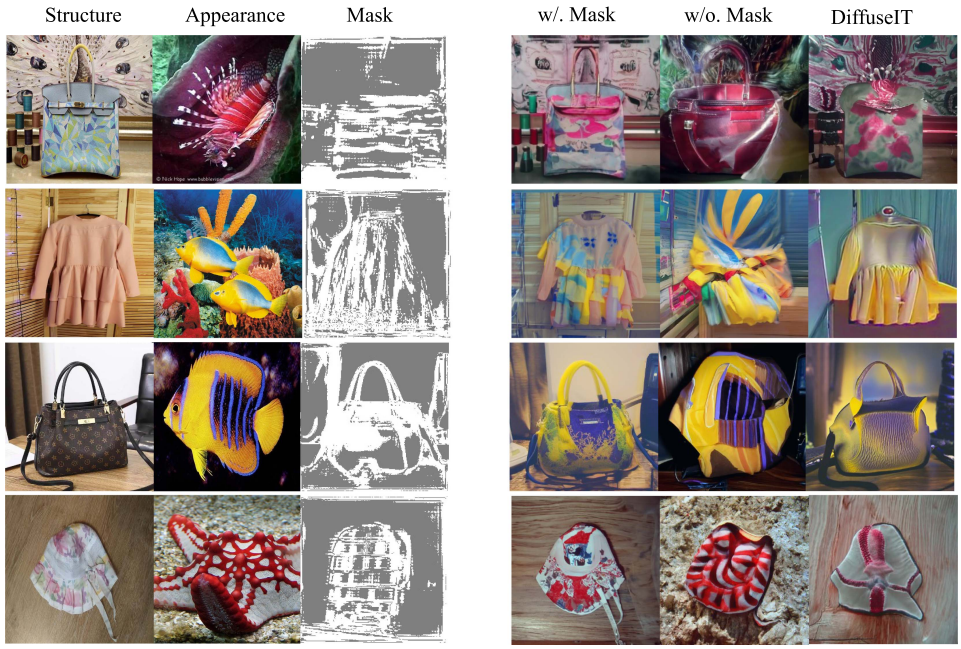


Fig. 8. Left three columns show hard examples with input images and generated messy masks. The right three columns show the fashion output with masks (w/. masks), without masks (w/o. masks), and DiffuseIT, respectively.

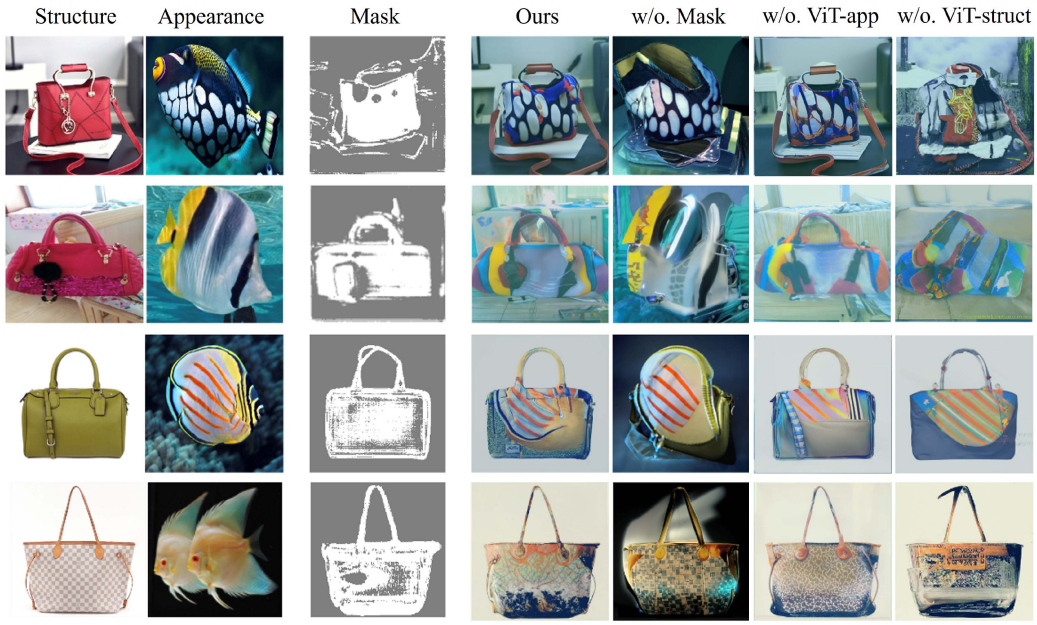


Fig. 9. Left three columns show examples with input images and generated masks. The right four columns show the fashion output with our full model, model without mask, model without ViT appearance guidance, and ViT without structure guidance, respectively.

decrease in appearance similarity. With ViT guidance, the CLS token in the last layer and the key vectors in the attention layers decouple the structure and appearance information very well. Moreover, as assumed in earlier sections, our image generation process mainly relies on the transfer of structural information. This is manifested as using ViT structure guidance leads to a significant increase in structure similarity, while only causing a slight decrease in appearance similarity. In addition, appearance guidance provides a little improvement in the metrics and increases the richness of local image appearance in some images.

3) Sensitivity Analysis: We conduct sensitivity analysis on some of the hyper-parameters including ViT structure guidance, mask threshold, and mask guidance ratio. We randomly sample four appearance images and five structure images to construct twenty pairs for evaluation. As shown in Fig. 11, our results indicate that within a certain range, variations in hyper-parameters

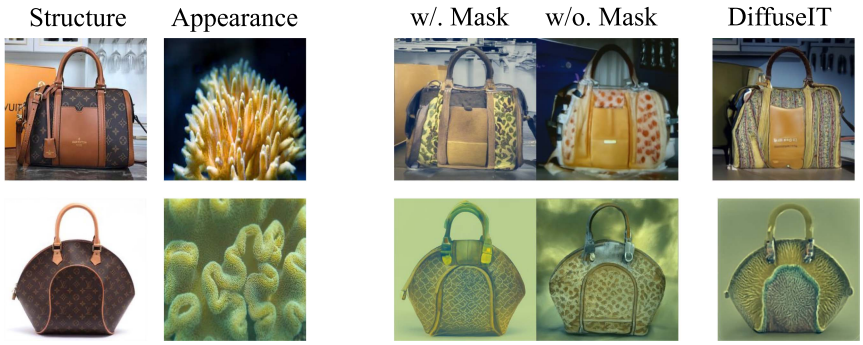
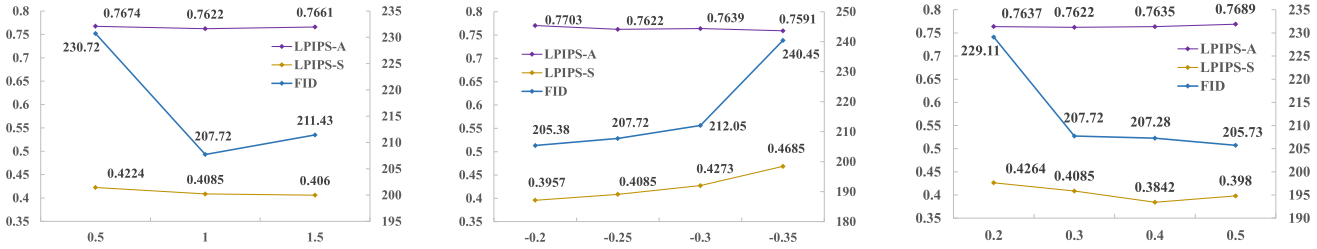


Fig. 10. Examples of failure cases where no mask guidance achieves better results.



(a) ViT structure guidance (b) Mask threshold (c) Mask guidance ratio

Fig. 11. Sensitivity analysis on some of the hyper-parameters. The left vertical axis represents LPIPS, while the right vertical axis represents FID.

do not cause abrupt changes or breakdowns in the results, showing that our algorithm exhibits excellent robustness.

VI. Failure Cases and Limitations

For a small number of cases where the source and reference images have simple or similar positional structures, ViT can be used to achieve structural transfer. In such cases, using a mask can lead to a loss of visual similarity, as shown in Fig. 10. Though our method can generate good fashion output, the method without mask guidance achieves slightly better performance. Also, due to the randomness of the diffusion, the generated masks may need further selection.

VII. Conclusion and Future Work

In this article, we tackle a new fashion design setting: designing new clothing fashion from a given clothing image and a natural appearance image, while keeping the structure of the clothing with a similar appearance to the natural image. We propose a novel diffusion-based image-to-image translation framework by swapping the input latent with structure transfer. And the model is guided by an automatically generated foreground mask and both structure and appearance information from the pre-trained DINO-ViT model. The experimental results have shown that our proposed method outperforms most baselines, demonstrating that our method can better balance authenticity and structure preservation while also achieving appearance migration. In the future, we will try to constrain the diffusion model using the information condition of other modalities to generate better masks.

References

[1] A. Ganesan et al., "Fashioning with networks: Neural style transfer to design clothes," 2017, *arXiv:1707.09899*.
[2] H. Yan, H. Zhang, J. Shi, J. Ma, and X. Xu, "Toward intelligent fashion design: A texture and shape disentangled generative adversarial network," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 19, no. 3, 2023, Art. no. 107.
[3] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, and C. Couprie, "Design: Design inspiration from generative networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 37–44.
[4] B.-K. Kim, G. Kim, and S.-Y. Lee, "Style-controlled synthesis of clothing segments for fashion image manipulation," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 298–310, Feb. 2020.
[5] H. Yan et al., "Toward intelligent design: An AI-based fashion designer using generative adversarial networks aided by sketch and rendering generators," *IEEE Trans. Multimedia*, vol. 25, pp. 2323–2338, 2023.
[6] D. Zhou, H. Zhang, Q. Li, J. Ma, and X. Xu, "CoutfitGAN: Learning to synthesize compatible outfits supervised by silhouette masks and fashion styles," *IEEE Trans. Multimedia*, 2022, early access, Jun. 23, 2023, doi: 10.1109/TMM.2022.3185894.
[7] C. Yuan and M. Moghaddam, "Garment design with generative adversarial networks," 2020, *arXiv:2007.10947*.
[8] Y. R. Cui, Q. Liu, C. Y. Gao, and Z. Su, "FashionGAN: Display your fashion design using conditional generative adversarial nets," *Comput. Graph. Forum*, vol. 37, pp. 109–119, 2018.
[9] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.
[10] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.
[11] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2479–2486.
[12] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 702–716.
[13] J. Ho, A. Jain, and P Abbeel, "Denoising diffusion probabilistic models," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[14] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Representations*, 2020.
[15] Y. Song et al., "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Representations*, 2020.
[16] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, *arXiv:2204.06125*.
[17] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.
[18] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *Proc. 11th Int. Conf. Learn. Representations*, 2022.
[19] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," in *Proc. Int. Conf. Learn. Representations*, 2023.
[20] N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel, "Splicing ViT features for semantic appearance transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10748–10757.
[21] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep ViT features as dense visual descriptors," 2021, *arXiv:2112.05814*.
[22] C. Saharia et al., "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH Conf.*, 2022, pp. 1–10.
[23] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 2089–2093.
[24] S. Jiang, J. Li, and Y. Fu, "Deep learning for fashion style generation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4538–4550, Sep. 2022.
[25] H. Yan, H. Zhang, J. Shi, J. Ma, and X. Xu, "Toward intelligent fashion design: A texture and shape disentangled generative adversarial network," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 19, no. 3, pp. 1–23, 2023.
[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.
[27] J.-Y. Zhu et al., "Toward multimodal image-to-image translation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.
[28] H. Dong, P. Neekhara, C. Wu, and Y. Guo, "Unsupervised image-to-image translation with generative adversarial networks," 2017, *arXiv:1701.02676*.
[29] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.
[30] K. Saito, K. Saenko, and M.-Y. Liu, "COCO-FUNIT: Few-shot unsupervised image translation with a content conditioned style encoder," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 382–398.
[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
[32] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.
[33] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Unsupervised image-to-image translation with generative prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18332–18341.
[34] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794.
[35] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2426–2435.
[36] V. Khrulkov and I. Oseledets, "Understanding DDPM latent codes through optimal transport," in *Proc. Int. Conf. Learn. Representations*, 2023.
[37] H. Risken and H. Risken, *Fokker-Planck Equation*. Berlin, Germany: Springer, 1996.
[38] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "DiffEdit: Diffusion-based semantic image editing with mask guidance," in *Proc. Int. Conf. Learn. Representations*, 2023.
[39] K. Kim and J. C. Ye, "Noise2score: Tweedie's approach to self-supervised image denoising without clean images," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, pp. 864–874.
[40] P. Aggarwal, "Fashion product images (small), data retrieved from Kaggle," 2018. [Online]. Available: https://www.kaggle.com/paramaggarwal/fashion-product-images-small
[41] A. Grigorev, "Clothing dataset (full, high resolution), data retrieved from Kaggle," 2020. [Online]. Available: https://www.kaggle.com/datasets/agrigorev/clothing-dataset-full
[42] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9036–9045.
[43] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5880–5888.
[44] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.
[45] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[46] Y. Jing et al., "Neural style transfer: A review," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 11, pp. 3365–3385, Nov. 2020.
[47] S. S. Kim, N. Kolkin, J. Salavon, and G. Shakhnarovich, "Deformable style transfer," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 246–261.
[48] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 768–783.
[49] T. Park et al., "Swapping autoencoder for deep image manipulation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, pp. 7198–7211.
[50] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[51] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2017.

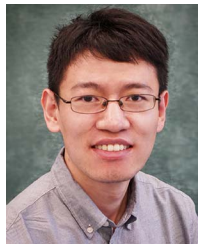
Shidong Cao received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China. He is currently working toward the M.S. degree with Zhejiang University - University of Illinois Urbana-Champaign Institute, Zhejiang University, Hangzhou, China. His research interests include machine learning and computer vision.

Wenhao Chai received the B.E. degree from Zhejiang University, Hangzhou, China. He is currently working toward the master's degree with the University of Washington, Seattle, WA, USA. His research interests include generative models and multimodality learning.

Shengyu Hao (Graduate Student Member, IEEE) received the M.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China. He is currently working toward the Ph.D. degree with Zhejiang University - University of Illinois Urbana-Champaign Institute, Zhejiang University, Hangzhou, China. His research interests include machine learning and computer vision.

Yanting Zhang received the B.E. and Ph.D. degrees from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, in 2015 and 2020, respectively. From 2018 to 2019, she was a visiting Scholar with the University of Washington, Seattle, WA, USA. She is currently an Assistant Professor with the School of Computer Science Technology, Donghua University, Shanghai, China. Her research interests include computer vision and video/image processing.

Hangyue Chen received the master's degree in industrial design engineering from Zhejiang University, Hangzhou, China, in 2010. He is currently a Lecturer with the Department of Digital Media Art, Hangzhou Dianzi University, Hangzhou, China. His research interests include product digital design and human-computer interaction design.



Gaoang Wang (Member, IEEE) received the B.S. degree from Fudan University, Shanghai, China, in 2013, the M.S. degree from the University of Wisconsin-Madison, Madison, WI, USA, in 2015, and the Ph.D. degree from the Information Processing Laboratory, Electrical and Computer Engineering Department, University of Washington, Seattle, WA, USA, in 2019. In 2020, he joined the International Campus of Zhejiang University, Hangzhou, China, as an Assistant Professor. He is also an Adjunct Assistant Professor with the University of Illinois Urbana-Champaign Champaign, Champaign, IL, USA. In 2019, he joined Megvii US Office as a Research Scientist working on multi-frame fusion. In 2019, he then joined Wyze Labs working on deep neural network design for edge-cloud collaboration. He has authored or coauthored papers in many renowned journals and conferences, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, European Conference on Computer Vision, ACM Multimedia, and International Joint Conference on Artificial Intelligence. His research interests include computer vision, machine learning, artificial intelligence, multi-object tracking, representation learning, and active learning.