

Detecting and Restoring Non-Standard Hands in Stable Diffusion Generated Images

Yiqun Zhang ^{*1} Zhenyue Qin ^{*1} Yang Liu ¹ Dylan Campbell ¹

Abstract

We introduce a pipeline to address anatomical inaccuracies in Stable Diffusion generated hand images. The initial step involves constructing a specialized dataset, focusing on hand anomalies, to train our models effectively. A finetuned detection model is pivotal for precise identification of these anomalies, ensuring targeted correction. Body pose estimation aids in understanding hand orientation and positioning, crucial for accurate anomaly correction. The integration of ControlNet and InstructPix2Pix facilitates sophisticated inpainting and pixel-level transformation, respectively. This dual approach allows for high-fidelity image adjustments. This comprehensive approach ensures the generation of images with anatomically accurate hands, closely resembling real-world appearances. Our experimental results demonstrate the pipeline's efficacy in enhancing hand image realism in Stable Diffusion outputs. We provide an online demo at fixhand.yiqun.io.

1. Introduction

Stable Diffusion (Rombach et al., 2022) has become increasingly prominent in generating human images. Its proficiency in creating realistic human representations, particularly for real-time applications like gaming or augmented reality, is noteworthy. However, a recurrent issue with this model is its tendency to produce inaccurate hand images, a problem we define as the "non-standard hand." Refer to Figure 1 for examples, where non-standard hands may display irregularities such as missing or extra fingers, disproportionate sizes, or structurally incorrect hands. These discrepancies, while minor, can greatly affect the perceived realism and authenticity of the images.

^{*}Equal contribution ¹Australian National University. Correspondence to: Zhenyue Qin <kf.zy.qin@gmail.com>, Yiqun Zhang <yiqun@admin.io>, Yang Liu <lyf1082@gmail.com>, Dylan Campbell <dylan.campbell@anu.edu.au>.

The concept of the uncanny valley (Mori, 1970) describes a sense of unease or discomfort when humanoid figures closely resemble humans but are not quite lifelike. This effect is relevant when considering the Stable Diffusion model's hand image generation, where images with "non-standard hands" may evoke the uncanny valley phenomenon. In fields like augmented reality (AR), virtual reality (VR), and gaming, where a seamless and realistic experience is crucial, such anomalies in hand images can be particularly unsettling for users. Non-standard hands may disrupt user engagement, detracting from the effectiveness of AR, VR, and gaming applications. Accurate hand representation is, therefore, essential not just for visual appeal, but also for the functionality and overall user satisfaction in interactive virtual settings.

Other models for creating images, like Variational Autoencoders (VAEs) (Kingma & Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), exist besides Stable Diffusion. But the easier training of the Diffusion model is a big plus over GANs, which need adversarial training, and VAEs, which need a variational posterior. The Stable Diffusion uses UNet architecture (Ronneberger et al., 2015) from image segmentation and shows stable loss during training and very good performance. We think Stable Diffusion is more useful for practical situations. That's why we focus on solving the problem of non-standard hands in Stable Diffusion.

Despite the evident necessity to resolve this challenge, there's a conspicuous scarcity of comprehensive research in this area. Our contribution is a structured method that initially identifies 'non-standard' hand depictions and subsequently corrects them to match the appearance of actual human hands, referred to here as "standard hands". We have devised an integrated pipeline for detection and correction. The initial phase employs bounding boxes to distinguish between "standard" and "non-standard" hands within images. Utilizing a specialized dataset and the advanced capabilities of the YOLOv8 algorithm (LLC, 2023), our approach attains notable accuracy in detection, with Figure 3 demonstrating an instance of the detection phase.

Upon identifying anomalies, our system progresses to the restoration phase. The process involves a set of defined



Figure 1. A compilation of images with non-standard hand anomalies, highlighting the varied manifestations of the issue.

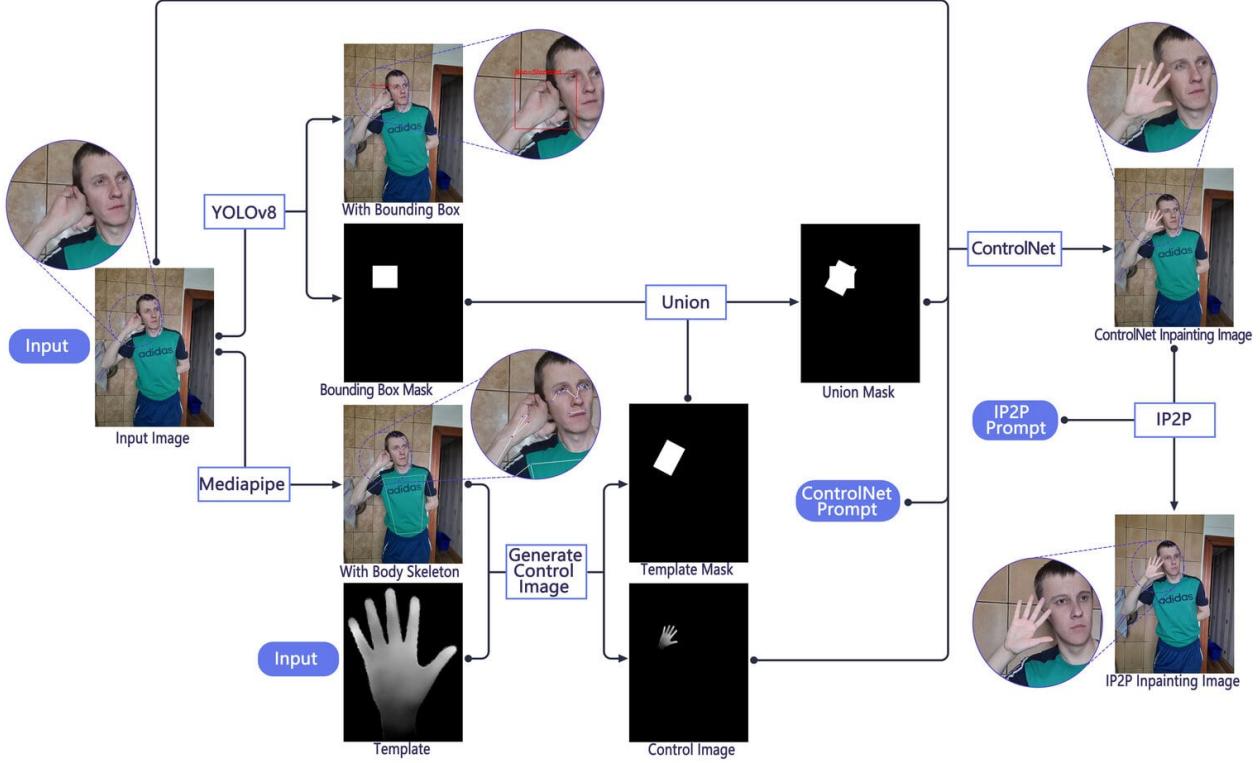


Figure 2. This flowchart outlines our proposed pipeline: Initially, an image with a non-standard hand as the input. We then employ YOLOv8 to delineate the non-standard hand using a bounding box, creating what we term the “bounding box mask”. MediaPipe is utilized to compute the body skeleton. Based on this skeleton, a template is accurately positioned over the non-standard hand to create the “control image”. The control image’s bounding box and the bounding box mask are combined to generate the “union mask”. Using this union mask, the control image, and a descriptive template prompt, we repair the area covered by the mask. Subsequently, IP2P and its associated prompt are used to refine the texture, resulting in the final output.

operations: *Body Pose Estimation*, utilizing Google’s Mediapipe (Lugaresi et al., 2019) to determine the hand’s position and movement; *Control Image Generation*, which provides

the Stable Diffusion model with directive images for better restoration outcomes; *ControlNet Inpainting*, offering an initial refinement based on the Control Image; and ultimately,

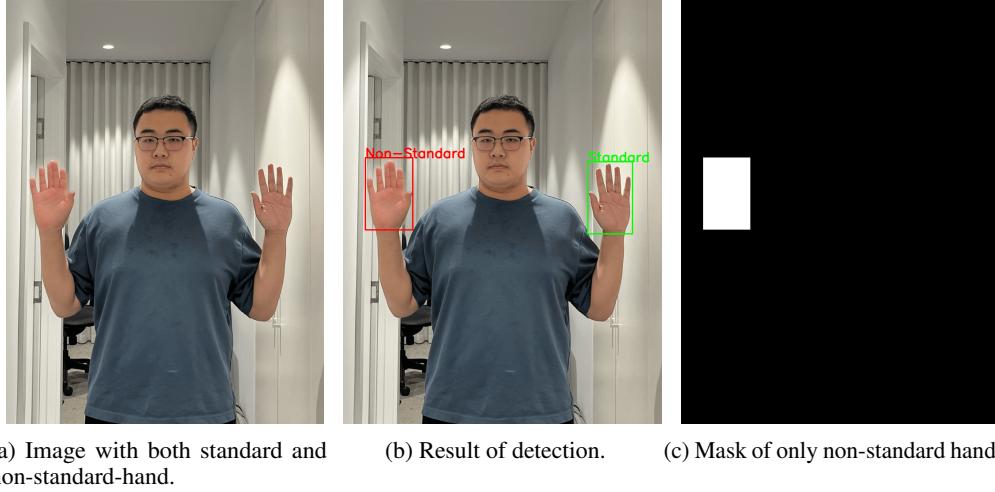


Figure 3. Examples of detection. Image is an author’s photo.

IP2P Inpainting, which enhances the images with realistic textures and accurate details. An illustration of this pipeline is depicted in Figure 4.

In this report, we present a solution that improves the image quality produced by Stable Diffusion, particularly the depiction of hands. Our finetuned YOLOv8 model effectively differentiates between non-standard and standard hand representations. Furthermore, in the restoration phase, our pipeline converts non-standard hands to more accurately resemble their standard counterparts.

We summarize our contributions as follows:

1. We create a pipeline to rectify anatomical inaccuracies in hand images. Our results demonstrate the effectiveness in producing anatomically accurate and realistic hand images in outputs from Stable Diffusion.
2. We finetune a detection model to locate and classify standard and nonstandard hands, and fine-tuned an InstructPix2Pix model to make high-fidelity adjustments to the images. We will make these models available.
3. We create a dataset featuring a diverse collection of hand images, both standard and nonstandard, to facilitate comprehensive model training. We plan to release this dataset publicly in the future.
4. Demo of this pipeline is available: fixhand.yiqun.io.

2. Method

Stable Diffusion occasionally generates images with atypical hands, defined as non-standard hands. Our method is to identify these variations and then adjust them to resemble real-world hands, defined as standard hands.

Our strategy consists of two main parts: detection and restoration. The detection stage involves pinpointing both standard and non-standard hands in images, highlighting them with bounding boxes. This step relies on a specialized dataset and a finely-tuned detection model. After detection, we proceed to the restoration phase. Our pipeline integrates these phases, including steps like *Non-standard Hand Detection*, *Body Pose Estimation*, *Control Image Generation*, *ControlNet Inpainting*, and *IP2P Inpainting*, as shown in Figure 1. These steps collectively enable the correction of non-standard hands. The following sections provide detailed insights into each part of the process. The entire workflow is depicted in Figure 2.

2.1. Non-Standard Hand Detection

Non-standard hand detection phase is designed to locate the bounding boxes of all hands present within an image. Moreover, it categorizes these hands into two distinct categories: non-standard hand and standard hand.

Non-Standard Hand Dataset

For accurate localization and categorization of hands in images, constructing a dedicated dataset is essential. This dataset should encompass a wide range of images featuring hands. Crucially, every hand in these images requires annotation. This includes both the bounding box around the hand and a classification label indicating whether it is a non-standard or standard hand. Such a dataset is fundamental to the effectiveness of our approach.

We begin with HaGRID (HAnd Gesture Recognition Image Dataset) (Kapitanov et al., 2022) as our foundational dataset. HaGRID comprises 552,992 RGB images, each containing hands in a variety of positions. Each image is annotated with bounding boxes to indicate hand locations. The dataset

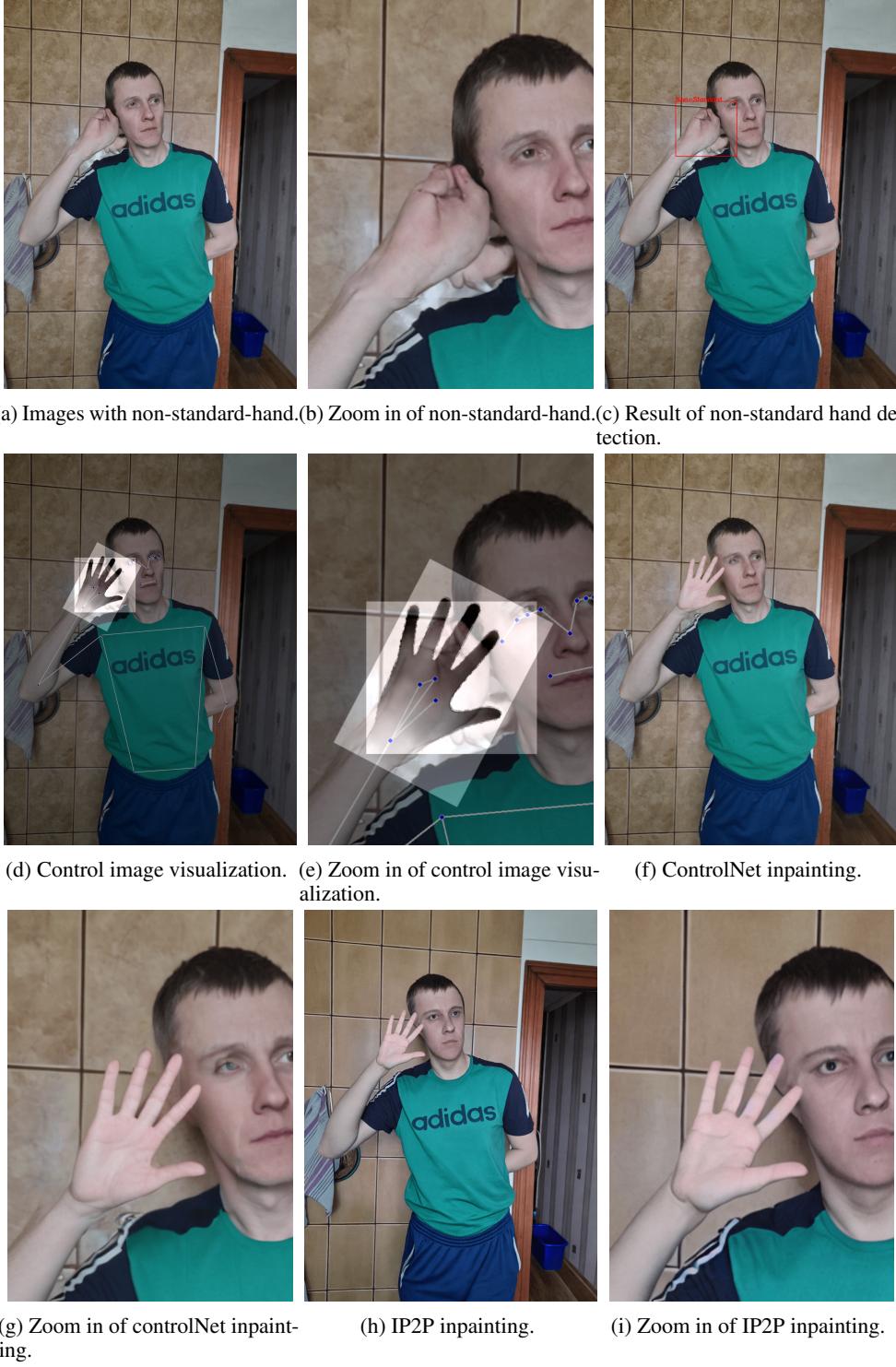


Figure 4. Pipeline overview.

represents a wide array of 18 hand gestures, performed by a diverse group of 34,730 individuals, aged 18 to 65. Some images depict one hand, while others show both. The photos were taken indoors under different lighting conditions,

providing a rich assortment of visual details. This diversity grants our dataset extensive coverage and high levels of generalizability, enhancing our model's capability to process hands in a wide range of appearances. Illustrations from



Figure 5. HaGRID dataset samples and samples with bounding boxes.

HaGRID, including examples of bounding box annotations, are shown in Figure 5.

Our approach began by randomly selecting 30,000 images from the HaGRID dataset. Since these images are authentic photographs, all hands within this chosen subset are classified as standard hands. We employed the Stable Diffusion model (Rombach et al., 2022) to recreate the hand areas, as outlined by their bounding boxes in HaGRID images. Bounding box information was directly obtained from the HaGRID dataset. Consequently, for every original image from HaGRID, we created a corresponding redrawn image. Examples of these redrawn images are displayed in Figure 6.

Due to certain limitations of the Stable Diffusion model, some hands in the redrawn images were classified as non-standard hands. From these redrawn images, we manually selected samples featuring non-standard hands, pairing each with its corresponding original image, which depicts a standard hand. We annotated each image in this set with labels and bounding boxes to highlight the presence and location of the hands. To enable a comprehensive evaluation of our model, we divided this data into training and testing sets. This dataset is specifically prepared for detecting non-standard hands.

Hand Detection

YOLOv8 (LLC, 2023) is one of the state-of-the-art models for object detection. To identify and classify hands as either non-standard or standard, we finetuned the YOLOv8 model using the training dataset described above. After implementing the trained YOLO model, we annotate the bounding boxes around the hands and classify them. Only the non-standard hands are chosen for further restoration. These bounding boxes are transformed into the 'bounding box masks', as shown in Figure 7. All examples are derived from the test set. The illustration demonstrates the finetuned model's ability to accurately locate and classify the hand.

2.2. Body Pose Estimation

Estimating body pose is crucial for determining the size position, and chirality of hands in our images. MediaPipe (Lugaresi et al., 2019) provides various solutions for vision tasks. Notably, its pose landmark detection feature is capable of detecting the human body skeleton. This includes a machine learning model skilled at identifying body landmarks such as hands, elbows, shoulders, hips, and more in images or videos, and their structural interconnections, as shown in Figure 8.



Figure 6. Redrawn samples by Stable Diffusion.

We employ MediaPipe because it provides three unique landmarks for the hand, as depicted in Figure 9a. These landmarks are crucial for accurately determining the hand’s size, position, and gesture chirality. In comparison, other 2D pose estimation models (Fang et al., 2022; Xu et al., 2022) typically provide information only up to the wrist. Moreover, MediaPipe’s detailed body skeleton detection remains reliable even when hand images in non-standard hand pictures are blurry or unclear, ensuring a certain degree of prediction accuracy.

Once detected, the skeletal landmarks were overlaid on the redrawn image. The hand region, highlighting these landmarks, is exhibited in Figure 9a. Four key landmarks are identified on the hand: a , b , c , and d . By defining vectors $v_1 = \vec{ac}$ and $v_2 = \vec{bd}$, we calculate their cross product $v_1 \times v_2$. The sign of this cross product helps classify hand orientation. A negative value indicates a *CW hand* (clockwise rotation from v_1 to v_2), while a positive value signals a *CCW hand* (counter-clockwise rotation from v_1 to v_2). This method simplifies the process of determining the hand’s chirality.

2.3. Control Image Generation

ControlNet (Zhang et al., 2023) enhances the Stable Diffusion model by introducing additional input modalities, including key landmarks, depth maps, and edge maps. This integration directs Stable Diffusion to produce images with more stable and realistic structures. Such advancements are crucial in restoring hand images, ensuring that their shape, size, and contours align accurately with real-world characteristics.

In refining the restoration process, we introduce a ControlNet image as an additional conditioning input. This image, tailored to enhance Stable Diffusion’s restoration of non-standard hands, pushes the model to its limits and im-

proves the output quality. Examples (Kamph, 2023) include opened-palm (see Figure 9b) and fist-back (see Figure 9c), collectively termed as hand templates. Adding more templates is feasible. A key part of this process is accurately placing these hand-templates at the correct hand locations, forming the Control Image. The accuracy of this placement is crucial as it significantly impacts ControlNet’s effectiveness in rendering a precise hand representation. Our methodology for this includes:

1. Like the hand landmarks identified by MediaPipe, we annotate four specific points (a', b', c', d') on the Templates, as shown in Figure 9d. This helps us form vectors $v'_1 = \vec{a'c'}$ and $v'_2 = \vec{b'd'}$.
2. We select an appropriate Template based on the gesture and context of the image undergoing restoration.
3. A background image, identical in size to the redrawn image but entirely black, is prepared (see Figure 10a). The Template is initially positioned at the top-left corner of this background (see Figure 10b). Simultaneously, a white image of the same size as the Template, called the template mask, is also placed at the top-left of the background (see Figure 10c).
4. The chirality (CW or CCW) of the hand is determined. If it does not match the Template’s chirality, the template is flipped.
5. **Scaling:** The Template and template mask are scaled by the ratio $\frac{|v_1|}{|v'_1|}$.
6. **Moving:** The Template and template mask are moved by the vector difference $v_1 - v'_1$.
7. **Rotation:** The Template and template mask are rotated by an angle θ . The angle is calculated as:

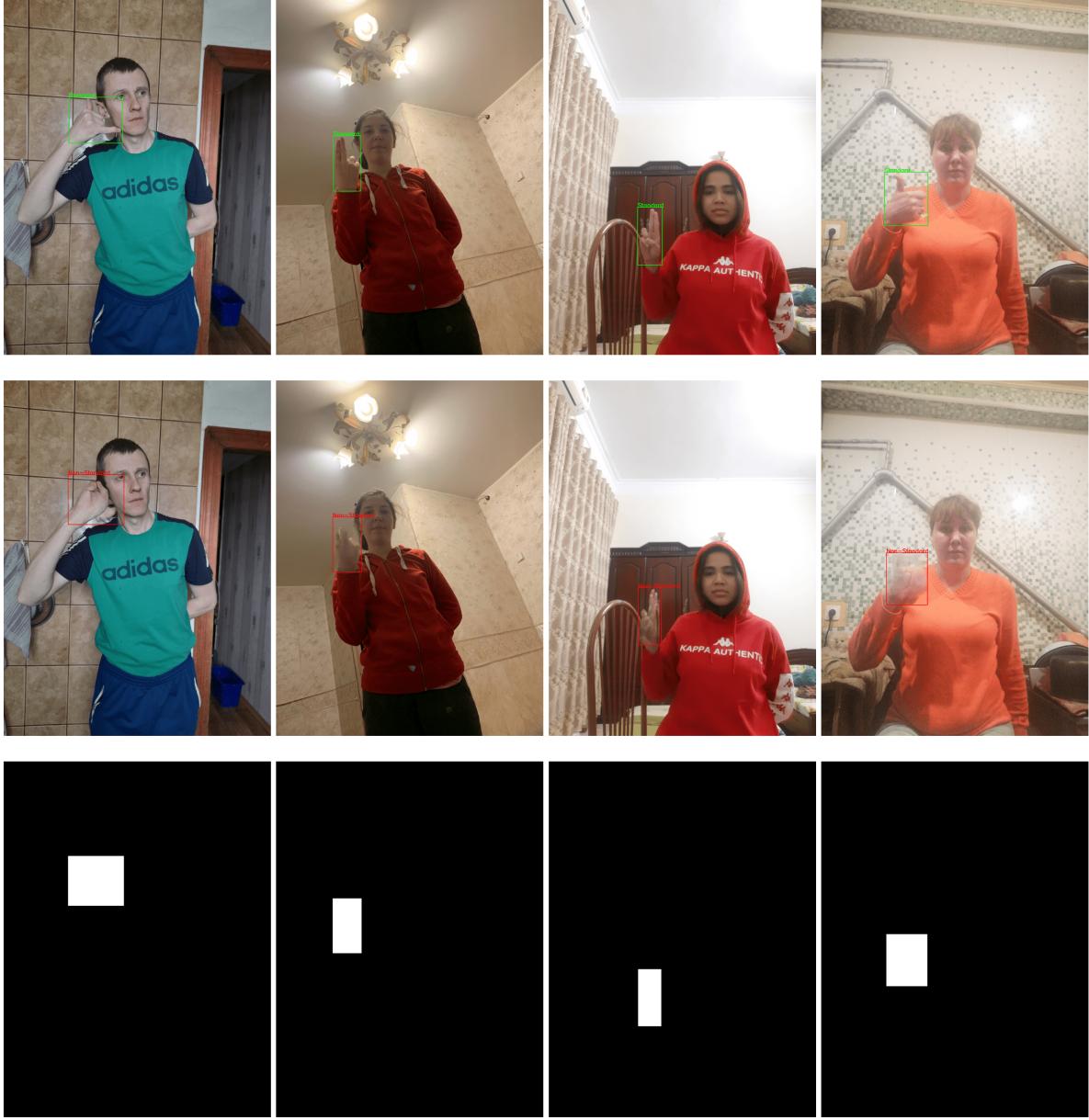


Figure 7. Original samples with detected bounding box, redrawn samples with detected bounding box and redrawn samples with bounding box mask.

$$\theta = \arccos \frac{aa' + cc'}{\sqrt{a^2 + c^2} \sqrt{a'^2 + c'^2}}$$

If $a \times c' - c \times a' > 0$, we rotate clockwise; otherwise, we rotate counterclockwise.

8. The final image with the Template is termed the control image (see Figure 11a).
9. The combination of the template mask and the bound-

ing box mask forms the union mask (see Figure 11b).

10. For a comprehensive view of the process, we compile samples of the redrawn images, body skeleton, control image, and union mask into a single visual representation (see Figure 11c).

Utilizing both YOLOv8 and MediaPipe results is crucial for this step. While MediaPipe detects occluded hands, it cannot differentiate between non-standard and standard



Figure 8. Redrawn samples with body skeleton detected by MediaPipe.

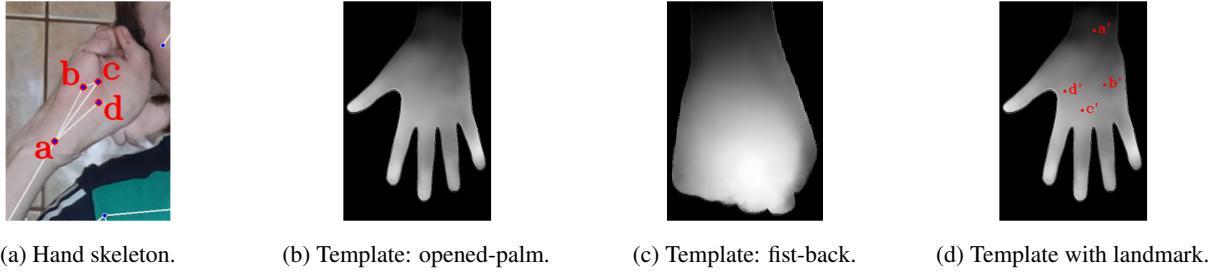


Figure 9. Hand and template details.

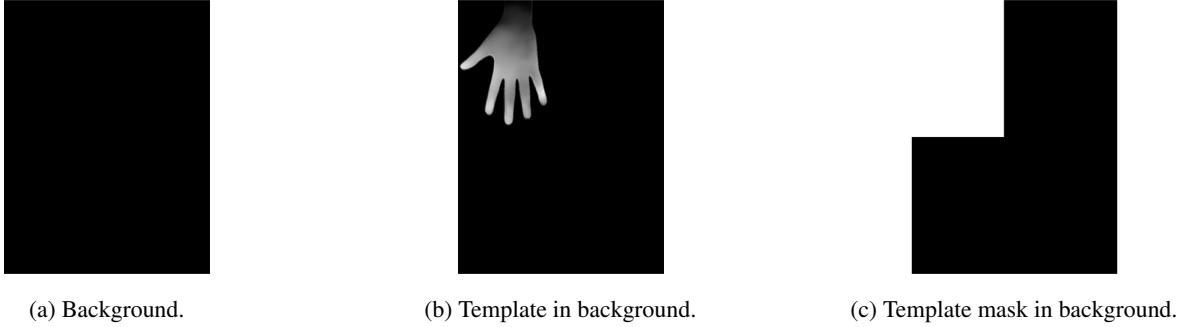


Figure 10. Background, template in background, and template mask in background.

hands. YOLOv8 compensates for this limitation. Through the methodology described above, we guarantee that the Control Image offers essential cues for ControlNet to accurately reconstruct hand images.

2.4. ControlNet Inpainting

After establishing the hand's position and shape, we engage ControlNet for image restoration. This advanced process relies on the Union Mask to focus restoration efforts on specific areas of the image for precise improvement.

The control image, accurately depicting the hand's position,

acts as a guide for ControlNet in restoring the redrawn image. We enhance this process with carefully selected prompts, aimed at providing detailed cues for desired image attributes. Our primary prompt is:

[TEMPLATE NAME], hand, realskin, photo-realistic, RAW photo, best quality, realistic, photo-realistic, masterpiece, an extremely delicate and beautiful, extremely detailed, 2k wallpaper, Amazing, finely detailed, 8k wallpaper, huge filesize, ultra-detailed, high-res, and extremely detailed.



Figure 11. Some samples in control image generation.

We also use ‘negative prompts’ to avoid unwanted outcomes, including: “deformed, EasyNegative, paintings, sketches, (worst quality:2), (low quality:2), (normal quality:2), low-res, normal quality, and (monochrome).”

These prompts, familiar within the community for generating high-quality images, reflect prompt engineering domain knowledge. Using both positive and negative prompts allows us to fully utilize ControlNet, producing images that are not only aesthetically pleasing but also realistic, as shown in [Figure 12](#).

2.5. IP2P Inpainting

The final phase of our method employs the *InstructPix2Pix* (IP2P) model ([Brooks et al., 2022](#)). Following the initial restoration with ControlNet, IP2P enhances the textures, focusing on giving the hands a more realistic and authentic appearance to blend seamlessly with the rest of the image.

Initially, the IP2P model is fine-tuned using our training set, which comprises 9623 pairs of images. Each pair includes a real image with a standard hand from the HaGRID dataset



Figure 12. ControlNet inpainting result.



Figure 13. IP2P inpainting result.

and a corresponding Stable Diffusion redrawn image with a non-standard hand. The real images were inputs during training, while the redrawn images served as outputs. Additionally, a prompt, “Turn the deformed hand into normal”, and its 50 variants (see ??) were also used as input. This fine-tuning enables the IP2P model to transform non-standard hands into standard ones with enhanced accuracy.

For restoration, the fine-tuned model is then applied. Unlike previous steps, no masking is used here. The entire image undergoes processing to ensure the hand’s texture matches the overall image style. This comprehensive approach employs the prompt: “Turn the deformed hand into normal”.

The outcome of this process, the final result of our restoration effort, is displayed in Figure 13.

3. Conclusion

Through the application of our approach, we observed encouraging results. The detection phase, utilizing the YOLO model (LLC, 2023), accurately pinpointed hand locations in images and classified them as non-standard or standard hands. We evaluated a test set of 2006 images, representing diverse scenes and gestures, using Precision and Recall metrics of 0.85, 0.90, 0.95. Our models performed well across these indicators, demonstrating the effectiveness and adaptability of our method in various real-world scenarios. This includes effectively detecting non-standard hands in images generated by Stable Diffusion.

The restoration phase further strengthened these outcomes. Each step, from body pose estimation to control image creation and inpainting processes, was precisely executed. We demonstrated using FID that the outcomes of ControlNet inpainting improved upon images with non-standard hands, and that IP2P inpainting further enhanced the results from ControlNet. Both ControlNet and IP2P inpainting proved

effective. We tested not only with redrawn samples from the HaGRID dataset but also with images generated by Stable Diffusion and real photographs. Our experiments show non-standard hand images transforming to more closely resemble standard hands.

References

- Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.
- Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., and Lu, C. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Kamph, S. Controlnet guidance tutorial. fixing hands?, 2023. URL https://youtu.be/wNOzW1N_Fxw.
- Kapitanov, A., Makhlyarchuk, A., and Kvachiani, K. Ha-grid - hand gesture recognition image dataset. *arXiv preprint arXiv:2206.08219*, 2022.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- LLC, U. Ultralytics yolov8, 2023. URL <https://github.com/ultralytics/ultralytics>. GitHub repository.
- Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.
- Mori, M. *Bukimi No Tani*. 1970.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.
- Xu, Y., Zhang, J., Zhang, Q., and Tao, D. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.
- Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models, 2023.