RHanDS: Refining Malformed Hands for Generated Images with Decoupled Structure and Style Guidance

Chengrui Wang[*1], **Pengfei Liu**[*1, 2†], **Min Zhou**[1], **Ming Zeng**[2‡], **Xubin Li**[1], **Tiezheng Ge**[1], **Bo Zheng**[1]

[1]Taobao & Tmall Group of Alibaba
[2]Xiamen University
{wangchengrui.wcr, zhukong.lpf, yunqi.zm}@taobao.com,
zengming@xmu.edu.cn, {lxb204722, tiezheng.gtz, bozheng}@taobao.com

Abstract

Although diffusion models can generate high-quality human images, their applications are limited by the instability in generating hands with correct structures. In this paper, we introduce RHanDS, a conditional diffusion-based framework designed to refine malformed hands by utilizing decoupled structure and style guidance. The hand mesh reconstructed from the malformed hand offers structure guidance for correcting the structure of the hand, while the malformed hand itself provides style guidance for preserving the style of the hand. To alleviate the mutual interference between style and structure guidance, we introduce a two-stage training strategy and build a series of multi-style hand datasets. In the first stage, we use paired hand images for training to ensure stylistic consistency in hand refining. In the second stage, various hand images generated based on human meshes are used for training, enabling the model to gain control over the hand structure. Experimental results demonstrate that RHanDS can effectively refine hand structure while preserving consistency in hand style.

Code — https://github.com/alimama-creative/RHanDS

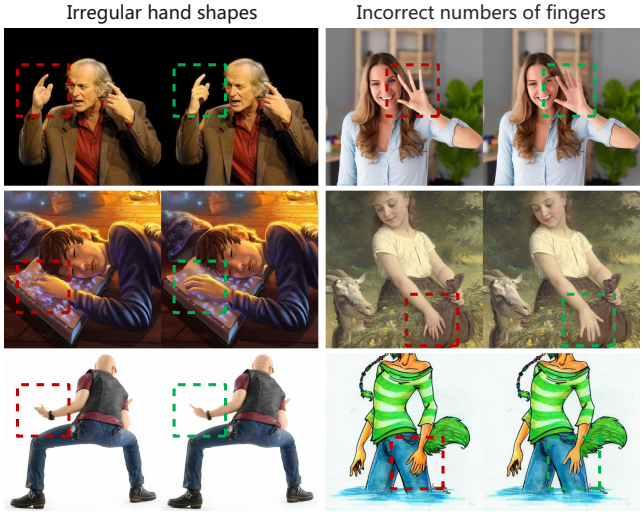

Figure 1: Examples of hands refined by our RHanDS (right in each pair) from the malformed hands (left in each pair).

1 Introduction

Text-to-image diffusion models (Nichol et al. 2022; Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022; Podell et al. 2023) have exhibited the remarkable ability to synthesize visually stunning images based on textual prompts, representing a significant advancement in image generation. Despite their impressive capabilities, current models still face challenges in effectively handling intricate structures like hands (Podell et al. 2023; Rombach et al. 2022). As illustrated in Figure 1, models often produce malformed hands with irregular shapes or incorrect numbers of fingers, deviating from the realistic 3D shape and physical limitations of human hands.

To address the problem, Lu et al. introduces a post-processing framework that employs an inpainting diffusion model to refine the malformed hands. During the refining process, a 3D hand mesh is reconstructed based on the malformed hand image and rendered into a depth image to provide structure guidance for correcting the malformed hand through ControlNet (Zhang, Rao, and Agrawala 2023). However, it is difficult to maintain the hand style as the refining process requires adding noise to the hand image. This issue becomes particularly serious when addressing specific styles like sculptures or cartoons. To enhance style consistency, it seems intuitive to utilize the image region of the original malformed hand. However, extracting style information from the malformed hand inevitably introduces incorrect structural information (Ye et al. 2023a).

In this paper, we propose a novel framework named RHanDS to refine malformed Hands with Decoupled Structure and Style guidance, which is adaptable to malformed hands of arbitrary styles. As shown in Figure 2, RHandDS utilizes diffusion models to repaint the hand region, effectively refining malformed hand structures while preserving the original hand style. In detail, the 3D mesh reconstructed from the deformed hand is rendered into a depth

[*]These authors contributed equally.
[†]Work done during the internship at Alibaba Group.
[‡]Corresponding author.

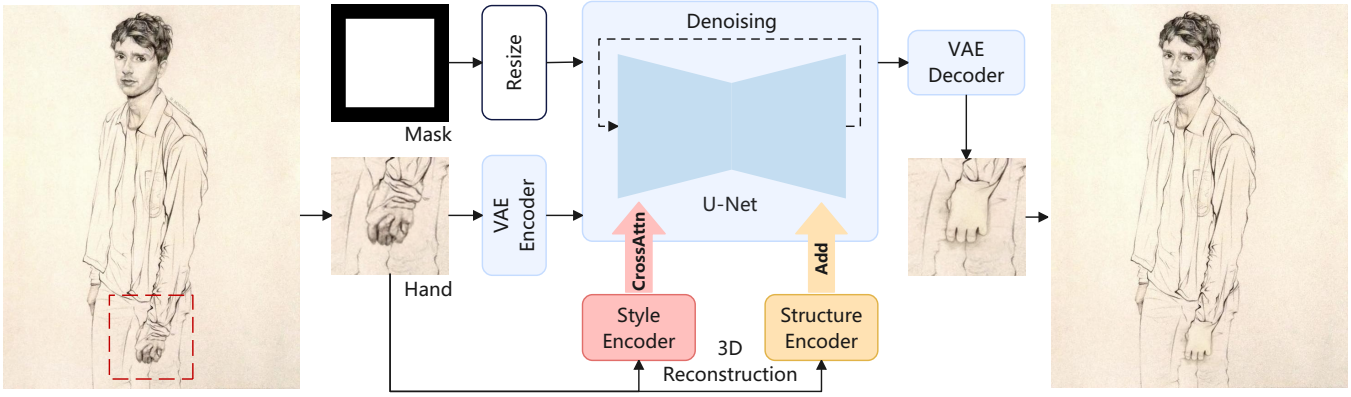

Figure 2: The RHanDS framework we propose contains four modules: a VAE for projecting images into a latent space and reconstructing images from the latent, a conditional U-net for predicting the denoised variant during the denoising process, a style encoder to extract hand style from the malformed hand and map it into the U-net via cross-attention, and a structure encoder to utilize the hand mesh reconstructed from the malformed hand to guide the hand structure. In addition, to achieve a fully automatic process, a hand detection model and a 3D hand reconstruction model are required.

image, which is then input into the structure encoder to provide structure guidance. Meanwhile, the style encoder extracts the style embedding from the malformed hand to offer style guidance. To alleviate the mutual interference between style and structure guidance, we introduce a two-stage training strategy and propose the corresponding multi-style hand datasets.

The first stage focuses on enabling the model to repaint hands in a specified style. To isolate hand styles from structural information, we collect a large number of realistic image pairs depicting the left and right hands of the same individual (i.e., images that share the same style but exhibit different structures). We then use one hand as a style reference to guide the generation of the other hand. In this stage, we only train the style encoder and U-net.

In the second stage, our objective shifts towards making the model generate hand images with additional structural control. Considering the limited style diversity in existing hand image datasets (with 3D meshes, data like these are called hand-mesh pairs for convenience in this paper), we propose a novel approach using the SMPL model and ControlNet to create a more diverse dataset. We initiate this process by generating hand meshes using the SMPL-H model and rendering the corresponding depth images. ControlNet is then utilized to generate hand images, controlled with depth images and prompts indicating various style categories. In this stage, we train the structure encoder while keeping the other parameters frozen.

To verify the ability of hand refining, we use a standard diffusion model to regenerate the hand regions in the HumanArt dataset (Ju et al. 2023), and propose a multi-style malformed hand dataset with mesh-image pairs for evaluation. Both qualitative and quantitative experiments demonstrate the effectiveness and superiority of RHanDS.

In summary, the contributions of this work are as follows:

- To refine malformed hand images, we propose a novel framework named RHandDS, which adopts decoupled structure and style guidance. By leveraging the direct and decoupled style guidance, our model can consistently maintains hand style while refining hand images across various styles.
- We propose a two-stage training strategy and create corresponding datasets, enabling the model to learn control over hand style and structure separately.
- Experiments demonstrate that RHanDS can effectively handle malformed hands across various styles and restore them to better structures.

2 Related Work

2.1 Diffusion Models for Image Synthesis

Recent advancements have been greatly promoted by text-to-image diffusion models. Specifically, the technique of Latent Diffusion Models (LDM) (Rombach et al. 2022) conducts diffusion in a latent image space (Van Den Oord, Vinyals et al. 2017), which largely reduces computational demands. Stable Diffusion (StabilityAI 2022b) is one of the highlighted works of LDM, which utilizes a pre-trained language encoder like CLIP (Radford et al. 2021) to encode the text prompt into latent space to guide the diffusion process.

Diverging from the text-to-image paradigm, image inpainting is tasked with synthesizing visually plausible content within the masked regions of the existing image, ensuring that the synthesized content maintains semantic coherence with the context of the unmasked regions. By fine-tuning the pre-trained text-to-image Stable Diffusion model, the specialized variant known as Stable Diffusion Inpainting model (StabilityAI 2022a) can synthesize content within the masked region that follows textual prompts. However, fine-grained control over the structure and style of the inpainting contents cannot be achieved solely through text guidance.

In order to provide fine-grained conditions during the generation process, different control methods have been explored in previous studies, and it has been demonstrated

that an additional module can be effectively plugged into the existing diffusion models to guide the image generation. ControlNet (Zhang, Rao, and Agrawala 2023) and T2I-Adapter (Mou et al. 2024) use spatial conditions such as canny, depth images, and body poses to control the structure of the generated image. To reduce the fine-tuning cost, Uni-ControlNet (Zhao et al. 2024) allows for the simultaneous utilization of different conditions within one single model. Previous subject-driven approaches such as DreamBooth (Ruiz et al. 2023), Textual Inversion (Gal et al. 2022), LoRA (Hu et al. 2021), and Concept Sliders (Gandikota et al. 2025) achieved stylization and customization by fine-tuning parameters. In contrast, IP-Adapter (Ye et al. 2023a) leverages pre-trained image encoder (Radford et al. 2021) to achieve the image prompt capability for diffusion models, generating images that resemble the reference images in content and style.

2.2 Plausible Hand Generation

Concept Sliders (Gandikota et al. 2025) use parameter-efficient training method (Hu et al. 2021) to learn better physical structure of hand by fine-tuning diffusion models with carefully selected data. Ye et al. synthesize hand-object interaction on existing images using the rough region of the palm and forearm as guidance. However, the inherent complex structure of 3D human hands makes it difficult for these methods to generate correct hands stably. Different from the methods that directly generate hands, some works (Ye et al. 2023b; Lu et al. 2024; Weng, Bravo-Sánchez, and Yeung 2023) introduce the structure information of the target hand into a model to prevent the model from generating malformed hands. HanDiffuser (Ye et al. 2023b) encodes the 3D hand parameters into text embedding for achieving text-to-image generation with realistic hand appearances. Diffusion-HPC (Weng, Bravo-Sánchez, and Yeung 2023) proposes a post-processing method that refines the generated malformed person through a conditional diffusion model with the help of a human depth image rendered based on the reconstructed human mesh. HandRefiner (Lu et al. 2024) proposes a similar post-processing method more suitable for dealing with malformed hands.

Our RHanDS is a post-processing method that leverages 3D hand mesh as the pixel-level condition to control hand structure. Compared to existing methods, RHanDS performs specifically on the hand region rather than the entire image, facilitating more precise refinement. Moreover, RHanDS enhances the perception of hand style, ensuring that the style of the refined hand seamlessly aligns with that of the original image for a coherent and authentic representation.

3 Methodology

In this section, we introduce the detail of our proposed framework RHanDS. Given a generated image containing malformed human hands, we aim to correct the structure of the malformed hand while preserving the style of the generated image. As shown in Figure 2, RHanDS mainly contains four modules: style encoder $\mathcal{E}_{style}$, structure encoder $\mathcal{E}_{struc}$, U-net ϵ_θ, and VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$. The inputs of the framework include a cropped malformed hand image x, a structure reference r_{struc}, a style reference r_{style}, and a mask $mask$. The malformed hand image x is cropped out from the entire image for effective refining. The structure reference r_{struc} is a depth image rendered from a 3D hand model. The 3D hand model can be manually created using 3D tools (ZhUyU1997 2023) or automatically reconstructed from malformed hand, and the structure reference is encoded by the structure encoder $\mathcal{E}_{struc}$ to guide the hand structure. The style reference r_{style} is the malformed hand itself and is encoded by the style encoder $\mathcal{E}_{style}$ to guide the hand style. The binary mask $mask$ indicates the malformed hand region that needs to be repainted. The training process is decomposed into two stages to decouple the structure and style guidance. In the first stage, the U-net and style encoder are trained using Multi-Style Paired Hand Dataset. In the second stage, the structure encoder is trained using the Multi-Style Hand-Mesh Dataset.

3.1 Preliminaries

RHanDS is built on stable diffusion, which efficiently performs the diffusion process in latent space rather than pixel space. Specifically, given an image x_0 in RGB space, the encoder $\mathcal{E}$ first encodes the image into a latent representation $z_0 = \mathcal{E}(x_0)$. During the forward process, normally-distributed noise ϵ_t is added into the latent z_0 to obtain the noisy latent z_t. During the reverse process, the stable diffusion model implements denoising U-net ϵ_θ as the backbone to predict the noise ϵ_t with noisy latent z_t and current timestep t. The simple loss function can be written as

$$\mathcal{L} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t)||_2^2]. \quad (1)$$

In this work, RHanDS employs a U-net with five additional input channels. Specifically, four channels are allocated for the encoded masked-image $\hat{z}_0$, and the remaining single channel is allocated for the resized mask $mask$. During reverse process, the U-net predicts the noise through $\epsilon_t = \epsilon_\theta(z_t, \hat{z}_0, mask, t)$, where $\hat{z}_0$ and $mask$ keep unchanged. After the reverse process, the decoder $\mathcal{D}$ reconstructs the hand image from the latent.

3.2 The First Stage: Style Guidance

As shown in Figure 3, the existing solutions are limited in perceiving hand style and exhibit poor generalization ability in images with anime, oil painting, and other styles. To solve the problem, we propose to encode the hand style into U-net, enabling the model to refine hands x based on style reference r_{style}. Similar to the approach used in IP-Adapter (Ye et al. 2023a), we use CLIP (Radford et al. 2021) image encoder to extract the global image embedding from r_{style}, and then use a linear projection network to transform the image embedding into a sequence of features to obtain the style embedding c_{style}. Overall, the style encoder $\mathcal{E}_{style}$ consists of a CLIP image encoder and a linear projection network. The style embedding c_{style} that can be formulated as follows:

$$c_{style} = \mathcal{E}_{style}(r_{style}) \quad (2)$$

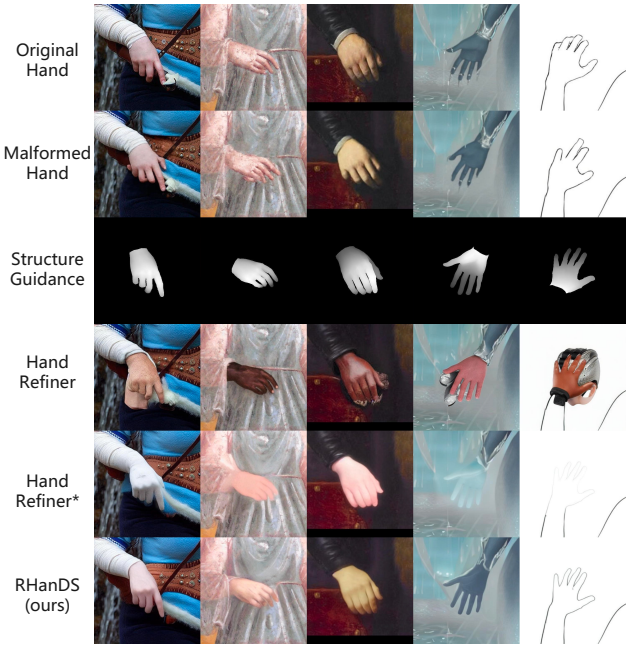

Figure 3: The visual comparison of RHanDS with other methods on different styles of malformed hands. The malformed hands are generated based on the original hands, and the structure guidances are reconstructed from the original hands.

Since the style embedding shares the same dimensions as the original text embedding in U-net, we directly substitute the text embedding with the style embedding. As shown in Figure 4, in this stage, the U-net and the style encoder are trained on multi-style paired hands. We randomly choose a hand from the hand pairs and add noise ϵ to obtain the noisy latent z_t, and use the other hand to supply style embedding c_{style} to prevent the model from learning hand structure from style guidance. The loss function is as follows:

$$\mathcal{L} = \mathbb{E}_{z_t,\epsilon,t}[||\epsilon - \epsilon_\theta(z_t, c_{style}, mask, t)||_2^2], \quad (3)$$

To build the multi-style paired hand dataset that meets the specified requirements, we collect character images of varying styles, and crop images of two hands from the same character, ensuring that each pair exhibits identical styles but different gestures. The details of the dataset can be found in Section 4.1.

3.3 The Second Stage: Structure Guidance

Under style guidance, our framework can repaints hands while preserving their original style. However, as shown in Figure 7, it is hard to correct the hand structure stably without explicit structure guidance due to the inherent complexity of hand structure. To address the problem, we introduce the representation of the correct hand structure by estimating the MANO-based (Romero, Tzionas, and Black 2017a) hand mesh from the malformed hand. For malformed hands that are difficult to reconstruct automatically, manual reconstruction can be performed using existing tools (ZhUyU1997

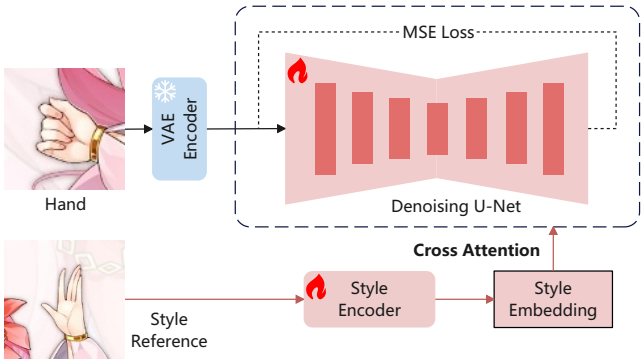

Figure 4: The first stage. In this stage, U-net and style encoder are trained using Multi-Style Paired Hand Dataset for style guidance.

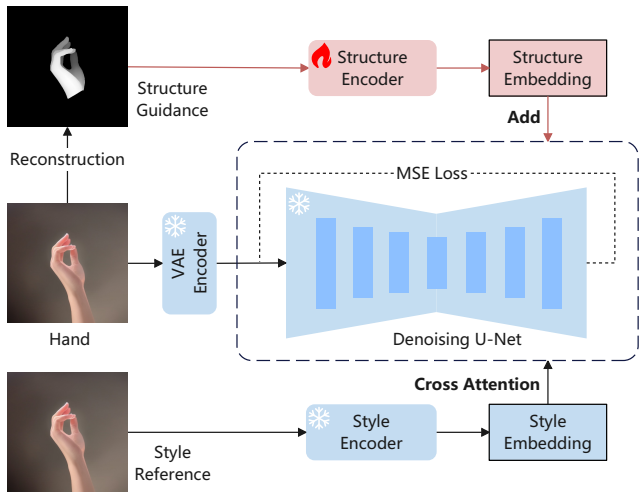

Figure 5: The second stage. In this stage, the structure encoder is trained using Multi-Style Hand-Mesh Dataset for structure guidance.

2023). Following Lu et al., we render the hand mesh into a depth image, serving as the structure reference r_{struc}, to guide the correction of hand structure.

We use the structure encoder $\mathcal{E}_{struc}$ to encode the structure reference r_{struc} into U-net for structure guidance, where $\mathcal{E}_{struc}$ has the same structure as ControlNet (Zhang, Rao, and Agrawala 2023). Moreover, we feed a learnable embedding c_l instead of text embedding into the cross-attention layers of $\mathcal{E}_{struc}$. The learnable embedding shares the same dimension as the text embedding and is integrated into the diffusion model similarly to how the text embedding is. Specially, the input of $\mathcal{E}_{struc}$ is consist of structure reference r_{struc}, noisy latent z_t and learnable embedding c_l. The output of $\mathcal{E}_{struc}$ is the structure embedding c_{struc} that can be formulated as follows:

$$c_{struc} = \mathcal{E}_{struc}(r_{struc}, z_t, c_l) \quad (4)$$

We feed the structure embedding c^i_{struc} into all the outputs out^i of the middle block and decoder blocks of U-net

through a 1×1 convolution layer $\mathbb{Z}^i$ with structure weight w:

$$out^i = out^i + w \cdot \mathbb{Z}^i(c^i_{struc}) \quad (5)$$

As shown in Figure 5, in this stage, we freeze the U-net and the style encoder to maintain the capability of style guidance and train the structure encoder on the multi-style hand-mesh dataset. We add noise ϵ into the hand image to obtain the noisy latent z_t and use the structure encoder to obtain the c_{struc} from the corresponding hand mesh.

$$\mathcal{L} = \mathbb{E}_{z_t,\epsilon,t}[||\epsilon - \epsilon_\theta(z_t, c_{style}, c_{struc}, mask, t)||^2_2]. \quad (6)$$

However, even if we only fine-tune the structure encoder, the performance of style guidance degrades when training with style-limited datasets datasets (Zimmermann and Brox 2017; Chen et al. 2022). Therefore, based on the depth-controlled text-to-image generation pipeline, we use stable diffusion and ControlNet to create a multi-style hand-mesh dataset by using the depth image rendered from random hand meshes. The details of this dataset can be found in Section 4.2.

4 Multi-Style Dataset

We have introduced a hand-refining model that decouples style and structure guidance, along with a two-stage training strategy. In this section, we will introduce a series of datasets purposefully constructed for this method. They encompass various data types and styles of hands. Detailed information about these datasets is provided below.

4.1 Multi-Style Paired Hand Dataset

During the first-stage training, to enable the model to refine hands under style guidance while preventing the leakage of hand structure, we propose a dataset consisting of hand pairs in various styles, each with the same style but different gestures. We find these pairs by detecting the two hands of the same person. Specifically, we employ the YOLOv8 (ultralytics 2022) to detect human hands and use mmpose (Contributors 2020) to detect human pose, which allows us to filter out hands belonging to the same person. This dataset includes 517,096 hand pairs from natural human images with various textures, and 2,423 hand pairs from anime images.

4.2 Multi-Style Hand-Mesh Dataset

During the second-stage training, to prevent the model from overfitting to the generation of hands in a single style, we first utilize the SMPL model with MANO (SMPL-H) (Loper et al. 2015; Romero, Tzionas, and Black 2017b) to create hand meshes and get corresponding depth images. Specifically, we randomly generated the body pose parameters of the SMPL model and hand pose parameters of the MANO model, then combined them to form a complete human model. We cropped out the arm and hand submesh and rendered them into depth images. Then, we employ the depth-conditioned ControlNet to generate multi-style hand images with text prompts indicating different style categories. We reconstruct the hand mesh from the generated hand images and calculate the metric MPJPE$(K, K') = \mathbb{E}||K - K'||$ to filter out poorly structured hands, where K, K' are the 2D projection coordinates of the hand joints, corresponding to the MANO hand mesh and the reconstructed hand mesh. Ultimately, this dataset includes 7 different style categories, with 8,000 hand-mesh pairs for each style.

4.3 Multi-Style Malformed Hand-Mesh Dataset

To evaluate the model's ability to handle multi-style hands, we extract hand images from the HumanArt (Ju et al. 2023) dataset and use SDEdit (Meng et al. 2022) to regenerate the hand regions as corresponding malformed hand images. For evaluation, the style reference is the malformed hand itself, while the structure reference is reconstructed from the original hands using (Chen et al. 2022). After manually filtering out the data with reconstruction failures, we collect 1440 image pairs in total, consisting of 1 natural style and 13 artificial styles.

5 Experiments

5.1 Implementation Details

In order to utilize the generation capability of existing model, the VAE and U-net are initialized from Stable Diffusion Inpainting v1.5, and the structure encoder is initialized from pretrained depth ControlNet. CLIP ViT-H/14 (Ilharco et al. 2021) is adopted as the style encoder, and the linear projection network is randomly initialized.

For the first stage, we use our Multi-Style Paired Hand Dataset to train U-net and style encoder for 15k iterations with a learning rate of 1e-5 and with a total batch size of 256 on 8 NVIDIA-A100 GPUs. The default token length for the style encoder is set to 8. The hand images x are resized to 512×512, and the style references r_{style} are resized to 224×224 with random horizontal and vertical flipping.

For the second stage, we use the combination of Static Gesture Dataset (SGD) (SynthesisAI 2022) and our Multi-Style Hand-Mesh Dataset to train structure encoder for 15k iterations with a learning rate of 2e-5 and with a total batch size of 256 on 8 NVIDIA-A100 GPUs. The weight w for the structure encoder is set to 1 during training. The structure references r_{struc} are resized to 512×512 and the depth values of the hand are normalized into [0.2, 1.0].

Additionally, the mask $mask$ we use to indicate the hand region covers the middle 9/16 of the entire image. In order to learn style guidance from the style encoder rather than the remaining skins in the image, we apply a strategy to randomly expand the region indicated by the mask to force the model to repaint all the skin in the image. Meanwhile, we randomly replace the style embedding with a learnable style embedding with a probability of 0.1 and set the mask to the entire image with a probability of 0.5. We adopt noise offset (Guttenberg 2023) and min-SNR strategy (Hang et al. 2023) for training.

During inference, we add the maximum noise (strength = 1.0) to image x and adopt DDPM (Ho, Jain, and Abbeel 2020) sampler for denoising. We use structure weight w = 0.6, denoising step $steps$ = 25 as default. The refining process takes approximately 3 seconds and requires 4GB GPU memory.

Method	FID ↓	MPJPE ↓	Conf. ↑
Text2Image Dataset			
Stable Diffusion†	77.60	-	0.93
HandRefiner†	74.12	-	0.94
RHanDS†	**73.63**	-	0.94
Image2Image Dataset			
HandRefiner†	13.97	7.87	0.97
RHanDS†	**13.54**	**6.89**	0.97
HandRefiner	33.84	12.02	0.94
RHanDS	**22.18**	**9.86**	**0.96**

Table 1: Quantitative comparison of the proposed RHanDS with other methods on datasets (Lu et al. 2024). The method with † represents calculating metrics on the entire image, while the method without † represents calculating metrics only on the hand region.

5.2 Evaluation Metrics

We use several metrics to evaluate the performances: (1) Fréchet Inception Distance (Heusel et al. 2017) (FID) focuses on the overall distribution statistics of the refined images and the Ground Truth, (2) Style loss(Gatys, Ecker, and Bethge 2016) evaluates the style consistency between the refined images and the Ground Truth (the loss we reported is multiplied by 100), (3) MPJPE (Ionescu et al. 2013) between reconstructed 2D hand poses measures the structure consistency, (4) Keypoint detection confidence scores (Conf.) of a hand detector (Lugaresi et al. 2019; Zhang et al. 2020) is used to indicate the plausibility of generated hands. The FID and style loss are calculated between the refined hands and the original hands. The MPJPE is calculated between the reconstructed hand mesh and the structure reference, where the hand mesh is reconstructed from the refined hand using MeshGraphormer (Lin, Wang, and Liu 2021) for test datasets proposed by HandRefiner and using MobRecon (Chen et al. 2022) for our multi-style multi-style malformed hand-mesh test dataset. By default, MPJPE is calculated based on the resolution of 512×512.

5.3 Results and Comparison

Comparison between different methods on HandRefiner Dataset. We evaluate our method on two datasets proposed by HandRefiner. The Text2Image Dataset includes 12K images generated with the text descriptions from HAGRID (Kapitanov et al. 2024), and the Image2Image Dataset includes 2K images sampled from HAGRID. Following the setting in HandRefiner, for each image, we extract the hand region and refine the hand, then paste it back to the original image and calculate metrics on the entire image. Meanwhile, we found that calculating metrics on the hand region extracted using a fixed strategy can eliminate the influence of hand proportion and image size on the evaluation metrics. Therefore, we also report the metrics calculated on the hand region. As shown in Table 1, our RHanDS outperform HandRefiner on both datasets.

Style-structure cross reference. To verify the robustness and generalization of RHanDS, we conducted experiments with various styles across different structure references.

Figure 6: Visualization of hands generated under the guidance of different style and structure references.

During inference, we mask the entire image to achieve the more intuitive result in Figure 6. Since the style and structure guidance are decoupled during training, we can generate hands with specified structures under any style reference without worrying about structure leakage.

Comparison between different methods on Multi-Style Malformed Hand-Mesh Dataset. We compare with HandRefiner on our multi-style malformed hand-mesh dataset to evaluate the generalization ability to various styles. For fair comparison, we also reimplement and train a variant version of handrefiner on our multi-style dataset, which is denoted as HandRefiner*. As shown in Figure 3, the hands generated by HandRefiner are monotonous in terms of style. After training with our multi-style dataset, HandRefiner* can generate hands of various styles while still struggling to generate hands with the consistent color of the original malformed hands. Through decoupled style and structure guidance, our RHanDS refines hand structure while maintaining consistency with the original malformed hand style. Furthermore, the quantitative experiments in Table 2 show that our RHanDS outperforms HandRefiner and HandRefiner* in all the evaluation metrics in terms of style and structure.

5.4 Ablation Study

In this subsection, we will validate the effectiveness of the two-stage training strategy and dataset construction through quantitative experiments. Additionally, we will analyze the roles of the style and structure modules within the network.

Effect of two-stage training. Without the first-stage training, we train the entire model on the multi-style hand-mesh dataset with a single hand as both hand image x and style reference r_{ref}. As shown in Table 2, this model can achieve performance comparable to the two-stage training model in terms of MPJPE and Conf., but the FID has in-

Method	Style Encoder	Structure Encoder	MPJPE ↓	Conf. ↑	FID ↓	Style Loss ↓
Malformed Dataset	-	-	40.29	0.86	23.84	2.77
HandRefiner	-	-	33.27	0.88	38.52	10.13
HandRefiner*	-	-	29.60	0.89	34.12	5.61
RHanDS	stage-2	stage-2	26.29	0.89	47.76	7.11
RHanDS	stage-1*	stage-2	29.48	0.88	33.68	5.35
RHanDS	stage-1	stage-2*	**26.17**	**0.91**	33.56	5.23
RHanDS	stage-1	stage-2	27.23	0.89	**32.59**	**5.04**

Table 2: Quantitative comparison of the proposed RHanDS with other methods on the Multi-Style Malformed Hand-Mesh Dataset. Stage-1* denotes using the hand image itself as a style reference for the first stage training, and stage-2* denotes using only the SGD dataset for the second stage training.

Figure 7: The visual comparison of different modules.

creased by 15.17, and the style loss has increased by 2.07. Additional experiment shows that compared with the the model without structure encoder (FID 33.33, Style Loss 5.50), the FID is increased by 14.43 and the style loss is increased by 1.61. These demonstrate that the first-stage training is beneficial for helping the model perceive hand style from style reference.

Effect of Paired Hand Dataset. To clarify the impact of hand structure information leaked from the style encoder on hand refinement, we use the unpaired hands for the first stage training, where the style reference r_{ref} uses the same image as the hand image x. As shown in Table 2, the model trained with paired hands achieves better performance on all metrics, especially on the metrics related to hand structure.

Effect of Multi-Style Hand-Mesh Dataset. To clarify the role of multi-style hand-mesh data in two-stage training, we conduct an experiment using only the SGD dataset for the second-stage training. As shown in Table 2, compared with using both SGD and the multi-style hand-mesh dataset, the model trained with the SGD dataset outperformed in terms of MPJPE and Conf. by 1.06 and 0.02, respectively. The reason for better MPJPE and Conf. is that, unlike the hand images generated using stable diffusion in the multi-style hand-mesh dataset, the hand images rendered from mesh in SGD are more stable in hand structure. However, due to the low style diversity of hand data in SGD, the FID increased by 0.97, and the style loss increased by 0.19.

Effects of different modules. In RHanDS, we use style encoder and structure encoder to guide the style and structure of the hands. To illustrate the effects of these two modules, we separately shield the structure encoder and the style encoder during inference. Specifically, we set the structure weight w to 0 to shield the structure encoder and use the learnable style embedding rather than extracting embedding from the style reference to shield the style encoder. Qualitative analysis in Figure 7 shows that the structure of the hand tends to be malformed without the structure encoder, and the style of the hand is easily changed without the style encoder.

w	MPJPE ↓	Conf. ↑	FID ↓	Style Loss ↓
0.0	67.19	0.833	33.33	5.50
0.2	48.32	0.867	31.50	4.96
0.4	31.50	0.895	**31.20**	**4.71**
0.6	**27.23**	**0.896**	32.59	5.04
0.8	27.29	0.891	34.46	5.62
1.0	29.73	0.881	37.46	6.22

Table 3: Ablation study on the impact of structure weight w.

Structure weights. The structure weight w is adjustable during inference. As shown in Table 3, the optimal value of structure weight is almost in the range of 0.4 to 0.6. Specifically, we choose the setting of w = 0.6 as optimal. This setting achieves the best MPJPE and Conf., which is important in controlling hand structure.

6 Conclusion

In this paper, we introduce a novel hand refining method RHanDS. It decouples the hand refining process into style and structure guidance and employs a two-stage training strategy to fully leverage existing data, improve the structure of repainted hands, and preserve the style of the original image. Furthermore, three datasets designed specifically for this task are built to facilitate research on multi-style hand refining. Qualitative and quantitative experiments demonstrate our method's superiority in adapting to images across various styles compared with existing methods.

Acknowledgements

This work is partially supported by National Natural Science Foundation (Grant No. 62072382), Yango Charitable Foundation, and Alibaba Group through Alibaba Research Intern Program.

References

Chen, X.; Liu, Y.; Dong, Y.; Zhang, X.; Ma, C.; Xiong, Y.; Zhang, Y.; and Guo, X. 2022. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20544–20554.

Contributors, M. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Gandikota, R.; Materzyńska, J.; Zhou, T.; Torralba, A.; and Bau, D. 2025. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, 172–188. Springer.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Guttenberg, N. 2023. Diffusion with offset noise.

Hang, T.; Gu, S.; Li, C.; Bao, J.; Chen, D.; Hu, H.; Geng, X.; and Guo, B. 2023. Efficient Diffusion Training via Min-SNR Weighting Strategy. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 7407–7417. IEEE.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP. https://doi.org/10.5281/zenodo.5143773. Accessed: YYYY-mm-dd.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Ju, X.; Zeng, A.; Wang, J.; Xu, Q.; and Zhang, L. 2023. Human-Art: A Versatile Human-Centric Dataset Bridging Natural and Artificial Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kapitanov, A.; Kvanchiani, K.; Nagaev, A.; Kraynov, R.; and Makhliarchuk, A. 2024. HaGRID–HAnd Gesture Recognition Image Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4572–4581.

Lin, K.; Wang, L.; and Liu, Z. 2021. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12939–12948.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.

Lu, W.; Xu, Y.; Zhang, J.; Wang, C.; and Tao, D. 2024. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7085–7093.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. arXiv:1906.08172.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 4296–4304. AAAI Press.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent

diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Romero, J.; Tzionas, D.; and Black, M. J. 2017a. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6): 245:1–245:17.

Romero, J.; Tzionas, D.; and Black, M. J. 2017b. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6).

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

StabilityAI. 2022a. Stable Diffusion Inpainting v1.5. https://huggingface.co/runwayml/stable-diffusion-inpainting/. Accessed: 2022.

StabilityAI. 2022b. Stable Diffusion v1.5. https://huggingface.co/runwayml/stable-diffusion-v1-5/. Accessed: 2022.

SynthesisAI. 2022. Static gestures dataset. https://synthesis.ai/static-gestures-dataset/. Accessed: 2022.

ultralytics. 2022. YOLOv8. https://github.com/ultralytics/ultralytics.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Weng, Z.; Bravo-Sánchez, L.; and Yeung, S. 2023. Diffusion-hpc: Generating synthetic images with realistic humans. arXiv:2303.09541.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023a. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv:2308.06721.

Ye, Y.; Li, X.; Gupta, A.; De Mello, S.; Birchfield, S.; Song, J.; Tulsiani, S.; and Liu, S. 2023b. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22479–22489.

Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.-L.; and Grundmann, M. 2020. Mediapipe hands: On-device real-time hand tracking. arXiv:2006.10214.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2024. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36.

ZhUyU1997. 2023. open-pose-editor. https://github.com/ZhUyU1997/open-pose-editor.

Zimmermann, C.; and Brox, T. 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 4913–4921. IEEE Computer Society.