2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 979-8-3503-0249-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPRW59228.2023.00358

Image Reference-guided Fashion Design with Structure-aware Transfer by Diffusion Models

Shidong Cao[1*] Wenhao Chai[1*] Shengyu Hao[1] Gaoang Wang[1†]

[1] Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University

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 are released at https://github.com/Rem105-210/DiffFashion.

1. Introduction

Image-based fashion design with AI [4, 7, 13, 15, 16, 19] has attracted increasing attention in recent years. One of the emerging tasks in fashion design is to add specific texture elements from non-fashion 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. However, there are two main challenges when applying the commonly used 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 [10, 14] is based on the similarity of the semantically related objects in vision transformer (ViT) [1] features. In this 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. Second, image transfer methods [12] 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.

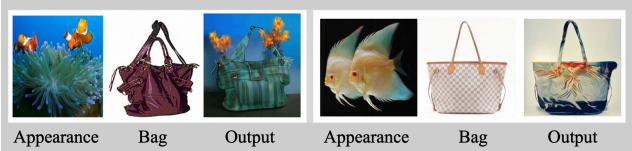

Figure 1. Two examples of a reference-based fashion design task. For a given image pair, *i.e.*, a bag and a referenced appearance image, our method can generate a new image with appearance similarity to the appearance image and structure similarity to the bag image.

To address these issues, we propose an unsupervised structure-aware transfer framework based on diffusion named *DiffFashion*. The proposed framework is based on denoising diffusion probabilistic models (DDPM) [5] 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 both appearance and structure guid-

[*] Equal contribution. [†] Corresponding author.

ance 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 show that our method achieves state-of-the-art in fashion design.

2. Proposed Method

2.1. 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 output images. 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 [11] pre-trained on ImageNet [3] for all the denoising processes in DiffFashion.

2.2. Mask Generation by Label Condition

To decouple the clothing and background, we generate a semantic mask for the input clothing image x_0^S with label conditions. The generated mask is used for preserving the structure information in later steps. Existing methods commonly use additional inputs to obtain the foreground region. Inspired by [2], we propose a mask generation approach that can obtain the foreground clothing area without external information or 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. For these different noise estimates, the noise tends 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, as shown in Fig. 2(a).

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. We use the foreground label to get the true noise estimation. Then we get an averaged noise map by using N different non-foreground label conditions to get an averaged noise map. After that, we calculate the difference between the noise estimations and set a threshold as -0.2 for binarization, which returns an editable semantic mask M for the foreground clothing region.

2.3. Mask-guided Structure Transfer Diffusion

It is difficult to transfer the appearance of the original image to a new fashion clothing image when the significant gap between the two domains is too large [14]. 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.

Inspired by [6], 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.

As shown in Fig. 2(b), we use the latent x_t^A of the reference appearance image to transfer more appearance information to the output fashion. At the first 30% steps in the denoising process, the semantic mask M obtained from the previous step is used to preserve the structure of the clothing image. We get the new prediction x_t^A from the diffusion model. 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.*,

$$\tilde{x}_t^A = M \cdot x_t^A + (1 - M) x_{qt}^S, \tag{1}$$

2.4. ViT Feature Guidance

Following [10, 14], we employ the $[CLS]$ tokens in the last layer of DINO-ViT to guide the appearance as follows:

$$\begin{aligned} &\mathcal{L}_{app}(x_0^A, \hat{x}_t^A) = \\ &||e_{[CLS]}^L(x_0^A) - e_{[CLS]}^L(\hat{x}_t^A)||_2 + \lambda_{MSE}||x_0^A - \hat{x}_t^A||_2, \end{aligned} \tag{2}$$

where $e_{[CLS]}^L$ is the last layer $[CLS]$ token, λ_{MSE} is the coefficient of global statistic loss, and $\hat{x}_t^A$ is the estimated source image predicted from x_t^A by Tweedie's method [8].

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,

$$\begin{aligned} &\mathcal{L}_{struct}(x_0^A, \hat{x}_t^A) = \\ &-\sum_i \log\left(\frac{\text{sim}(k_i^{l,S}, k_i^{l,A})}{\text{sim}(k_i^{l,S}, k_i^{l,A}) + \sum_{j\neq i}\text{sim}(k_i^{l,S}, k_j^{l,A})}\right), \end{aligned} \tag{3}$$

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

By using the loss in Eq. (3), we minimize the loss between keys at the same position of two images while maximizing it of different positions. The weight of appearance loss and structure loss is set to 0.1 and 1 in our experiments.

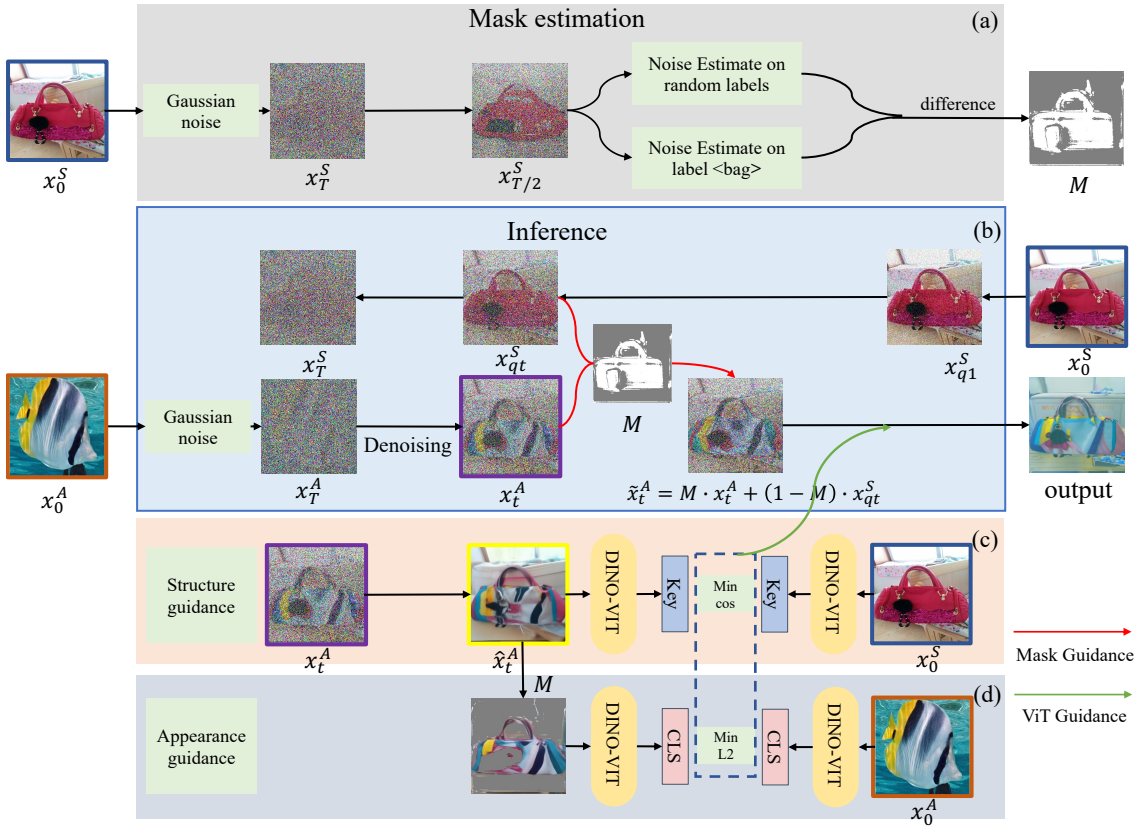

Figure 2. The pipeline of our approach. (a): We add noise to clothing image x_0^S and then use different label conditions to estimate the noise in the denoising process. The semantic mask of the x_0^S can be obtained from the noise difference. (b): We denoise the reference appearance image x_0^A. In 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 $\hat{x}_t^A$ and x_0^S, appearance loss between $\hat{x}_t^A$ and x_0^A, to guide the denoising process. Purple dots and yellow dots represent the denoising process with the same timesteps respectively.

Figure 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.

3. Experiments

In all experiments, we use a diffusion step of $T = 120$. In a single RTX 3090, it takes to about 2 minutes to generate each image. 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. We screened 30 image pairs from *OceanBag* and ImageNet for experiments.

In the guidance part, we set the probability of applying mask guidance to 0.8 and use three different sets of labels to generate masks. To ensure a fair comparison, We run the baseline DiffuseIT [10] three times as ours.

3.1. Experimental Results

We perform both quantitative and qualitative evaluations. We compare our model with SplicingViT [14], DiffuseIT [10] , WCT2 [17] and STROTSS [9]. Fig. 4 shows qualitative results for all methods. We use the LPIPS [18] score between the input clothing images or appearance images and our results to verify the content preservation performance and the appearance similarity. We also apply Mask-RCNN pre-trained on the COCO dataset to detect the mask of the object of each method. The results are shown in Table 1.

We also conduct a user study to obtain subjective evaluations from participants. Specifically, we ask 30 users to score all the output fashion images from all methods for each input pair in realism structure preservation and appearance similarity. The scores range from 0 to 100. We show the averaged subjective evaluation results in Table 2.

In most examples, it can be seen that in terms of fashion design, our method has achieved better performances

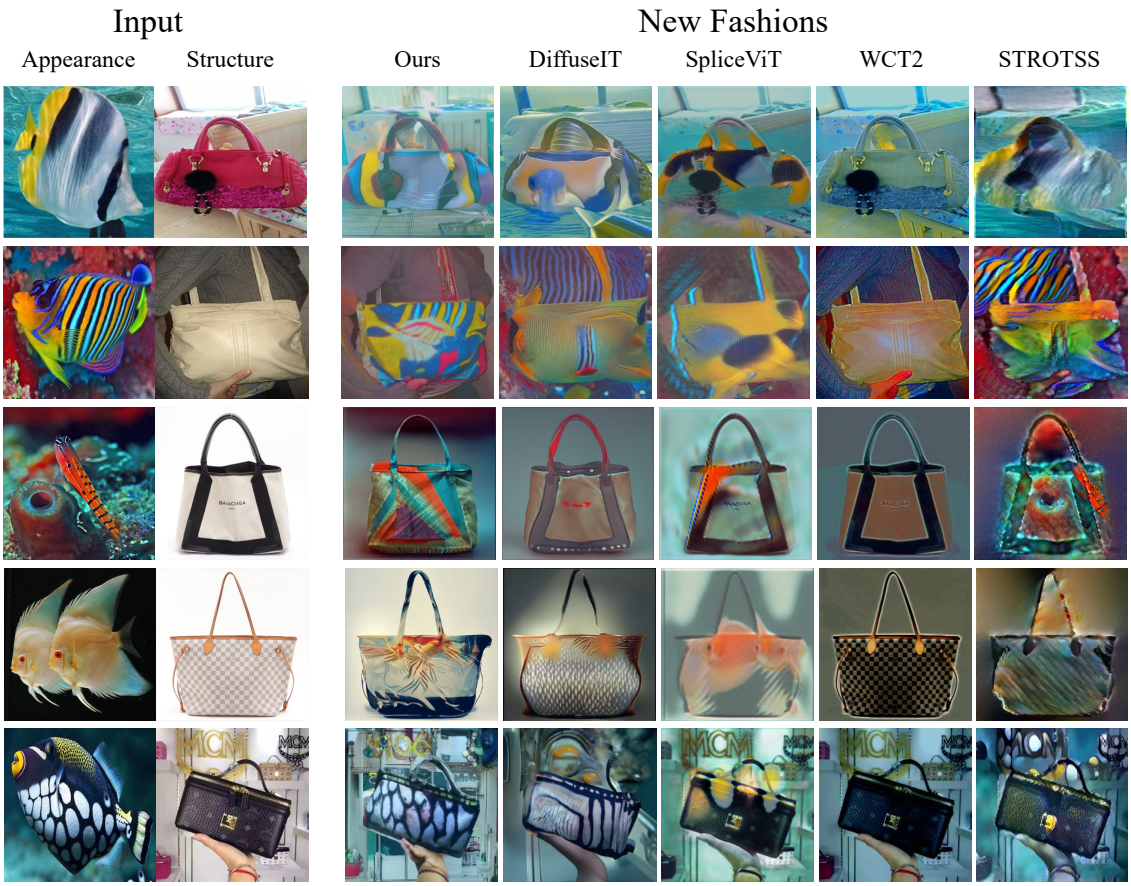

Figure 4. Comparison with other SOTA methods. Our results show better performance in both appearance and structure similarity.

Table 1. Evaluation results based on other models. "L.app", "L.struct", "M.recall" and "M.prec." represent LPIPS with appearance image, LPIPS with clothing image, Mask-RCNN recall, and Mask-RCNN precision, respectively.

Method	L.app↓	L.struct↓	M.recall ↑	M.precision↑
DiffuseIT	0.682	0.622	0.17	0.16
SpliceViT	0.739	0.644	0.03	0.03
WCT2	0.722	**0.435**	**0.53**	**0.51**
STROTSS	**0.600**	0.678	0.06	0.06
Ours	0.666	0.613	0.2	0.17

Table 2. Results of the user study. The output fashion images are evaluated based on the following aspects, ranging from 0 to 100.

Method	Realism	Structure	Appearance
DiffuseIT	75.53 ± 4.68	88.46 ± 5.40	37.27 ± 8.36
SpliceViT	68.44 ± 4.36	80.80 ± 7.82	33.70 ± 8.32
WCT2	$\mathbf{82.89} \pm 5.42$	$\mathbf{95.76} \pm 1.84$	17.69 ± 5.77
STROTSS	63.33 ± 6.43	82.55 ± 6.65	43.11 ± 8.93
Ours	81.15 ± 4.76	91.07 ± 3.82	$\mathbf{52.89} \pm 7.92$

in terms of realism and structure, while completing appearance transfer.

Our model obtains the best score in the overall performance and 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. Our model demonstrates the second-best appearance similarity after STROTSS. But STROTSS transfers the whole image, and its results often suffer from color bleeding artifacts and thus show less authenticity.

4. Conclusion

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 information from the DINO-ViT. The experimental results show that our proposed method outperforms most baselines.

Acknowledgment This work is supported by National Key R&D Program of China under Grant No.2022ZD0162000 and National Natural Science Foundation of China under Grant No.62106219.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 1

[2] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[4] Ashwinkumar Ganesan, Tim Oates, et al. Fashioning with networks: Neural style transfer to design clothes. *arXiv preprint arXiv:1707.09899*, 2017. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[6] Valentin Khrulkov and Ivan Oseledets. Understanding ddpm latent codes through optimal transport. In *International Conference on Learning Representations*, 2023. 2

[7] Bo-Kyeong Kim, Geonmin Kim, and Soo-Young Lee. Style-controlled synthesis of clothing segments for fashion image manipulation. *IEEE Transactions on Multimedia*, 22(2):298–310, 2019. 1

[8] Kwanyoung Kim and Jong Chul Ye. Noise2score: tweedie's approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021. 2

[9] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 3

[10] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. In *International Conference on Learning Representations*, 2023. 1, 2, 3

[11] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[12] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1

[13] Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, and Camille Couprie. Design: Design inspiration from generative networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[14] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 1, 2, 3

[15] Han Yan, Haijun Zhang, Linlin Liu, Dongliang Zhou, Xiaofei Xu, Zhao Zhang, and Shuicheng Yan. Toward intelligent design: An ai-based fashion designer using generative adversarial networks aided by sketch and rendering generators. *IEEE Transactions on Multimedia*, 2022. 1

[16] Han Yan, Haijun Zhang, Jianyang Shi, Jianghong Ma, and Xiaofei Xu. Toward intelligent fashion design: A texture and shape disentangled generative adversarial network. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2022. 1

[17] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 3

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[19] Dongliang Zhou, Haijun Zhang, Qun Li, Jianghong Ma, and Xiaofei Xu. Coutfitgan: learning to synthesize compatible outfits supervised by silhouette masks and fashion styles. *IEEE transactions on multimedia*, 2022. 1