DiffFashion: Reference-based Fashion Design with Structure-aware Transfer by Diffusion Models

Shidong Cao*, Wenhao Chai*, Shengyu Hao, 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 a new fashion design task, where we aim to transfer a reference appearance image onto a clothing image while preserving the structure of the clothing image. It is a challenging task since there are no reference images available for the newly designed output fashion images. Although diffusion-based image translation or neural style transfer (NST) has enabled flexible style transfer, it is often difficult to maintain the original structure of the image realistically during the reverse diffusion, especially when the referenced appearance image greatly differs from the common clothing appearance. To tackle this issue, we present a novel diffusion model-based unsupervised structure-aware transfer method to semantically generate new clothes from a given clothing image and a reference appearance image. In specific, we decouple the foreground clothing with automatically generated semantic masks by conditioned labels. And the mask is further used as guidance in the denoising process to preserve the structure information. Moreover, we use the pre-trained vision Transformer (ViT) for both appearance and structure guidance. Our experimental results show that the proposed method outperforms state-of-the-art baseline models, generating more realistic images in the fashion design task. Code and demo can be found at https://github.com/Rem105-210/DiffFashion.**

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

I. Introduction

Image-based fashion design with artificial intelligence (AI) techniques [1]–[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 clothing image, a designer may want to generate a new clothes 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] can be adopted in the common fashion design tasks to generate new clothes. However, GAN-based methods can hardly have good control over the appearance and shape of clothes when transferring from non-fashion domain images. Recently, diffusion models [9]–[11] have been widely explored due to the realism and diversity of their results, and have been applied in various generative areas, such as text image generation [12], [13] and image translation [14]. Some approaches [15], [16] consider both structure and appearance in image transfer. Kwon et al. [15] 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 shown in Fig. 1. First, common image transfer methods only consider the translation between semantically similar images or objects. For example, the transformation in [15] is based on the similarity of the semantically related objects in vision transformer (ViT) [17] features. In the reference-based fashion design task, the semantic features of reference appearance images are always far different from 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 hardly converts the appearance to a suitable texture material by using a non-clothing image. Second, image transfer methods [18] usually require a large number of samples from both source and target domains. 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 structure-aware transfer framework based on diffusion named *DiffFashion*, which semantically generates new clothes from a given clothing image and a reference appearance image. The proposed framework is based on denoising diffusion probabulistic models (DDPM) [9] and preserves the structural information of the input clothing image when transferring the reference appearance with three steps. First, we decouple the foreground clothing with automatically generated semantic masks by conditioned labels. Then, we encode the appearance image with DDPM which is proven to be the optimal transport process to keep the high-appearance similarity and denoise the image with mask guidance to transfer the structural information. Moreover, we use the ViT for

* Equal contribution.

† Corresponding author: Hangyue Chen, and Gaoang Wang.

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

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

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

Gaoang Wang is with the Zhejiang University-University of Illinois Urbana-Champaign Institute, and College of Computer Science and Technology, Zhejiang University, China (e-mail: gaoangwang@intl.zju.edu.cn).

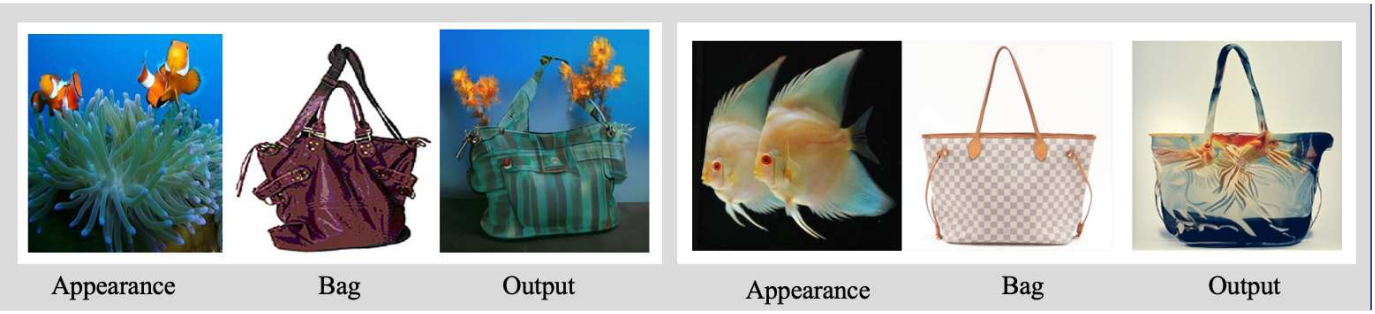

Fig. 1. Two examples of a reference-based fashion design task. For a given image pair, *i.e.*, a bag and a referenced appearance image, our method can generate a new image with appearance similarity to the appearance image and structure similarity to the bag image.

both appearance and structure guidance during the denoising process. This process is illustrated in Fig. 2.

Our contributions are summarized as follows:

- We propose a novel structure-aware image transfer framework, which generates structure-preserving fashion designs without knowledge about output domains.
- We keep the appearance information by the optimal transport properties of the DDPM encoder.
- We employ mask guidance and ViT guidance to transfer structural information in the denoising process.
- Extensive experimental results verify that our method achieves state-of-the-art (SOTA) performance in clothing design.

The outline of the paper 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 VI.

II. Related Work

A. Fashion Design

Fashion design models aim to design new clothing from a given clothing collection. Sbai et al. [3] use GAN to learn the encoding of clothes, and then use the latent vector to perform the stylistic transformation. Cui et al. [8] use the sketch image of the clothes to control the generated structure. Good results have been achieved in terms of structural control. Yan et al. [2] use a patch-based structure to implement texture transfer on generated objects. However, they cannot use other images as texture references and their tasks is limited to generating new samples from existing clothes collections. As a result, due to the unreliable training problem of GAN, more advanced methods are needed to achieve improved realism in generated effects.

B. GAN-based Image Transfer

The image-to-image translation aims to learn the mapping between the source and the target domains, often using a GAN network. Paired data methods like [19], [20] use the target image corresponding to each input for the condition in the discriminator. Unpaired data methods like [21]–[23] 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 [23]–[25] or have lost information from the original domain [26].

C. Diffusion Model-based Image Transfer

Recently, denoising diffusion probabilistic models (DDPMs) have emerged as a promising alternative to GANs in image-to-image translation tasks. Palette [18] 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 [15] uses the same DINO-ViT guidance as [16], which greatly improves the realism of the transformation. However, it still cannot solve the problem of lacking matching objects in the clothing design task.

D. 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 [27], [28], 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 [29], [30], 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. [16] 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 [9]–[11] 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 x_t for $t = 1, ..., T$ as a sequence. Noise added to data in each step is sampled from a Gaussian distribution, and the transmission can be represented as $q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$, 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,

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{(1-\overline{\alpha}_t)}\epsilon, \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

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 θ to reverse the Gaussian transitions by predicting x_{t-1} from x_t as follow:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma^2 I). \tag{2}$$

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

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)), \tag{3}$$

where $\epsilon_\theta(x_t, t)$ is the diffusion model trained by optimizing the objective, *i.e.*,

$$min_\theta \mathcal{L}(\theta) = E_{t,x_0,\epsilon}[(\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0) + \sqrt{1-\overline{\alpha}_t}\epsilon, t))^2]. \tag{4}$$

In the image translation task, there are two mainstream methods to complete the translation. One is using the conditional diffusion model, which takes extra conditions, such as text and labels as input in the denoising process. Then the diffusion model ϵ_θ in Eq. (3) and Eq. (4) can be replaced with $\epsilon_\theta(x_t, t, y)$, where y is the condition. The other type of method [31] 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 $p_\phi(y|x_t)$, the denoising process $\mu_\theta(x_t, t)$ in Eq. (3) can be supplemented with the gradient of the classifier, *i.e.*, $\hat{\mu}_\theta(x_t, t) = \mu_\theta(x_t, t) + \sigma_t \nabla log p_\phi(y|x_t)$.

IV. Proposed Method

A. Overview of Fashion Design with DiffFashion

Given a clothing image x_0^S and a reference appearance image x_0^A, our proposed *DiffFashion* aims to design a new clothing fashion that preserves the structure in x_0^S and transfers the appearance from x_0^A while keeping it natural, as shown in Fig. 2. We list two main challenges in this task. First, there are no given reference images for the output result since there is no standard answer for fashion design. Without the supervision of the ground truth, it is difficult to train the model. Second, preserving the structure information from the given input clothing image while transferring the appearance is also being under-explored. To address those two challenges, we present the *DiffFashon*, which is a novel structure-aware transfer model with the diffusion model. We use the diffusion model [32] pre-trained on Imagenet [33] for all the denoising processes in DiffFashion. First, we decouple the foreground clothing with a generated semantic mask by conditioned labels, as shown in Fig. 2 (a). Then, we encode the appearance image x_0^A with DDPM, and denoise it with mask guidance to preserve the structure information, as shown in Fig. 2 (b). Moreover, we use the DINO-ViT [17] for both appearance and structure guidance during the denoising process, as shown in Fig. 2 (c) and (d). The details are illustrated in the following sections.

B. Mask Generation by Label Condition

To decouple the foreground clothing and background, we generate a semantic mask for the input clothing image x_0^S with label conditions. The generated semantic mask is also used for preserving the structure information in later steps. Existing methods commonly use additional inputs to obtain the foreground region. However, this leads to increased annotation expenses. Inspired by [34], we propose a mask generation approach that can obtain the foreground clothing area without external information or segmentation models. Our approach leverages the label-conditional diffusion model to obtain the desired result.

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 *bag*. For these different noise estimates, the regions of the foreground object that are denoised tend to vary little in background regions but greatly in object 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).

Instead of generating a mask with the latent of the forward process like [34], we observe that in the denoising process, x_t^S has less perceptual appearance information than x_{qt}^S (the image in the forward process with timestamp t). Therefore, we generate a mask from the image in the denoising process x_t^S instead of the image x_{qt}^S in the forward process. Although the structure of 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 clothing image x_0^S into the diffusion model. After DDPM encoding in the forward process, we obtain the image latent $X_{T/2}^S$ in 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

$$M_p = \epsilon_\theta(\hat{x}_{T/2}^S, T/2, y_p), \tag{5}$$

where $\hat{x}_{T/2}^S$ is the estimated source image predicted from $x_{T/2}^S$ by Tweedie's method [35], *i.e.*,

$$\hat{x}_t = \frac{x_{T/2}}{\sqrt{\bar{\alpha}_{T/2}}} - \frac{\sqrt{1-\bar{\alpha}_{T/2}}}{\sqrt{\bar{\alpha}_{T/2}}}\epsilon_\theta(x_{T/2}, T/2, y_p). \tag{6}$$

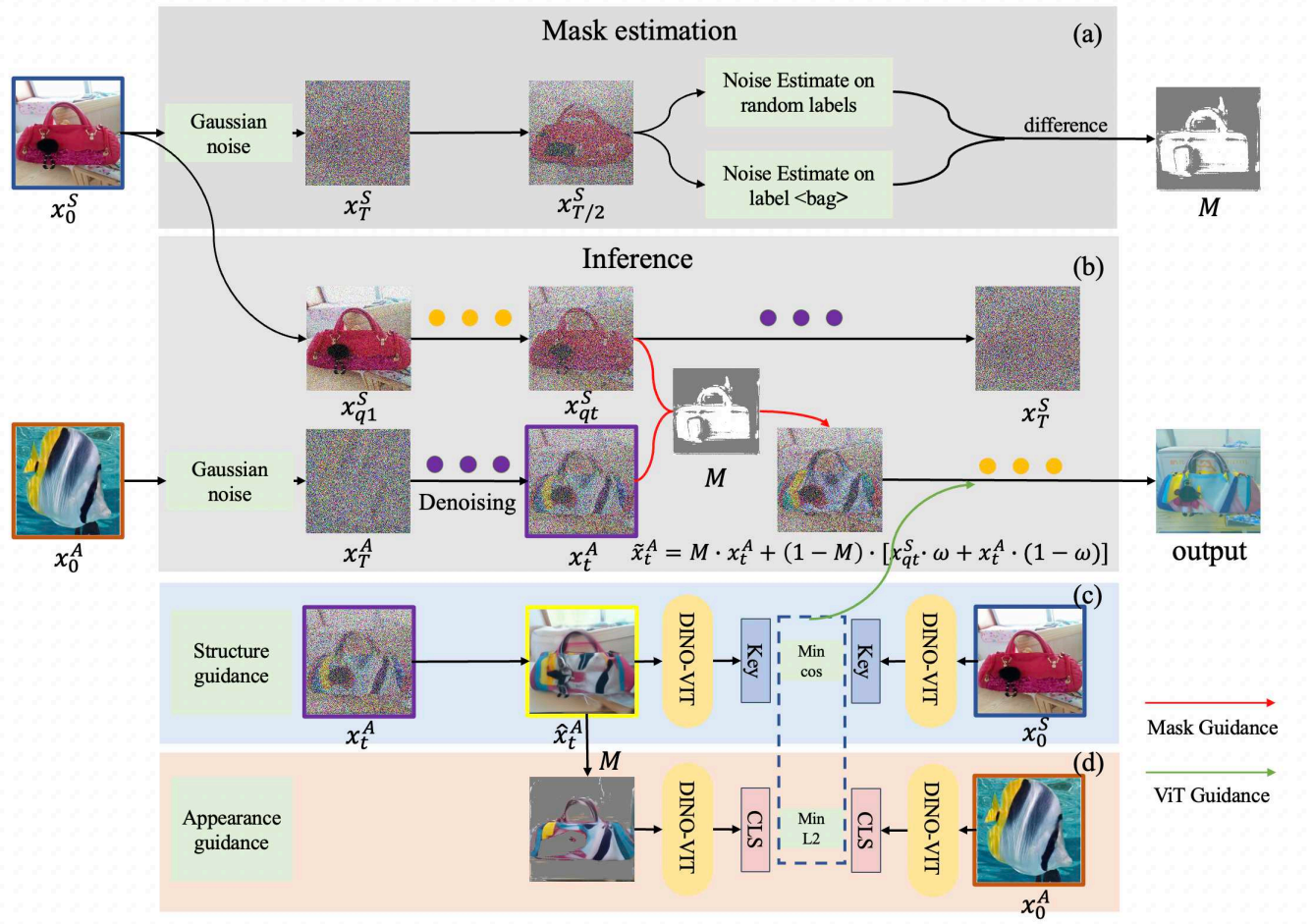

Fig. 2. The pipeline of our approach. (a): We add noise to clothing image x_0^S, and then use different label conditions to estimate the noise in the denoising process. The semantic mask of the x_0^S can be obtained from the noise difference. (b): We denoise the reference appearance image x_0^A. In the denoising process, we use the mask in (a) to replace the background with pixel values obtained from the encoding process at the same timestamp. (c) and (d): We use DINO-VIT features to compute structure loss between x_t^A and x_0^S, appearance loss between x_t^A and x_0^A, to guide the denoising process. Purple dots and yellow dots represent the denoising process with the same timesteps respectively.

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.*,

$$M_n = \frac{1}{N}\sum_{i=1}^{N}\epsilon_\theta(\hat{x}^S_{T/2}, T/2, y_i), \tag{7}$$

where $i \in \{1, ..., N\}$. The difference between the two noise maps M_p and M_n can be obtained. Then we set a threshold for binarization, which returns an editable semantic mask M for the foreground clothing region.

C. Mask-guided Structure Transfer Diffusion

It is difficult to transfer the appearance of the original image to a new fashion clothing image when the gap between the two domains is too large [16]. Because such methods control the appearance by a single loss of guidance, the redundant appearance information of the structure clothing reference image cannot be completely eliminated. Besides, when using a natural non-clothing image for appearance reference, the generated texture maybe not be suitable for clothing. Because these models only transfer the style or appearance. The appearance cannot be converted to a suitable texture material like cotton for clothing. In DiffFashion, to address this problem, rather than transferring from the input clothing image x_0^S, we transfer from the reference appearance image x_0^A to the output fashion clothing image with the guidance of the structural information of the input clothing image.

Inspired by [36], it has been shown that for the same DDPM encoding latent with different label conditions used for denoising, the resulting natural images have similar textures and semantic structures. We use the latent x_t^A of the reference appearance image to transfer more appearance information to the output fashion. Besides, the texture of the appearance image can be transferred more realistic and suitable for clothing in the denoising process. Meanwhile, the semantic mask M obtained from the previous step is used to preserve the structure of the clothing image. As shown in Fig. 2(b), the appearance image x_0^A is first used to encode 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 x_t^A from the diffusion model as follows,

$$x_t^A = \frac{1}{\sqrt{\alpha_{t+1}}}(x_{t+1}^A - \frac{1-\alpha_{t+1}}{\sqrt{1-\bar{\alpha}_{t+1}}}\epsilon_\theta(x_{t+1}^A, t+1, y_p)). \quad (8)$$

Then we combine the transferred foreground appearance x_t^A and the clothing image of corresponding timestamp x_{qt}^S with the generated mask M as guidance, *i.e.*,

$$\tilde{x}_t^A = M \cdot x_t^A + (1-M) \cdot [\omega_{mix} \cdot x_{qt}^S + (1-\omega_{mix}) \cdot x_t^A], \quad (9)$$

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.

D. ViT Feature Guidance

As mentioned in [15], [16], the structure features and appearance features can be separated with DINO-ViT [17]. We use both appearance guidance and structure guidance in the denoising process to keep the output image realistic.

Following [15], [16], we employ the $[CLS]$ tokens in the last layer of ViT to guide the semantic appearance information as follows,

$$\begin{aligned}&\mathcal{L}_{app}(x_0^A, \hat{x}_t^A) = \\ &||e_{[CLS]}^L(x_0^A) - e_{[CLS]}^L(\hat{x}_t^A)||_2 + \lambda_{MSE}||x_0^A - \hat{x}_t^A||_2,\end{aligned} \quad (10)$$

where $e_{[CLS]}^L$ 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 M to remove the background pixel of $\hat{x}_t^A$ in Eq. 10, and only compute the appearance loss of the object within the mask.

In addition, we adopt a patch-wise method in the structural loss 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 $k_i^l(x_t)$, to guide the structural information of the i-th patch of the original clothing image as follows,

$$\begin{aligned}&\mathcal{L}_{struct}(x_0^A, \hat{x}_t^A) = \\ &-\sum_i \log\left(\frac{\text{sim}(k_i^{l,S}, k_i^{l,A})}{\text{sim}(k_i^{l,S}, k_i^{l,A}) + \sum_{j\neq i}\text{sim}(k_i^{l,S}, k_j^{l,A})}\right),\end{aligned} \quad (11)$$

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

$$\text{sim}(k_i^{l,S}, k_j^{l,A}) = \exp\left(\cos\left(k_i^l(x_0^S), k_j^l(\hat{x}_t^A)\right)/\tau\right), \quad (12)$$

and τ is the temperature parameter. By using the loss in Eq. (11), 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 as follow:

$$\mathcal{L}_{total} = \lambda_{struct}\mathcal{L}_{struct} + \lambda_{app}\mathcal{L}_{app}, \quad (13)$$

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

TABLE I
OVERALL INFORMATION ABOUT THE *OceanBag* DATASET.

Dataset	Quantity	Image size	Complex ratio
Handbag	6,000	256×256	0.16
Marine life	2,400	256×256	0.43

TABLE II
RESULTS OF THE USER STUDY. 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.

Method	**Overall**	Realism	Structure	Appearance
DiffuseIT	67.09	75.53±4.68	88.46±5.40	37.27±8.36
SpliceViT	60.98	68.44±4.36	80.80±7.82	33.70±8.32
WCT2	65.45	**82.89**±5.42	**95.76**±1.84	17.69±5.77
STROTSS	63.00	63.33±6.43	82.55±6.65	43.11±8.93
Ours	**75.04**	81.15±4.76	91.07±3.82	**52.89**±7.92

TABLE III
EVALUATION RESULTS BASED ON OTHER MODELS. THE BEST PERFORMANCE IS SHOWN IN BOLD AND THE SECOND BEST IS SHOWN IN LIGHT BLUE. "C.LOSS", "M.RECALL" AND "M.PREC." REPRESENT CLASSIFICATION LOSS, MASK-RCNN RECALL, AND MASK-RCNN PRECISION, RESPECTIVELY.

Method	C.loss	M.recall	M.precision	CDH
DiffuseIT	7.62 ± 4.38	0.17	0.16	0.13
SpliceViT	6.03 ± 3.68	0.03	0.03	0.07
WCT2	9.56 ± 4.16	**0.53**	**0.51**	0.23
STROTSS	11.20 ± 3.52	0.06	0.06	**0.43**
Ours	**5.93** ± 4.72	0.2	0.17	0.22

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

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 photos of handbags in various scenes and 2,400 pictures of various marine lives in the real world, among various marine scenes such as fish swimming on the ocean floor. The 2,400 marine scene images contain more than 80 kinds of marine organisms, 50% of which are fish, as well as starfish, crabs, algae, and other sea creatures, as shown in Fig. 3. In our experiments, we screened 30 images for experiments based on diversity such as background complexity, species, and quantity of organisms.

We refer to images with solid backgrounds as simple backgrounds, while those with real scenes are referred to as complex backgrounds. The complex ratio in Table I shows the proportion of complex background images in the dataset. The

Fig. 3. 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 bag images.

complex background of the marine biological dataset is usually real ocean pictures such as the seabed and the deep sea. For the bag images in the dataset, the complex backgrounds often include scenes of mall containers or tables.

B. Experimental Setup

We conduct all experiments using a label-conditional diffusion model [32] pre-trained on the ImageNet dataset [33] with 256×256 resolution. In all experiments, we use a diffusion step of $T = 60$ and re-sampling repetitions of $N = 10$. In a single RTX 3090 unit, it takes 20 seconds to generate each mask and 120 seconds to generate each image. For fairness of comparison, other parameters in the diffusion model are kept the same as [15].

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 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 [15] 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 50% steps of the denoising stage, and the mix ratio ω_{mix} is set to 0.98. In the ViT guidance part, we set the coefficient of appearance loss λ_{app} to 0.1 and 1 for structure loss λ_{struct}. And we keep other parameters the same as DiffuseIT [15].

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 [15], [16], [37]–[40], which rely on human perceptual evaluation to validate the results.

D. Experimental Results

We perform both quantitative and qualitative evaluations on the *OceanBag* dataset. We compare our model with SplicingViT [16], DiffuseIT [15] , WCT2 [41] and STROTSS [42]. Fig. 4 shows qualitative results 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. SplicingViT transfers the appearance well, but its results are far away from realistic fashion images. NST methods like STROTSS and WCT2 effectively retain the structure of the source image, but WCT2 outputs exhibit limited changes apart from color adjustments. Although STROTSS successfully transfers the appearance, its results often suffer from color bleeding artifacts and thus show less 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) Is the image's structure similar to 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 Table II. 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 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 [16], we also adopt other pre-trained models to evaluate the result. We use the classifier pre-trained with the ImageNet dataset given by improved DDPM [32] and calculate the average classification loss. We also apply Mask-RCNN pre-trained on the COCO dataset to detect the mask

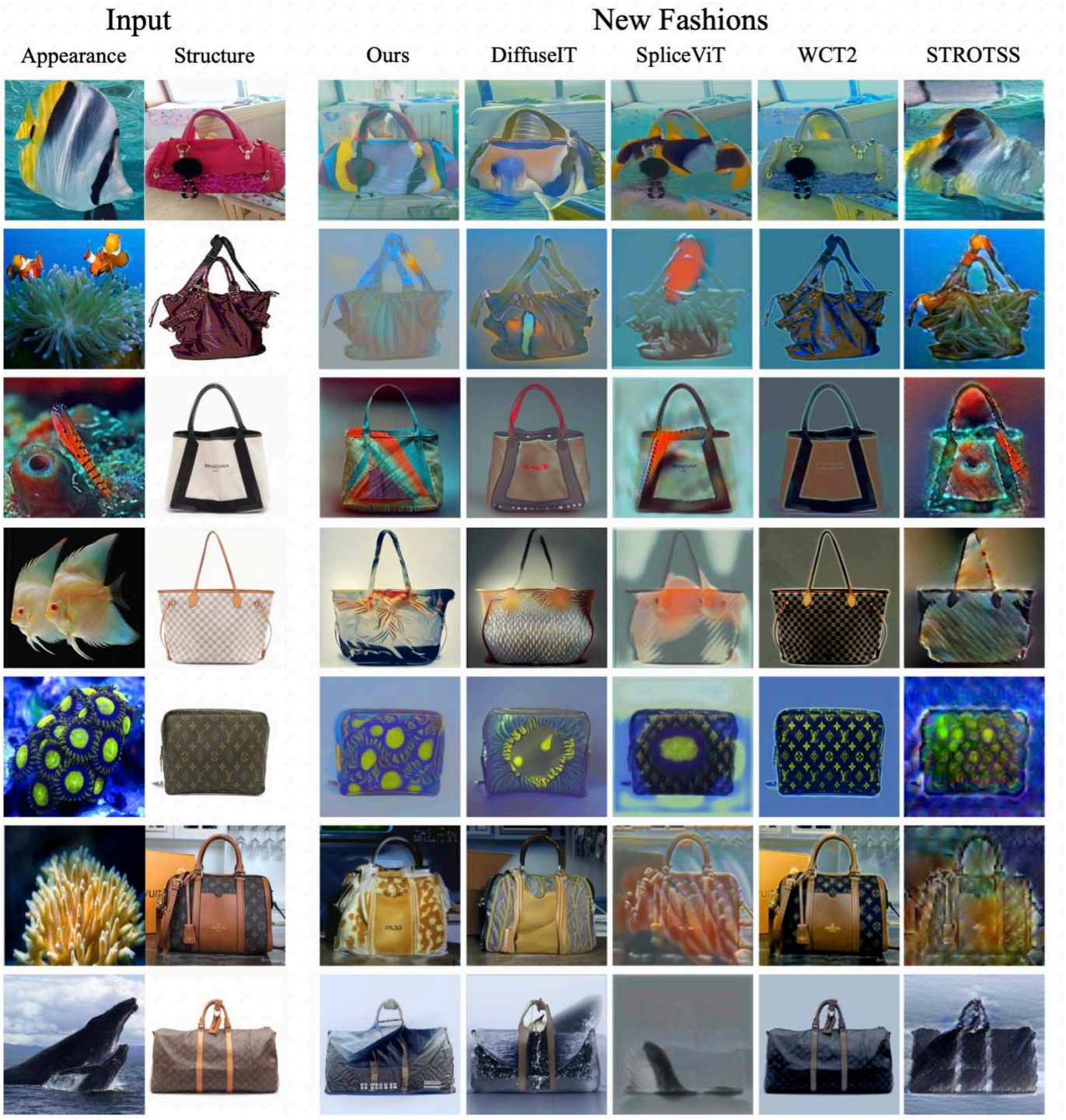

Fig. 4. Comparison with other state-of-the-art (SOTA) methods. Our results show better performance in both appearance and structure similarity.

of the object of each method. The results are shown in Table III. Our model achieves the lowest classification loss. 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 performance after WCT2, but WCT2 only transforms the color for the whole image. Besides, we calculate the color difference histogram (CDH) [43] between the result and appearance image for each method. Our method achieves better appearance similarity than image translation methods. Although NST methods like STROTSS have a better CDH, they tend to transfer the whole image with simple color transformation, as shown in Fig. 4.

E. Ablation Study

In order to verify the effectiveness of the method, we study the individual components of our technical designs through

Fig. 5. Illustration of Mask Generation by Label Condition.

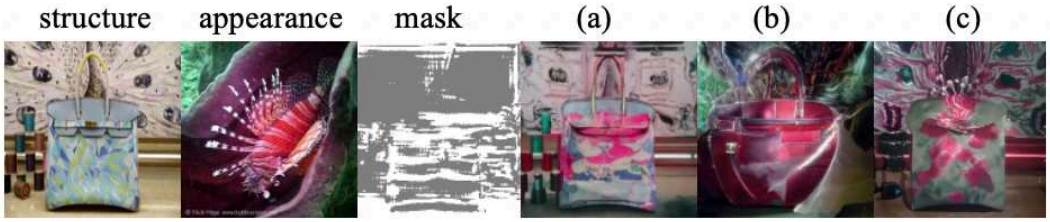

Fig. 6. An example of fashion output with a generated messy mask. (a) and (b) are our results with and without mask guidance, respectively. (c) is the result of DiffuseIT.

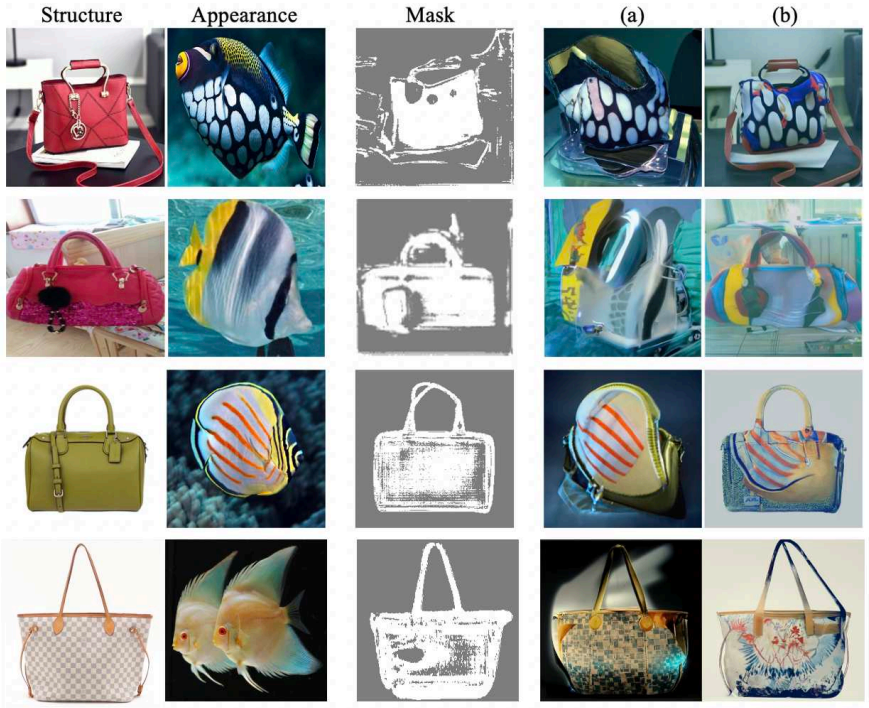

Fig. 7. Examples that show the mask effectiveness. (a) and (b) show the results of our method with or without mask guidance, respectively

several ablation studies as illustrated from Fig. 5 to Fig. 9.

1) Mask Generation: We randomly select several bag images with backgrounds from ImageNet and our dataset. We keep the same experimental setup as Section V-B and show the masks in Fig. 5. For most images, it can generate a foreground object mask that is suitable for our models. Due to the randomness of diffusion, in the last column, we show the scene where the mask is not good enough. But even so, our model still outperforms other models, as shown in Fig. 6.

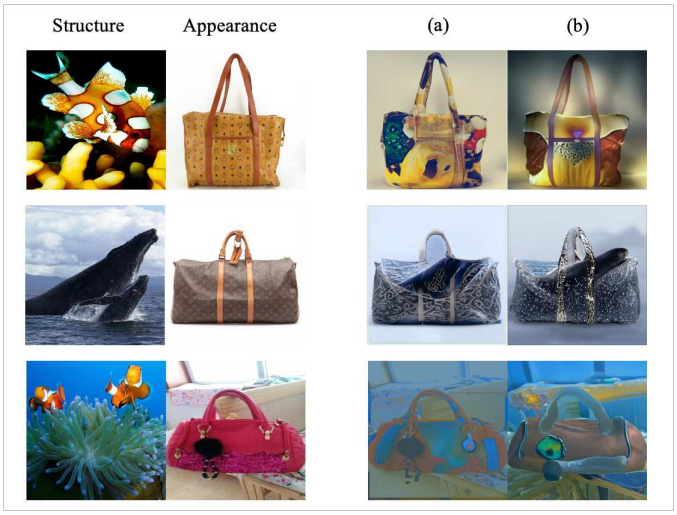

Fig. 8. Comparison with label-conditional DiffuseIT. Our results and results from DiffuseIT with label-conditional diffusion models are shown in (a) and (b), respectively.

Fig. 9. Examples of the DiffuseIT model with the text guidance. "Handbag" to "Handbag with marine life pattern" and "Ocean style Handbag" are prompts for (a) and (b), respectively.

2) Mask Guidance: We conduct an experiment on our model without the mask guidance part, as shown in Fig. 7. Fig. 7(a) shows the result without mask guidance and Fig. 7 (b) presents the outputs of our model with mask guidance. Without mask guidance, in many images, the structure of the bag is destroyed during diffusion. In the last row of the figure, we show that for some images, using a mask may reduce the correlation of appearance, but this is still enough to complete the transfer task. In order to solve a small number of such problems, we set the probability of 0.2 when applying without using mask guidance.

3) Label-Condition: Because our model uses the diffusion model with label-condition, for a fair comparison, we replace the diffusion model of diffuseIT with the same model as ours and use the label "bag" for the condition in the denoising stage. Fig. 8(a) shows the results of DiffuseIT with label condition, and Fig. 8(b) presents our method. Our method still shows better results in structure preservation, appearance similarity, and authenticity. In addition, We show some results of a multi-modal guided diffusion model trained on the same amount of data. Fig. 9 shows the result of DiffuseIT with the text guidance. "Handbag" to "Handbag with marine life pattern" and "Ocean style Handbag" are prompts for (a) and (b), respectively. We can see that a text-guided model cannot complete the task well.

VI. Conclusion and Future Work

We tackle a new problem set in the context of fashion design: designing new clothing fashion from a given clothing image and a natural appearance image, and 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. Due to the randomness of diffusion, the mask cannot guarantee good results every time. In the future, we will try to constrain the diffusion model using the information condition of other modalities to generate better masks.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62106219) and Natural Science Foundation of Zhejiang Province (QY19E050003).

References

[1] A. Ganesan, T. Oates, et al., Fashioning with networks: Neural style transfer to design clothes, arXiv preprint arXiv:1707.09899 (2017).
[2] H. Yan, H. Zhang, J. Shi, J. Ma, X. Xu, Toward intelligent fashion design: A texture and shape disentangled generative adversarial network, ACM Transactions on Multimedia Computing, Communications and Applications (2022).
[3] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, C. Couprie, Design: Design inspiration from generative networks, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.
[4] B.-K. Kim, G. Kim, S.-Y. Lee, Style-controlled synthesis of clothing segments for fashion image manipulation, IEEE Transactions on Multimedia 22 (2) (2019) 298–310.
[5] H. Yan, H. Zhang, L. Liu, D. Zhou, X. Xu, Z. Zhang, S. Yan, Toward intelligent design: An ai-based fashion designer using generative adversarial networks aided by sketch and rendering generators, IEEE Transactions on Multimedia (2022).
[6] D. Zhou, H. Zhang, Q. Li, J. Ma, X. Xu, Coutfitgan: learning to synthesize compatible outfits supervised by silhouette masks and fashion styles, IEEE transactions on multimedia (2022).
[7] C. Yuan, M. Moghaddam, Garment design with generative adversarial networks, arXiv preprint arXiv:2007.10947 (2020).
[8] Y. R. Cui, Q. Liu, C. Y. Gao, Z. Su, Fashiongan: Display your fashion design using conditional generative adversarial nets, in: Computer Graphics Forum, Vol. 37, Wiley Online Library, 2018, pp. 109–119.
[9] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in Neural Information Processing Systems 33 (2020) 6840–6851.
[10] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, arXiv preprint arXiv:2010.02502 (2020).
[11] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, arXiv preprint arXiv:2011.13456 (2020).
[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 (2022).
[13] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al., Photorealistic text-to-image diffusion models with deep language understanding, arXiv preprint arXiv:2205.11487 (2022).
[14] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, D. Cohen-Or, An image is worth one word: Personalizing text-to-image generation using textual inversion, arXiv preprint arXiv:2208.01618 (2022).

[15] G. Kwon, J. C. Ye, Diffusion-based image translation using disentangled style and content representation, in: International Conference on Learning Representations, 2023.

[16] N. Tumanyan, O. Bar-Tal, S. Bagon, T. Dekel, Splicing vit features for semantic appearance transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10748–10757.

[17] S. Amir, Y. Gandelsman, S. Bagon, T. Dekel, Deep vit features as dense visual descriptors, arXiv preprint arXiv:2112.05814 2 (3) (2021) 4.

[18] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, M. Norouzi, Palette: Image-to-image diffusion models, in: ACM SIGGRAPH 2022 Conference Proceedings, 2022, pp. 1–10.

[19] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[20] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, Advances in neural information processing systems 30 (2017).

[21] H. Dong, P. Neekhara, C. Wu, Y. Guo, Unsupervised image-to-image translation with generative adversarial networks, arXiv preprint arXiv:1701.02676 (2017).

[22] X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 172–189.

[23] K. Saito, K. Saenko, M.-Y. Liu, Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder, in: European Conference on Computer Vision, Springer, 2020, pp. 382–398.

[24] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[25] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, M.-H. Yang, Diverse image-to-image translation via disentangled representations, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 35–51.

[26] S. Yang, L. Jiang, Z. Liu, C. C. Loy, Unsupervised image-to-image translation with generative prior, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18332–18341.

[27] L. A. Gatys, A. S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2414–2423.

[28] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1501–1510.

[29] C. Li, M. Wand, Combining markov random fields and convolutional neural networks for image synthesis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2479–2486.

[30] C. Li, M. Wand, Precomputed real-time texture synthesis with markovian generative adversarial networks, in: European conference on computer vision, Springer, 2016, pp. 702–716.

[31] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Advances in Neural Information Processing Systems 34 (2021) 8780–8794.

[32] A. Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR, 2021, pp. 8162–8171.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[34] G. Couairon, J. Verbeek, H. Schwenk, M. Cord, Diffedit: Diffusion-based semantic image editing with mask guidance, in: International Conference on Learning Representations, 2023.

[35] K. Kim, J. C. Ye, Noise2score: tweedie's approach to self-supervised image denoising without clean images, Advances in Neural Information Processing Systems 34 (2021) 864–874.

[36] V. Khrulkov, I. Oseledets, Understanding ddpm latent codes through optimal transport, in: International Conference on Learning Representations, 2023.

[37] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, M. Song, Neural style transfer: A review, IEEE transactions on visualization and computer graphics 26 (11) (2019) 3365–3385.

[38] S. S. Kim, N. Kolkin, J. Salavon, G. Shakhnarovich, Deformable style transfer, in: European Conference on Computer Vision, Springer, 2020, pp. 246–261.

[39] R. Mechrez, I. Talmi, L. Zelnik-Manor, The contextual loss for image transformation with non-aligned data, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 768–783.

[40] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Efros, R. Zhang, Swapping autoencoder for deep image manipulation, Advances in Neural Information Processing Systems 33 (2020) 7198–7211.

[41] J. Yoo, Y. Uh, S. Chun, B. Kang, J.-W. Ha, Photorealistic style transfer via wavelet transforms, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9036–9045.

[42] N. Kolkin, J. Salavon, G. Shakhnarovich, Style transfer by relaxed optimal transport and self-similarity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10051–10060.

[43] G.-H. Liu, J.-Y. Yang, Content-based image retrieval using color difference histogram, Pattern recognition 46 (1) (2013) 188–198.