HandCraft: Anatomically Correct Restoration of Malformed Hands in Diffusion Generated Images

Anonymous WACV Algorithms Track submission

Paper ID 774

Abstract

Generative text-to-image models, such as Stable Diffusion, have demonstrated a remarkable ability to generate diverse, high-quality images. However, they are surprisingly inept when it comes to rendering human hands, which are often anatomically incorrect or reside in the "uncanny valley". This paper proposes a method—HandCraft—for restoring such malformed hands. This is achieved by automatically constructing masks and depth images for hands as conditioning signals using a parametric model, allowing a diffusion-based image editor to fix the hand's anatomy and adjust its pose while seamlessly integrating the changes into the original image, preserving pose, color, and style. Our plug-and-play hand restoration solution is compatible with existing diffusion models, and the restoration process facilitates adoption by eschewing any fine-tuning or training requirements. We also contribute MalHand datasets that contain generated images with a wide variety of malformed hands in several styles for training and benchmarking, and demonstrate through qualitative and quantitative evaluation that HandCraft not only restores anatomical correctness but also maintains the integrity of the overall image.

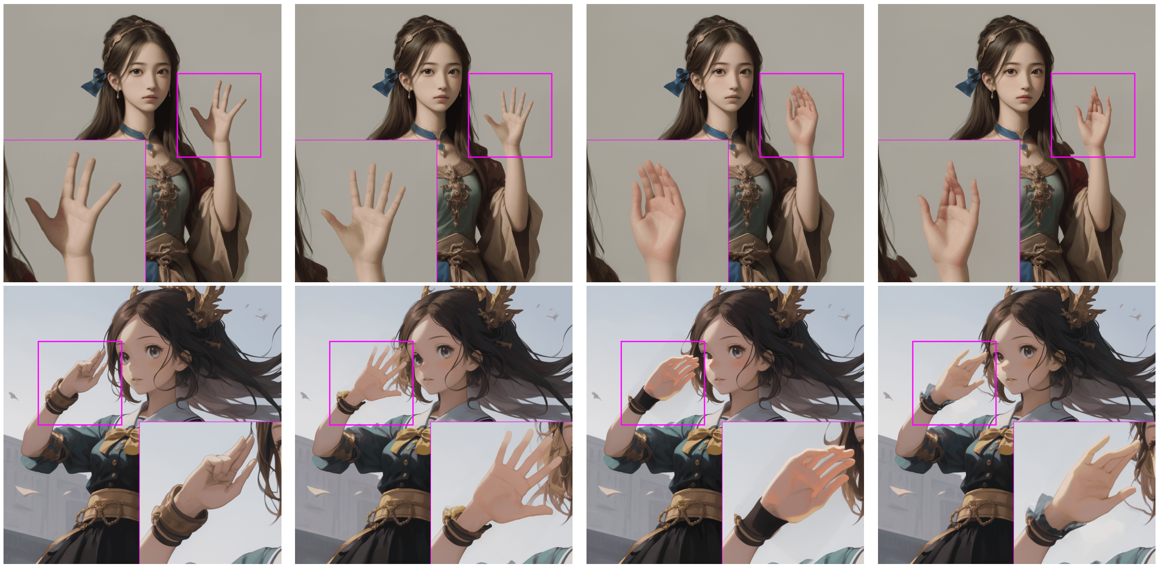
1. Introduction

Text-to-image diffusion models, such as Stable Diffusion [25], have gained wide popularity due to their remarkable capability to generate diverse, high-quality images across a wide range of styles [23, 27]. However, they struggle to accurately render human hands, often producing anatomically incorrect or highly unusual forms [22]. These errors can include hands with supernumerary or missing digits, atypical relative finger lengths, and other distortions. Fig. 1 illustrates two cases of such malformed hands, with a missing finger in the top row and abnormal relative finger lengths in the bottom row. These examples highlight the discrepancy between the generated depictions and human anatomy.

Due to humans' high sensitivity to deviations from the expected human form, generating malformed hands often leads to an "uncanny valley" [21] effect, which affects the realism of these images. This in turn hinders the use of these models as artistic tools. We note here that we do not use the term "malformed" in the pejorative sense, since we recognize that a wide variety of hand shapes are naturally present in the human population or may arise from misadventure. That is, the model is inadvertently forming the hands atypically, rather than intentionally depicting the difference that exists in the human population.

Diffusion models' propensity for generating malformed hands has been widely recognized [3, 19, 22]. There has been growing interest for techniques to repair these malformed hands, reflected by a large number of tutorials across various languages for this purpose [1, 2, 7, 15]. However, the restoration methods proposed in these tutorials often necessitate human intervention. For instance, repeatedly inpainting the manually-annotated affected areas until a satisfactory outcome is achieved [12]. The requirement for human involvement makes the correction process laborious. Prompt engineering has also emerged as a popular strategy to mitigate the issue of malformed hands in images generated by diffusion models [17, 26]. By meticulously designing and refining text prompts, users attempt to guide the model towards generating more anatomically accurate hands [5]. Despite these efforts, even well-crafted prompts often fail to prevent the occurrence of malformed hands [4].

We introduce an end-to-end framework designed to repair malformed hands in generated images while minimizing the need for human intervention. To achieve this, we propose an approach for generating a hand shape as a conditioning image to guide ControlNet [29], a diffusion-based image editing method, in correcting malformed hands. Our method is capable of responsively adjusting the size and angle of the hand shape, ensuring that the restored hand seamlessly integrates with the original human figure, while preserving the surrounding regions of the image unaltered. Experimental results demonstrate the robustness of our approach. Furthermore, our restoration process is designed to



(a) Malformed Hand (b) Spread Palm (c) Curved Gesture (d) Relaxed Pose

Figure 1. Images generated by Stable Diffusion [25] often exhibit anatomically incorrect hands (a), for example, a missing finger (top) or abnormal relative finger lengths (bottom). Our method—HandCraft—is able to correct the hands in a controllable manner, allowing for a variety of output gestures while following the style of the original image (b–d). The resulting images feature naturally-posed hands, improving the quality of the AI-generated portraits and restoring the illusion of reality.

be plug-and-play, requiring no further fine-tuning or training, and is therefore easy to integrate into various diffusion models. Our contributions are

1. HandCraft, a framework for detecting and restoring malformed hands generated by diffusion models while minimizing alterations to other image regions;
2. a simple yet robust control image generation method to construct a mask and an aligned depth image for the hand region as condition signals, enabling a diffusion-based image editor to restore malformed hands; and
3. the MalHand datasets, comprising portraits with malformed hands across diverse styles, that can be used to train a malformed hand detector and thoroughly evaluate baseline models.

HandCraft achieves state-of-the-art performance on both the MalHand-realistic and MalHand-artistic datasets.

2. Related Work

In this section, we provide a brief overview of image synthesis and editing techniques before discussing approaches for restoring malformed hands in generated images.

Image Synthesis. After earlier successes with Variational Autoencoders (VAEs) [14] and Generative Adversarial Networks (GANs) [8], diffusion models [9] have emerged as a powerful new class of generative models. They are characterized by their ability to map noise into complex images through a gradual denoising process. This technique was refined by the development of Latent Diffusion Models (LDMs) [25], which tackle the computational challenges by operating in a latent space, significantly improving both efficiency and the quality of generated images. By leveraging pretrained autoencoders, LDMs offer a versatile and flexible architecture that supports a wide range of conditioning inputs, such as text descriptions. This advance enables efficient and adaptable image synthesis models like Stable Diffusion [25]. However, despite the impressive capabilities of these models, generated images of humans often exhibit malformed hands. The complex structure and fine details of hands pose a challenge for these models, often resulting in anatomically incorrect hand representations [22].

Image Editing has emerged as an application of generative models, enabling users to modify existing images according to their preferences. Early deep learning-based image editing methods employed encoder-decoder architectures, where the input image is encoded into a latent representation, manipulated, and then decoded to produce the edited output [11, 28]. More recent techniques have explored the use of GANs [8] for image editing [10, 30]. ControlNet [29] is a recent work that leverages diffusion models for image editing by incorporating spatially-localized conditioning controls. The primary objective of ControlNet is to provide users with a means of introducing conditions, such as Canny edges and human poses, to guide the generation and editing of images from pretrained diffusion models.

Malformed Hand Restoration. Contemporary work has also tackled the issue of correcting malformed hands in images generated by Stable Diffusion. HandDiffuser [22] focuses on generating humans with non-malformed hands from text, instead of restoring existing images. HandRefiner [19], in contrast, has a more similar objective to our work, albeit with a diverging approach. It modifies the entire image to rectify the malformed hands, affecting the overall composition and potentially altering unintended aspects of the image. In contrast, our research focuses specifically on the malformed hand area, aiming to correct these imperfections with minimal impact on the rest of the image. This targeted approach allows for precise corrections that maintain the integrity and originality of the generated artwork, distinguishing our work from existing solutions.

3. Detecting and Restoring Malformed Hands

In this section, we detail the proposed HandCraft framework. We begin by establishing the notation and providing an overview of the framework's pipeline. Subsequently, we delve into the generation of a conditional hand shape that ensures anatomical plausibility and accurate positioning. This is followed by a discussion on defining the restoration region within the image for localized correction while preserving the overall image integrity.

3.1. The HandCraft Framework

Our HandCraft framework is designed to address restoring malformed hands in images generated by diffusion models, which is illustrated in Fig. 2. This framework consists of three stages: malformed hand detection, control image generation, and hand restoration.

At the malformed hand detection stage, a hand detector is applied to the input image I to identify the region of interest, producing a bounding box mask M_d for the malformed hand. Any pretrained hand detection model, such as YOLOv8 [18], can be used as the hand detector, but both standard and malformed hands will be detected. By fine-tuning the malformed hand detector on stylized images with standard and malformed hands, our HandCraft framework can accommodate diverse image styles, such as anime, and avoid modifying the images when the generated hand is not malformed. In addition to the malformed hand detector, a body pose estimator (Mediapipe [20]) is also used to predict the body pose S, which facilitates the correct positioning and orientation of the hand template T. This is used instead of a hand pose estimator, since the latter regularly fails when applied to malformed hands.

At the control image generation stage, the primary objective is to create a control image I_c and a corresponding control mask M that will guide the hand restoration process. To this end, a control image generator aligns a pre-defined hand template T, which is a depth map of a hand, using the extracted body pose S to create the control image I_c, as well as a template mask M_t. The control mask M is obtained by the union of the bounding box mask M_d extracted from the original image and the hand template mask M_t, to precisely localize the hand. The depth image I_c and the mask M for hand are crucial conditioning signals for the editing process to achieve a seamless integration of the restored hand.

The final stage is hand restoration. A pretrained ControlNet [29] model with frozen weights is provided with the control image I_c and mask M to adjust the input image I, given text prompt P that describes the shape of the hand template T. This restoration process focuses only on the hand region, while preserving the integrity of the rest of the image. The output of this stage is the restored image I', where the malformed hand has been restored to match the desired shape and pose specified by the hand template and text prompt. The restored hand blends with the original image's style and aesthetics, resulting in a more realistic and anatomically correct representation.

Our framework's versatility is evidenced by its successful application to various instances of Stable Diffusion models, demonstrating its efficacy across diverse image styles.

3.2. Control Image Generation

The two main challenges when generating the conditioning signals for hand restoration are (1) ensuring the hand template T is anatomically plausible for the body pose; and (2) accurately positioning T in the input image to make a seamless generation, inclusive of its rotation and handedness (left or right hand). Our detailed solutions to address these two challenges are provided below.

Ensuring Anatomical Plausibility. To guarantee the anatomical plausibility of T, we utilize a predefined library [12] to randomly select an appropriate hand template. While relying on predefined templates might constrain diversity, it significantly enhances the restoration's anatomical accuracy—a critical factor since methods like mesh fitting (e.g., using MeshFormer [16]) for severely deformed hands in M_d can lead to unnatural hand shapes that deviate from typical human anatomy. Such deviations are evident when observing inputs with malformed hands in which fingers may appear unnaturally bent or fused, as shown in Fig. 3. In contrast, the template-based approach of aligning T within the region defined by M_d and subsequently adjusting within the union mask $M = M_d \cup M_t$ ensures a more faithful restoration, demonstrating the method's efficacy in maintaining anatomical fidelity.

In addition to the default random selection method, we also propose a silhouette-based method to select the hand template. Although the goal of HandCraft is to ensure anatomical correctness of hand renders and seamless integration with the original image, not consistency with the original (corrupt) hand render, we also provide an option to encourage consistency between the malformed and restored hand renders. We do so by generating multiple renders with

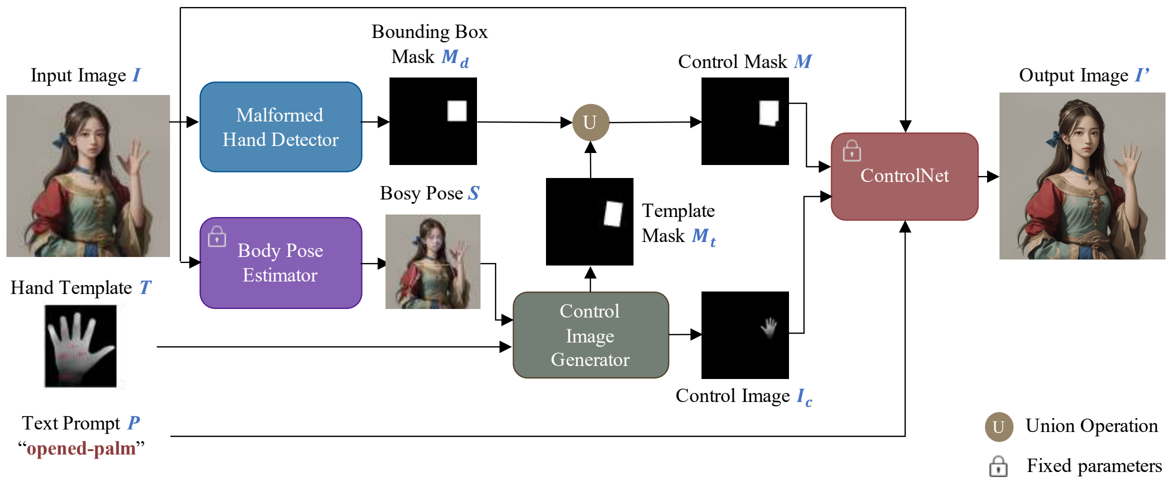
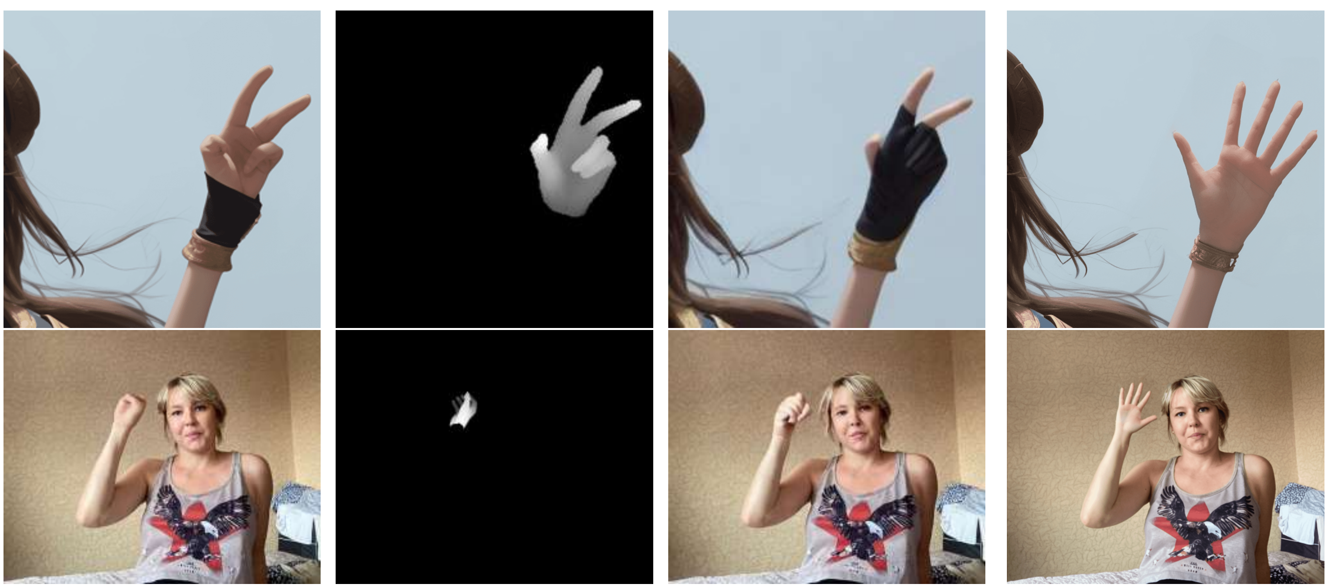


Figure 2. HandCraft flowchart. The framework has three stages for correcting malformed hands in images. (1) Hand detection. A malformed hand detector is employed to detected the bounding box of the hand and a body pose estimator is used to predict the landmarks on hands with the prior of the whole body pose. (2) Control image generation. The extracted body pose and a parametric hand template are given to a control image generator to obtain a control image I_c and a template mask M_t. The final control mask M is obtained by doing a union operation between the bounding box mask M_d and the template mask M_t. (3) Hand restoration. The final output image with corrected hand is generated using ControlNet given the input image, a text prompt, control mask and control image as the conditioning.

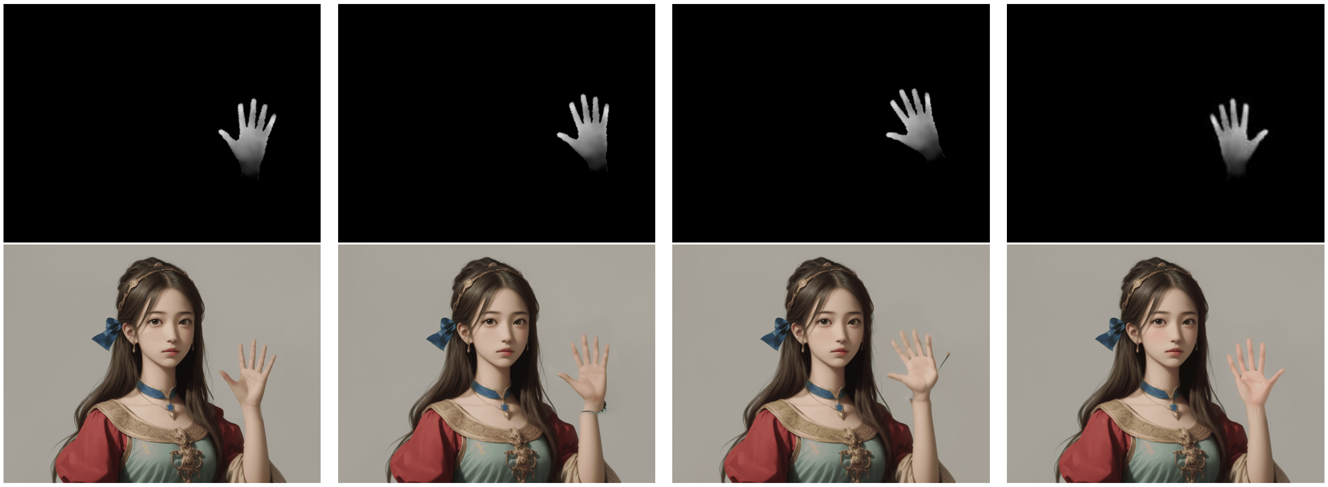


(a) Malformed hand (b) Mesh fitting (c) Restore via mesh (d) Using template

Figure 3. Comparison of hand restoration methods: (a) The original image with a deformed hand where fingers are bent in an unnatural manner or are missing. (b) The result of mesh fitting, mimicking the incorrect finger alignment and positioning from the original, resulting in a hand orientation that does not match the natural pose. (c) The outcome of attempting restoration with the flawed mesh, maintaining the unnatural bending of the fingers, or resulting in a malformed hand inconsistent with the mesh condition. (d) The hand restored using a predefined template, which achieves a natural-looking hand pose and maintains anatomical accuracy.

different hand template parameters, and automatically selecting the render that most closely matches the silhouette, as shown in Fig. 6.

Accurate Positioning and Orientation. The restoration of deformed hands necessitates that the hand is not only anatomically accurate but also precisely positioned and oriented. This means that the hand template T should align correctly in terms of location, rotation, and handedness (left or right hand), corresponding to the detected deformation. Fig. 4 illustrates that an inaccurate rotation will result in misalignment between the hand and the wrist, and that the



(a) No Rotation (b) 15° Rotation (c) 30° Rotation (d) Flipping

Figure 4. **Impact of Template Rotation on Hand Restoration.** (a) The original hand without any rotation. (b) The effect of a 15° rotation, leading to a misalignment at the wrist and disrupting the natural flow from the forearm to the hand. (c) The effect of a 30° rotation, which further exaggerates the misalignment and creates an unnatural hand shape. (d) The effect of flipping the hand template, resulting in a mirrored appearance that is anatomically incorrect for that side of the body. These examples highlight the importance of accurate rotational alignment and proper handedness, where even minor inaccuracies can significantly compromise the restoration's anatomical precision.

Detected Keypoints Alignment Keypoints Overlaid Template

Figure 5. **Alignment of Hand Keypoints.** The left image illustrates a real hand with keypoints s_1, s_2, s_3 and s_4 detected by a pose estimation algorithm. These points correspond to critical anatomical landmarks necessary for accurate hand posing. The right image displays a hand template with annotated keypoints t_1, t_2, t_3, and t_4, which are intended to align with the keypoints of the real hand after correcting for scale, position, and rotation. The process involves scaling the template based on vector lengths, moving it to match the keypoint positions, and rotating it accordingly.

Input Silhouette Restored

Figure 6. Silhouette-guided consistent hand restoration.

wrong handedness (chirality) leads to generation errors.

As shown in Fig. 5, the procedure is as follows.

1. **Identification of Keypoints:** Utilizing pose estimation (*e.g.*, MediaPipe [20]), we identify keypoints $S_h = \{s_1, s_2, ..., s_n\} \subset S$ on the deformed hand within M_d. Corresponding keypoints are also defined on the hand template T, denoted as $T_h = \{t_1, t_2, ..., t_n\}$.
2. **Scaling:** The template T is scaled to match the size of the detected hand, based on the ratio of distances between key pairs in S_h and T_h, ensuring T fits the size of the deformation in M_d.
3. **Translation:** T is then translated such that its reference point (e.g., the base of the palm) aligns with the equivalent point in S.
4. **Rotation:** To correct for orientation, a rotation matrix $R(\theta)$ is applied to T, where θ is the angle calculated from the orientation discrepancy between S_h and T_h. For handedness correction, a conditional mirroring transformation may also be applied if necessary.

This ensures that the transformed template T aligns accurately with the orientation and position of the detected hand within the image I, effectively correcting the deformation within the region defined by $M = M_d \cup M_t$. This procedure highlights the significance of precise alignment in hand restoration, where even minor deviations can lead to an unnatural appearance. By applying these steps, we ensure that the restored hand is correctly aligned and integrated within the input image, maintains anatomical accuracy and seamlessly blends into the original scene, thus preserving the overall authenticity of the image.

3.3. Hand Restoration

To ensure that the hand restoration process is targeted and does not inadvertently modify other parts of the image, we introduce the concept of a restoration region. This guides ControlNet [29] to concentrate exclusively on the identified deformity. Such a strategy reflects our goal of maintaining the original image's integrity, ensuring only the malformed hand is rectified.

The restoration region should: (1) encompass the entire area of the malformed hand, ensuring comprehensive correction, as represented by the mask M_d; and (2) be sufficiently large to accommodate the corrected hand shape, *i.e.*, the hand template T, as represented by the template bounding box mask M_t. This ensures that the restoration does not introduce any spatial constraints that could compromise the correction's effectiveness. The final restoration region, denoted as M, is then given by the union of M_d and M_t ($M = M_d \cup M_t$). This union ensures that the restoration region is optimally sized to cover both the detected deformity and the area required for the corrected hand shape.

This methodical definition of the restoration region is pivotal, as it directly influences the restoration's quality and the preservation of the image's overall composition. Experimental analyses affirm the efficacy of this dual-region approach, highlighting its superiority in achieving precise and aesthetically cohesive hand restorations within the broader context of the original images.

4. Experiments

In this section, we present our experiments, where we evaluate the performance of HandCraft qualitatively and quantitatively and compare it to a baseline method.

4.1. Dataset

We propose MalHand datasets, including training and evaluation datasets. The evaluation dataset is divided into photorealistic images (MalHand-realistic) and stylized artistic images (MalHand-artistic). Here, we outline the process of generating these datasets.

Training data. While using a pretrained hand detection model, such as YOLOv8 [18], provides satisfactory performance to detect the malformed hands, greater accuracy can be achieved by finetuning the model. To do so, it is necessary to compile a training dataset with malformed hands and their locations. For this purpose, we utilize the HaGRID dataset [13], which contains portrait photos featuring hands, along with bounding boxes that mark the hand positions.

To create instances of malformed hands for our training data, we leveraged Stable Diffusion [25] to modify the hands within the provided bounding boxes, using "hands" as the guiding text prompt. This process aimed to generate a variety of hand abnormalities similar to those encountered in images produced by Stable Diffusion models. After generating these modified hands, we manually selected images that clearly displayed malformed hands. This selection process resulted in a dataset comprising 60,000 images, each featuring at least one malformed hand.

Additionally, to ensure the model can distinguish between malformed and normal hands, we also evaluate our metrics on the unaltered images from HaGRID [13]. The bounding boxes from the original dataset were preserved to provide the locations of both normal and malformed hands.

Evaluation data. To assess the effectiveness of our algorithm in restoring malformed hands across various styles, we compiled a dataset comprising 1,500 portrait images featuring malformed hands. This dataset is divided into two categories: 1,000 images in a realistic style (MalHand-realistic) and 500 images in artistic styles (MalHand-artistic). By incorporating a mix of styles, we aim to evaluate the algorithm's robustness and adaptability to different visual representations. The realistic images consist of those sourced from the HaGRID dataset [13]. The generating process is similar to the training data, using a different set of images. These images are used for quantitative evaluation.

We also generated artistic-style portrait images using Stable Diffusion for qualitative evaluation, including Japanese anime and Disney cartoon styles, among others. Complete prompts and instructions for generation are provided in the supplement. We then manually identified portraits with malformed hands. Bounding boxes for these malformed hands were obtained via crowdsourcing, with detailed instructions provided to ensure consistency and accuracy. These instructions are included in the supplement.

4.2. Evaluation Metrics

Mean hand pose confidence. This metric assesses the naturalness of the hand's anatomy by averaging the confidence scores predicted by Mediapipe [20] across all detected hand *keypoints*. Let c_{ij} denote the confidence of the j^{th} hand keypoint out of the set of detected keypoint indices $\mathcal{J}_i$ for hand i in a dataset containing N hands. Then the mean hand pose confidence is given by

$$\bar{c}_{\text{pose}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{J}_i|}\sum_{j\in\mathcal{J}_i} c_{ij}. \tag{1}$$

Mean hand classifier confidence. This metric assesses the model's ability to generate hands that can be confidently classified as non-malformed by our YOLOv8-based hand detector [18]. Let c'_i denote the confidence of the hand classifier for hand i in a dataset containing N hands. Then the mean hand classifier confidence is given by

$$\bar{c}_{\text{classifier}} = \frac{1}{N}\sum_{i=1}^{N} c'_i. \tag{2}$$

Masked peak signal-to-noise ratio (PSNR) / masked structural similarity index measure (SSIM). These metrics assess how well the image outside the restored area is

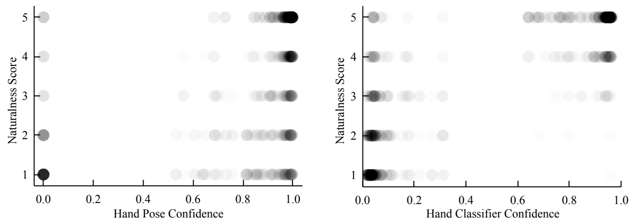
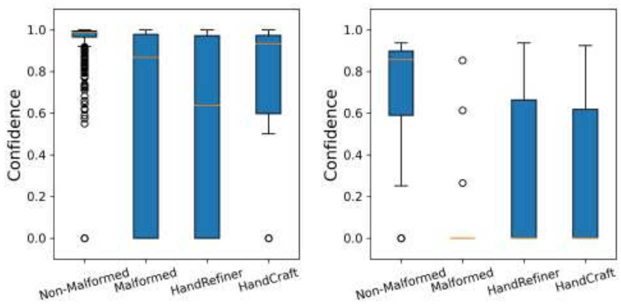


Figure 7. Human hand naturalness score has positive correlation with c_{pose} and $c_{\text{classifier}}$



(a) c_{pose} (b) $c_{\text{classifer}}$

Figure 8. Box plots demonstrating the performance of different methods in restoring anatomical correctness to images with malformed hands. (a) The hand pose confidence c_{pose}, where HandCraft shows a notable improvement over HandRefiner, closely aligning with the non-malformed control group. (b) The hand classifier confidence $c_{\text{classifier}}$, where HandCraft's restorations achieve comparable confidence levels to HandRefiner.

preserved, at a per-pixel and structural level: whether the restoration has altered or corrupted the rest of the image.

Validation. User studies were conducted to investigate whether the confidence scores (*i.e.*, c_{pose} and $c_{\text{classifier}}$) correlated with naturalness. Seven unaffiliated individuals rated the naturalness of hands in each image, from a stratified random subset of 200 restored images, on a scale of 1 to 5, where 1 is least natural and 5 is most natural. As shown in Fig. 7 (opacity 1%), the hand pose confidence score (c_{pose}) and the hand classifier confidence score ($c_{\text{classifier}}$) correlates with perceived naturalness. The average Pearson's correlation coefficient between c_{pose} and human-rated naturalness scores is 0.44 and the average correlation for $c_{\text{classifier}}$ is 0.82.

4.3. Comparison with Control Images

We quantitatively compare the anatomical accuracy of HandCraft's restored images to a control group of images without malformed hands. Two sets are assembled for evaluation: a control set $\mathcal{D}_{\text{C}}$ composed of the real images from the HaGRID dataset [13] and a set $\mathcal{D}_{\text{R}}$ composed of realistic images with malformed hands from the MalHand-realistic dataset, which have undergone hand restoration with HandCraft. For each image, we calculate the hand pose confidence c_{pose} and hand classifier confidence $c_{\text{classifer}}$, reflecting

Table 1. Quantitative comparison of hand restoration methods on the MalHand-realistic dataset. We report the mean hand pose confidence ($\bar{c}_{\text{pose}}$) to assess the accuracy of the hand restoration, the mean hand classifier confidence ($\bar{c}_{\text{classifier}}$) to assess whether a classifier deems the hand as non-malformed, the masked peak signal-to-noise ratio (PSNR) to assess the visual fidelity outside of the hand region, and the masked structural similarity index measure (SSIM) to assess any change in structural content. The latter two are calculated between the input image I with malformed hands and restored images.

Method	$\bar{c}_{\text{pose}} \uparrow$	$\bar{c}_{\text{classifier}} \uparrow$	PSNR ↑	SSIM ↑
Null Intervention	0.68	0.00	N/A	N/A
HandRefiner [19]	0.54	**0.25**	12.93	0.3839
HandCraft (Ours)	**0.79**	**0.25**	**23.40**	**0.6462**

anatomical correctness as perceived by hand detection and pose estimation algorithms, with higher scores indicating closer resemblance to authentic hand anatomy.

Fig. 8 shows overlapping hand pose confidence intervals between the restored and Non-Malformed images. While there is a statistically significant difference between the two groups, this is partly due to inherent differences between real and restored generated images. The overlap in mean hand pose confidence scores between HandCraft restorations and the Non-Malformed group indicates HandCraft's superior performance in restoring anatomical correctness compared to HandRefiner.

4.4. Comparison with HandRefiner

To demonstrate the effectiveness of HandCraft, we compare its performance with the current state-of-the-art, HandRefiner [19], on the MalHand-realistic dataset. The results in Tab. 1 indicate that HandCraft outperforms the null intervention (no restoration) and HandRefiner [19] in terms of hand pose confidence and is comparable in terms of hand classifier confidence. HandCraft also achieves a higher masked PSNR and SSIM in the non-hand regions, reflecting its ability to preserve the integrity of the image while correcting the hand anatomy. These results validate the efficacy of our method in generating realistic and natural-looking hand postures without compromising the quality of the original image. The results are further supported by qualitative comparisons conducted on the MalHand-artistic dataset, as shown in Fig. 9. These visual examples clearly demonstrates that HandCraft not only corrects the malformed hands but also does so with a seamless integration into the original portrait, surpassing the performance of HandRefiner [19]. As shown at the bottom right of Fig. 9, we infrequently observe artifacts at the border of the regenerated area. An algorithm to mitigate this is presented in the supplement, alongside more qualitative results, failure cases and a discussion of limitations.

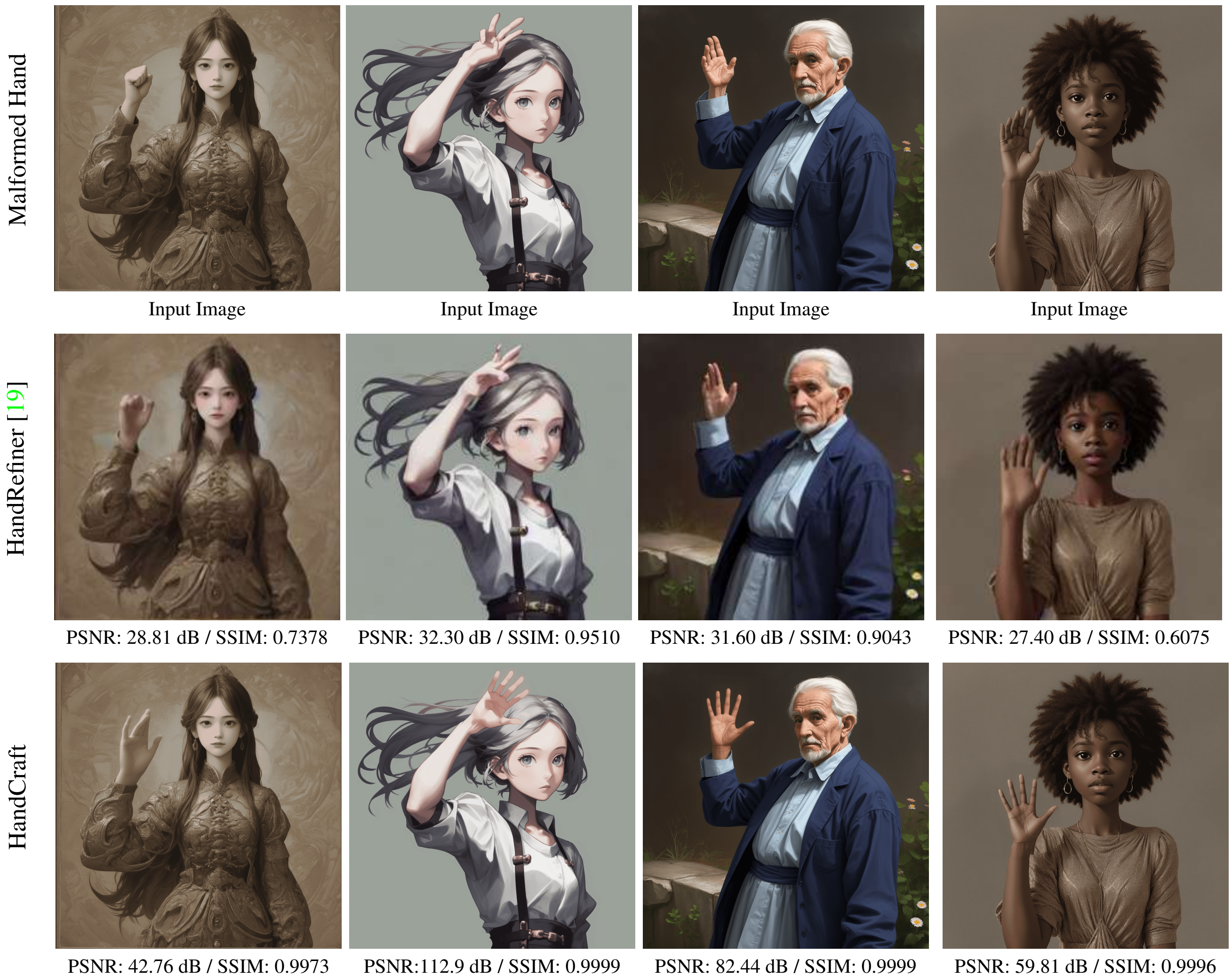


Figure 9. Comparison between various hand restoration techniques. HandCraft demonstrates superior performance in hand rectification tasks, seamlessly integrating corrected hand into the original portraits. PSNR and SSIM metrics reveals that HandCraft achieves this while minimizing perturbations to non-hand regions, significantly outperforming the baseline HandRefiner [19] method.

5. Discussion and Conclusion

Despite advances in generative image models, the issue of malformed hands has persisted across generations of such models, even in recently-released state-of-the-art systems like Stable Diffusion XL [24] and Sora [6]. We provide visualizations in the supplement highlighting instances of anatomically incorrect hands produced by these latest models. As such, we expect HandCraft to remain useful for some time, since this issue appears to be quite persistent. Even if this challenge is overcome by a new generative model, our method provides functionality to change hand gestures and poses for creative control, offering utility beyond just correcting malformed hands.

Our HandCraft framework addresses the challenge of correcting malformed hands in images generated by text-to-image diffusion models. Through the use of a parametric hand model to guide a diffusion-based image editor, we achieve seamless anatomical corrections that integrate with the original image's aesthetics. Our approach is both effective and accessible, requiring no additional training. The accompanying Malhand datasets further enriches the field by providing resources for training and benchmarking. Furthermore, comparisons with a state-of-the-art method HandRefiner [19] showed that HandCraft not only surpassed it in restoring hand anatomy but also maintained the integrity of the rest of the image. We hope that the proposed HandCraft will be useful for artists, designers, and developers.

References

[1] Stable diffusion finger repair method with various gesture posture adjustments (in chinese), 2024. https://youtu.be/MRiVy2MgWCU?si=Wjp72LIblKrNjDmb. 1

[2] Ahfaz Ahmed. How to fix hands in stable diffusion. In *OpenAI Journey*, Feb 2024. 1

[3] Endangered AI. Ultimate way to fix hands, 2024. https://youtu.be/sKaEd2XfqZE?si=7KI-qIHar7tNopI9. 1

[4] Mikhail Avady. Bad hands, 2024. https://civitai.com/models/116230/bad-hands-5. 1

[5] Mikhail Avady. Stable diffusion negative prompts, 2024. https://github.com/mikhail-bot/stable-diffusion-negative-prompts. 1

[6] T Brooks, B Peebles, C Homes, W DePue, Y Guo, L Jing, D Schnurr, J Taylor, T Luhman, E Luhman, et al. Video generation models as world simulators. *OpenAI*. 8

[7] DigitalDreamer. Fix hand, 2023. https://civitai.com/models/103942/fix-hand. 1

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2

[10] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Eur. Conf. Comput. Vis.*, pages 172–189, 2018. 2

[11] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2

[12] Sebastian Kamph. Controlnet guidance tutorial. fixing hands?, 2023. https://youtu.be/wNOzW1N_Fxw. 1, 3

[13] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid–hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4572–4581, 2024. 6, 7

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Machine Learning (ICML)*, 2013. 2

[15] SUJEET KUMAR. Ways to fix hand in stable diffusion, 2024. https://www.pixcores.com/2023/05/fix-hand-in-stable-diffusion. 1

[16] Yuan Li, Xiangyang He, Yankai Jiang, Huan Liu, Yubo Tao, and Lin Hai. Meshformer: High-resolution mesh segmentation with graph transformer. In *Computer Graphics Forum*, volume 41, pages 37–49. Wiley Online Library, 2022. 3

[17] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2022. 1

[18] Ultralytics LLC. Ultralytics yolov8, 2023. https://github.com/ultralytics/ultralytics. 3, 6

[19] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. *arXiv preprint arXiv:2311.17957*, 2023. 1, 3, 7, 8

[20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 3, 5, 6

[21] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley from the field. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012. 1

[22] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, and Minh Hoai1. Handiffuser: Text-to-image generation with realistic hand appearances. In *IEEE/CVF Conference on Computer Visision Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Stability AI*, 2023. 8

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 6

[26] Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[27] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023. 1

[28] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *IEEE/CVF Conference on Computer Visision Pattern Recognition (CVPR)*, pages 5505–5514, 2018. 2

[29] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 6

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251. IEEE, 2017. 2