Diffusion-based Interacting Hand Pose Transfer

Junho Park[1,2*], Yeieun Hwang[1*], and Suk-Ju Kang[1✉]

[1] Department of Electronic Engineering, Sogang University, South Korea
[2] AI Lab, CTO Division, LG Electronics, South Korea
{junho18.park, yeieun1213}@gmail.com
sjkang@sogang.ac.kr

Abstract. We propose a new interacting hand pose transfer model, IHPT, which is a diffusion-based approach designed to transfer hand poses between source and target images. IHPT can generate a new target image with the target hand pose while maintaining the source image's texture and quality, leading to improved semantic understanding and generalizability in generating target hand poses. Experiments show that IHPT produces physically plausible and robust results for various text prompts and poses. Additionally, training a 3D hand mesh reconstruction network with IHPT-generated images enhances the performance in real-world scenarios, addressing the lack of in-the-wild 3D hand datasets and bridging gaps between indoor and outdoor environments.

Keywords: Pose Transfer · Image Generation · 3D Hand Mesh Reconstruction

1 Introduction

As large-scale foundation models [1–4] are developed, the AI community has evolved radically and tremendously. Therefore, they greatly impact multi-modal understanding, zero-shot learning, and transfer learning. Unfortunately, the correlation between hand-related research and foundation models is quite weak. In particular, after we discovered Stable Diffusion (SD) [4] generates hands bizarrely and weirdly, the need for study on foundation models with hand-related field has emerged.

Accordingly, several diffusion-based hand generation models [5–16] have been proposed in recent years. However, there are no studies for the hand pose transfer among them. Note that the pose transfer is a task of generating the target image from the target pose based on the source image. It has a wide range of applications including entertainment, virtual reality, fashion e-commerce, and human-computer interaction. Although it has been actively studied with respect to the person image synthesis [17–22], fewer studies for the hand image synthesis have been explored.

* Equal contribution.
✉ Corresponding author.

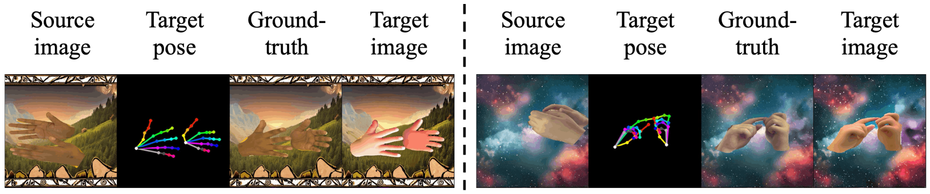

Fig. 1: Interacting hand pose transfer. It aims to generate a new target image similar to the ground-truth by transferring the hand of the source image and the pose of the target pose.

Thus, we first propose a diffusion-based interacting hand pose transfer model, named IHPT, as shown in Fig. 1. Specifically, IHPT first makes the background of the source image based on the text prompt by leveraging the visual fidelity of the large-scale pre-trained SD. Next, IHPT generates a new target image for a given target pose, while maintaining the texture, complexity, and quality of the background-added source hand image. Therefore, IHPT leads to maximizing semantic understanding of the source image by enhancing the generalizability of target hand image generation.

In the experiments, IHPT demonstrates the capability of hand image transfer with more physically plausible results. In particular, IHPT shows robust image generation, given any text prompts and target poses. Moreover, we additionally trained off-the-shelf 3D interacting hand mesh reconstruction network [23] with images generated by IHPT. As a result, the improvement of performance is verified on in-the-wild scenes. It implies generating diverse in-the-wild hand images with IHPT can alleviate the lack of in-the-wild 3D hand datasets and overcome domain gaps between indoor and outdoor environments, leading to make a positive contribution to downstream applications.

2 Method

We introduce a novel diffusion-based interacting hand pose transfer model, IHPT. As shown in Fig. 2, IHPT is composed of three modules: (1) Background Image Generator, (2) Source Image Generator, and (3) Target Image Generator.

2.1 Background Image Generator (BIG)

BIG is a module that generates a new background image I_{back} from a given text prompt T. It is designed based on Stable Diffusion [4], so that we can create high-quality and faithful images. Moreover, since I_{back} is utilized as the input of Source Image Generator to generate various source images, our model can contribute to the downstream task, such as 3D hand mesh reconstruction in the wild. BIG can be expressed as follows:

$$I_{back} = BIG(T). \tag{1}$$

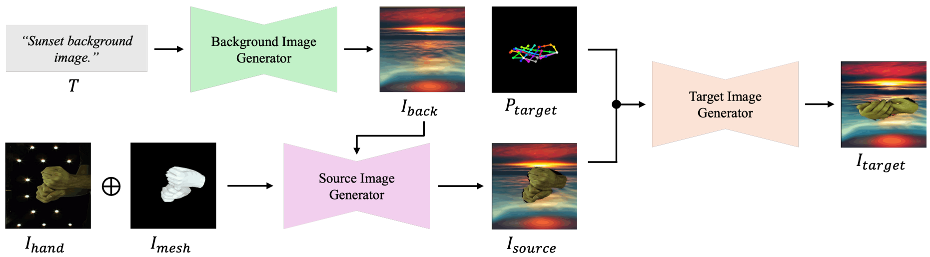

Fig. 2: The overall pipeline of IHPT. IHPT has three modules: Background Image Generator, Source Image Generator, and Target Image Generator.

2.2 Source Image Generator (SIG)

SIG is a module that generates a new source image I_{source} from the background image I_{back}, the hand image I_{hand}, and the corresponding mesh image I_{mesh}. Specifically, I_{mesh} can be obtained by projecting MANO [24]-based 3D label into the 2D image space. Hence, by thresholding the intensity of pixels of I_{mesh}, we can fill high-intensity pixels with I_{hand} and low-intensity pixels with I_{back} to make I_{source}. In other words, it is possible to generate I_{source} reflecting T and I_{hand}. SIG can be expressed as follows:

$$I_{source} = SIG(I_{back}, I_{hand}, I_{mesh}). \tag{2}$$

2.3 Target Image Generator (TIG)

TIG is a module that creates a generated target image based on the source image and the target pose. For the training phase, note that we notate variables with $\wedge$ (i.e., *hat*) for readability. First, the source image $\hat{I}_{source}$, the source pose $\hat{P}_{source}$, the ground-truth target image $\hat{I}_{target}$, and the target pose $\hat{P}_{target}$ are needed. Specifically, we extract the feature map $\hat{f}_{source}$ by passing $\hat{I}_{source}$ through the image backbone network. Next, we obtain the source feature embedding $\hat{e}_{source}$ by passing $\hat{f}_{source}$ through the feature encoder. In addition, $\hat{P}_{source}$ and $\hat{P}_{target}$ are passed through a pose encoder to obtain the pose embedding $\hat{e}_{pose}$. Moreover, $\hat{f}_{source}$ and $\hat{e}_{pose}$ are passed through the feature decoder to obtain the visual prompt $\hat{c}$ for the diffusion process. Additionally, the noisy latent $\hat{z}_t$ for the diffusion process can be obtained by adding noise $\hat{\epsilon}$ by timestep $\hat{t}$ to the image latent $\hat{z}_0$, which is obtained by passing $\hat{I}_{source}$ and $\hat{I}_{target}$ through the image encoder. As a result, the denoising network ϵ_θ is optimized as follows:

$$\mathcal{L} = \mathbb{E}_{\hat{z}_0, \hat{c}, \hat{e}_{source}, \hat{e}_{pose}, \epsilon, \hat{t}}[||\hat{\epsilon} - \epsilon_\theta(\hat{z}_{\hat{t}}, \hat{t}, \hat{c}, \hat{e}_{source}, \hat{e}_{pose})||_2^2]. \tag{3}$$

Next, for the inference phase, only I_{source} and P_{target} are needed. Similar to the training phase, I_{source} is passed through the image backbone network to extract the feature map f_{source}. f_{source} is passed through the feature encoder to obtain the feature embedding e_{source}, and P_{target} is passed through the pose encoder to obtain the pose embedding e_{pose}. Moreover, f_{source} and e_{pose} are passed through the feature decoder to obtain the visual prompt c for the diffusion process. In addition, the noisy latent z_t for the diffusion process can be obtained from the random Gaussian distribution $\mathcal{N}(0,1)$ by timestep t. Finally, with the trained denoising network ϵ_θ, we obtain the predicted noise ϵ as follows:

$$\epsilon = \epsilon_\theta(z_t, t, c, e_{source}, e_{pose}). \tag{4}$$

Therefore, the predicted image latent z_0 can be obtained based on ϵ and z_t, and a newly generated target image I_{target} can be obtained by passing it through the image decoder.

3 Experiments

Interacting Hand Pose Transfer. We adopted two popular interacting hand datasets: InterHand2.6M (IH2.6M) [25] and Re:InterHand (ReIH) [26]. We demonstrated the qualitative results as shown in Fig. 3. We can see that target images are generated robustly and plausiblly on multiple background images generated from diverse text prompts. In addition, hands in target images are well generated for random target poses without distortion. This tendency is revealed both on IH2.6M and ReIH. It implies that IHPT properly handles semantic information of the source image and complex geometric information of the target pose.

3D Hand Mesh Reconstruction. We adopted MSCOCO [27], which is a representative dataset of in-the-wild scenes. Hence, it is appropriate to evaluate the generalizability of images generated by IHPT. We trained an off-the-shelf 3D hand mesh reconstruction network [23] with the new data generated by IHPT. We demonstrated the qualitative results as shown in Fig. 4; the case of applying IHPT showed better performance than the case without applying it. It implies that in-the-wild hand images generated by IHPT positively contribute to the downstream task (i.e., 3D hand mesh reconstruction).

4 Conclusion

We presented IHPT, a diffusion-based model for interacting hand pose transfer. IHPT treated the hand pose transfer as a series of diffusion processes that progressively adjust the hand from the source image to match the target pose. Initially, IHPT created the background for the source image using text prompts, leveraging the high visual fidelity of a large-scale pre-trained Stable Diffusion. It then generated a new target image with the desired target hand pose while preserving the texture, complexity, and quality of the background-added source hand image. In our experiments, IHPT demonstrated its ability to produce more physically plausible transferred hand images. It showed strong and robust image generation capabilities, effectively handling various text prompts and target

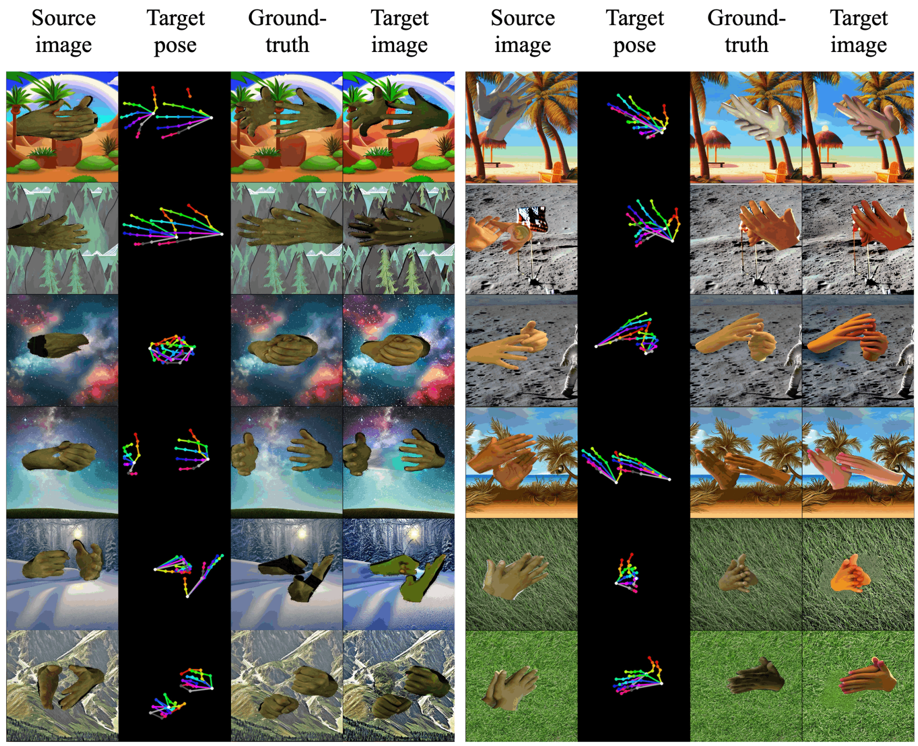

Fig. 3: Qualitative results for the hand pose transfer on IH2.6M [25] (left) and ReIH [26] (right).

Fig. 4: Qualitative results for the 3D hand mesh reconstruction on MSCOCO [27]. Red and green boxes indicate wrong and correct 3D hand mesh, respectively.

poses. Additionally, we trained an off-the-shelf 3D interacting hand mesh reconstruction network with IHPT-generated images and proved the improvement of performance in real-world scenarios. Therefore, utilizing IHPT to generate diverse hand images can help alleviate the shortage of in-the-wild 3D hand datasets and bridge the domain gaps between indoor and outdoor environments, thereby benefiting downstream applications.

References

1. Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and Krueger, Gretchen and Sutskever, Ilya. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021.
2. Nichol, Alex and Jun, Heewoo and Dhariwal, Prafulla and Mishkin, Pamela and Chen, Mark. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.
3. Kirillov, Alexander and Mintun, Eric and Ravi, Nikhila and Mao, Hanzi and Rolland, Chloe and Gustafson, Laura and Xiao, Tete and Whitehead, Spencer and Berg, Alexander C and Lo, Wan-Yen and Dollár, Piotr and Girshick, Ross. Segment anything. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4015–4026, 2023.
4. Rombach, Robin and Blattmann, Andreas and Lorenz, Dominik and Esser, Patrick and Ommer, Björn. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022.
5. Ye, Yufei and Li, Xueting and Gupta, Abhinav and De Mello, Shalini and Birchfield, Stan and Song, Jiaming and Tulsiani, Shubham and Liu, Sifei. Affordance diffusion: Synthesizing hand-object interactions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22479–22489, 2023.
6. Ye, Yufei and Hebbar, Poorvi and Gupta, Abhinav and Tulsiani, Shubham. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Int. Conf. Comput. Vis.*, pages 19717–19728, 2023.
7. Zhang, Mengqi and Fu, Yang and Ding, Zheng and Liu, Sifei and Tu, Zhuowen and Wang, Xiaolong. HOIDiffusion: Generating realistic 3D hand-object interaction data. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8521–8531, 2024.
8. Narasimhaswamy, Supreeth and Bhattacharya, Uttaran and Chen, Xiang and Dasgupta, Ishita and Mitra, Saayan and Hoai, Minh. Handiffuser: Text-to-image generation with realistic hand appearances. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2468–2479, 2024.
9. Lee, Jihyun and Saito, Shunsuke and Nam, Giljoo and Sung, Minhyuk and Kim, Tae-Kyun. InterHandGen: Two-hand interaction generation via cascaded reverse diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 527–537, 2024.
10. Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, and Chi-Wing Fu. HandBooster: Boosting 3D hand-mesh reconstruction by conditional synthesis and sampling of hand-object interactions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10159–10169, 2024.
11. Cheng, Wencan and Tang, Hao and Van Gool, Luc and Ko, Jong Hwan. HandDiff: 3D hand pose estimation with diffusion on image-point cloud. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2274–2284, 2024.

12. Cha, Junuk and Kim, Jihyeon and Yoon, Jae Shin and Baek, Seungryul. Text2HOI: Text-guided 3D motion generation for hand-object interaction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1577–1585, 2024.
13. Zhang, Wenqian and Huang, Molin and Zhou, Yuxuan and Zhang, Juze and Yu, Jingyi and Wang, Jingya and Xu, Lan. BOTH2Hands: Inferring 3D hands from both text prompts and body dynamics. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2393–2404, 2024.
14. Ye, Yufei and Gupta, Abhinav and Kitani, Kris and Tulsiani, Shubham. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1911–1920, 2024.
15. Li, Mengcheng and Zhang, Hongwen and Zhang, Yuxiang and Shao, Ruizhi and Yu, Tao and Liu, Yebin. HHMR: Holistic hand mesh recovery by enhancing the multimodal controllability of graph diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 645–654, 2024.
16. Park, Junho and Kong, Kyeongbo and Kang, Suk-Ju. AttentionHand: Text-driven Controllable Hand Image Generation for 3D Hand Reconstruction in the Wild. *Eur. Conf. Comput. Vis.*, 2024.
17. Bhunia, Ankan Kumar and Khan, Salman and Cholakkal, Hisham and Anwer, Rao Muhammad and Laaksonen, Jorma and Shah, Mubarak and Khan, Fahad Shahbaz. Person image synthesis via denoising diffusion model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5968–5976, 2023.
18. Han, Xiao and Zhu, Xiatian and Deng, Jiankang and Song, Yi-Zhe and Xiang, Tao. Controllable person image synthesis with pose-constrained latent diffusion. In *Int. Conf. Comput. Vis.*, pages 22768–22777, 2023.
19. Lu, Yanzuo and Zhang, Manlin and Ma, Andy J and Xie, Xiaohua and Lai, Jianhuang. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6420–6429, 2024.
20. Zhou, Xinyue and Yin, Mingyu and Chen, Xinyuan and Sun, Li and Gao, Changxin and Li, Qingli. Cross attention based style distribution for controllable person image synthesis. In *Eur. Conf. Comput. Vis.*, pages 161–178, 2022.
21. Ren, Yurui and Fan, Xiaoqing and Li, Ge and Liu, Shan and Li, Thomas H. Neural texture extraction and distribution for controllable person image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13535–13544, 2022.
22. Zhang, Pengze and Yang, Lingxiao and Lai, Jian-Huang and Xie, Xiaohua. Exploring dual-task correlation for pose guided person image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7713–7722, 2022.
23. Moon, Gyeongsik. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17028–17037, 2023.
24. Romero, Javier and Tzionas, Dimitris and Black, Michael J. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), 2017.
25. Moon, Gyeongsik and Yu, Shoou-I and Wen, He and Shiratori, Takaaki and Lee, Kyoung Mu. Interhand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *Eur. Conf. Comput. Vis.*, pages 548–564, 2020.
26. Moon, Gyeongsik and Saito, Shunsuke and Xu, Weipeng and Joshi, Rohan and Buffalini, Julia and Bellan, Harley and Rosen, Nicholas and Richardson, Jesse and Mize, Mallorie and De Bree, Philippe and Simon, Tomas and Peng, Bo and Garg, Shubham and McPhail, Kevyn and Shiratori, Takaaki. A dataset of relighted 3D interacting hands. *Adv. Neural Inform. Process. Syst.*, 36, 2023.

27. Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014.