Semi-supervised Hand Appearance Recovery via Structure Disentanglement and Dual Adversarial Discrimination

– Supplementary Material –

Zimeng Zhao Binghui Zuo Zhiyu Long Yangang Wang*

Southeast University, China

A. Overview

In this supplementary document, we first introduce the hand saliency network to assist our ViT sketcher's training in Sec. B. Then, we explain the way to construct the partner domain to assist our dual adversarial discrimination (DAD) scheme in Sec. C. After that, we describe the data augmentation strategies used in each piece of training in Sec. D. We further add more experimental results of our methods (Sec. E), as well as the discussions about its failure cases (Sec. F). They were not included in the main paper due to the page limit.

B. Hand Saliency Estimation

Utilization. Our estimator regresses the visible hand saliency $M(X) \in [0,1]^{(h,w)}$ from a hand-centered image $X \in \mathbb{R}^{(3,h,w)}$. Because S is also an image domain containing bare hand structure, the estimator is also compatible with regressing $M(S)$ from S. Compared with those generic instance segmentation approaches [2, 8], $M(X)$ may be imperfect, but it is sufficient to reduce the effect of the background. Furthermore, as shown in Fig. 6, $M[S(X)]$ is more valuable for our task, which retains the complete bare hand structure defined in S. In our framework, $M(X)$ plays the following roles: (i) $\boldsymbol{m}(X)^{\star} = \mathrm{MaxPool}(M(X), p)$ is used as a teacher to provide patch-wise saliency supervision during training of the sketcher's MLP. (ii) $M[S(X)]$ is used as a mask to compute MSE only for the background part during the training of the translator. (iii) $M[S(X)]$ is utilized as a mask to extract domain features only for the non-background part within DNN perceptual metrics.

Architecture. As illustrated in Fig. 1, a header composed of two parallel convolution layers is first used to extract features from input images. Subsequently, those two branches are concatenated and fed into an encoder. The encoder is composed of 5 residual blocks, each of them combined with 2D convolution layers and rectified linear unit functions. As a result, receptive fields with gradual enlargement are obtained. The decoder adopts a symmetric structure similar to the encoder that also consists of 5 stacked residual blocks but with up-sampling behind each block. For the first decoder layer, after scaling up by up-sampling, the feature map produced by the last encoder layer was concatenated with the later encoder layer. Similarly, other decoder layers do the same up-sampling and concatenate with the encoder output with the same resolution. Except for the last layer, leaky-ReLU is adopted for activation. Finally, a hand saliency with 256×256 is estimated.

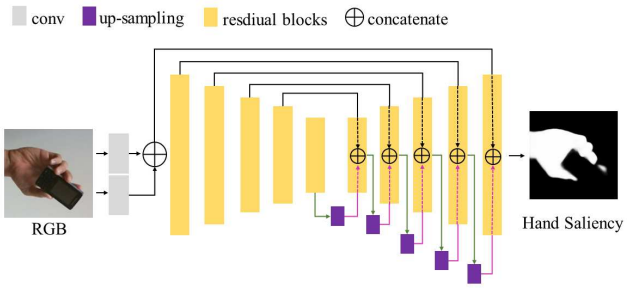

Figure 1. **Hand saliency estimation network**. We use an encoder-decoder architecture network to estimate hand saliency. We do the skip connection between the encoder and decoder with the same resolution.

C. Partner Domain Construction

The partner domain $\tilde{\mathbf{B}}$ provides a bridge for unpaired translation from source domain $\mathbf{A}$ to target domain $\mathbf{B}$ according to our DAD scheme. During the training of our

*Corresponding author. E-mail: yangangwang@seu.edu.cn. This work was supported in part by the National Natural Science Foundation of China (No. 62076061), in part by the Natural Science Foundation of Jiangsu Province (No. BK20220127).

translator, $X_{\tilde{\mathbf{B}}} \in \tilde{\mathbf{B}}$ is augmented by $X_{\mathbf{B}} \in \mathbf{B}$ as follows:

$$\begin{aligned} X_{\tilde{\mathbf{B}}} &= \tilde{N}(X_{\mathbf{B}}) \\ &= (1 - M[S(X)]) \odot X_{\mathbf{B}} + M[S(X)] \odot D(X_{\mathbf{B}}) \end{aligned} \tag{1}$$

where $M[S(X)]$ ensures that the degradations only occur in the hand region. $D(X_{\mathbf{B}})$ is used to simulate various markers, gloves, and objects. In practice, we create a variety of hand-specific degradations: (i) Spot degradations centered at visible hand joints. The locations are estimated by the off-the-shelf 2D key-point estimator [11]. (ii) Line degradations distributed along visible hand affinities [3], *i.e.* between adjacent visible joints. (iii) Region degradations approximated by randomly enclosing polygons in $M[S(X)]$; (iv) Whole degradation in $M[S(X)]$. The number and color of each degradation type are random. They are interpolated randomly with the original pixel value. Some examples are shown in Fig. 5.

D. Training Data Augmentation

Image domain. During the training of the MLP, discrete VAE$\{\mathcal{T}_s, \mathcal{F}_s\}$, attention decoder $\mathcal{E}_s$, and translator $\mathcal{G}$, the following data enhancement methods are adopted to augment the input RGB: (i) flip randomly up-down and left-right; (ii) rotate $\theta \in U(-\pi, \pi)$ randomly centered on the hand area; (iii) scale $s \in U(0.8, 1.2)$ randomly centered on the hand area; (iv) blur randomly with a kernel size $k \in U(3, 9)$.

Structure domain. The structure map represented as depth S_d or IUV S_{uv} used in ablations is augmented in the same way as the RGB image. For normal map S_n, because the pixel with the 2D coordinate (u, v) records the normal direction of the hand surface point in the camera coordinate system, it also changes after a flip or rotation augmentation:

$$S_n(u, v)' = R(\theta) \cdot S_n(u, v) \tag{2}$$

where θ is the accumulated angle from flip and rotation. Other forms of augmentation do not cause changes in the normal vector.

E. More Experiments

Human perceptual metrics. The human perceptual survey about the translation authenticity of the results from different frameworks is completed in Amazon Mechanical Turk (AMT). At the beginning of the questionnaire, participants were instructed to select one "real" candidate in each question that best matches the appearance of the bare hand and is most consistent with the semantics in the source image. We collect 2K questionnaires and use the percentage of each method's score divided by the total number of people (2K) as the evaluation of the translation quality of each method, with higher scores representing better results. As shown in Tab. 1, our framework obtains the majority of votes for best translating from both $\mathbf{A}_1 \rightarrow \mathbf{B}$ and $\mathbf{A}_2 \rightarrow \mathbf{B}$. It is worth noting that not all participants evaluated all six methods due to the random assignment process. However, the number of participants in each method evaluated ranged from 55% to 60% of the total number of participants, for this reason, our numbers may be different from the original baselines. The translation quality of CUT [7] is second, and several other methods [1, 5, 6, 12] based on cycle consistency have poorer quality in our task.

Tasks	$\mathbf{A}_1 \rightarrow \mathbf{B} \uparrow$	$\mathbf{A}_2 \rightarrow \mathbf{B} \uparrow$
CycleGAN [12]	$18.46\% \pm 0.9\%$	$22.59\% \pm 1.7\%$
GANerated [5]	$14.51\% \pm 2.3\%$	$17.47\% \pm 2.0\%$
H-GAN [6]	$11.12\% \pm 1.7\%$	$9.91\% \pm 2.8\%$
UAG [1]	$13.93\% \pm 0.4\%$	$12.59\% \pm 1.2\%$
CUT [7]	$21.26\% \pm 1.6\%$	$23.19\% \pm 3.2\%$
Ours	$\mathbf{28.16\% \pm 1.1\%}$	$\mathbf{32.37\% \pm 2.3\%}$

Table 1. **AMT perceptual evaluation.** AMT *real vs fake* test on $\mathbf{A}_1 \rightarrow \mathbf{B}$ and $\mathbf{A}_2 \rightarrow \mathbf{B}$.

Estimator	Openpose [3]		SRNet [11]	
Dataset Version	Original	Translated	Original	Translated
FPHAB [4]	0.81	0.85	0.86	0.91
FreiHand [13]	0.87	0.92	0.89	0.93

Table 2. **Hand pose estimation performance** on the original datasets and their appearance recovery version translated by our framework. PCK0.2 score is adopted as the accuracy criterion.

Improvement to pose estimation. We quantitatively evaluated the effect of recovering the bare appearance on the accuracy of hand pose estimation. Two different 2D pose estimators [3, 11] are adopted to estimate the accuracy of key-points localization before and after restoring the bare appearance for data from two datasets [4, 13]. As shown in Tab. 2, the accuracy of the pose estimators is generally improved on the datasets with recovered appearance. This is one of the most direct ways our framework can contribute to downstream tasks.

Universality of structure domain. In our framework, a bare structure prior defined on a standardized domain is built explicitly. Interestingly, in Splice-ViT [9], a similar domain is constructed in its implicit appearance wrapping process. As shown in Fig. 2, in the first few iterations, the translator in Splice-ViT tends to translate the structural reference to a uniform domain that contains only visible structure information. This finding indirectly confirms that the design of our framework to disentangle the bare structure is reasonable and efficient.

More recovery results. We show more qualitative results for our framework in hand appearance recovery from $\mathbf{A}_1 \rightarrow \mathbf{B}$ in Fig. 7 and $\mathbf{A}_2 \rightarrow \mathbf{B}$ in Fig. 8. The sampled hand region

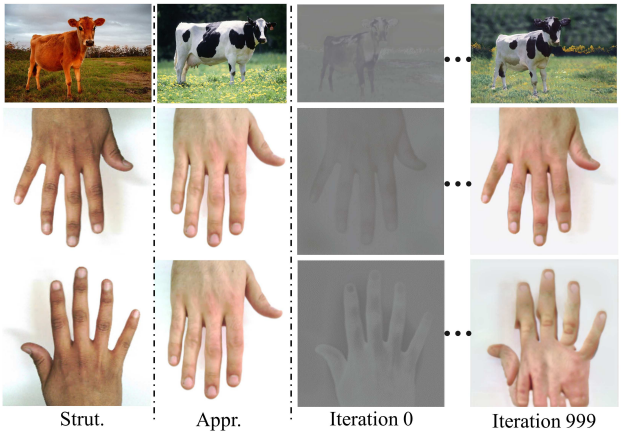

Figure 2. **Splice-ViT wrapping process**. From left to right: the structure reference, the appearance reference, the wrapping results in the first few iterations, and the final wrapping results.

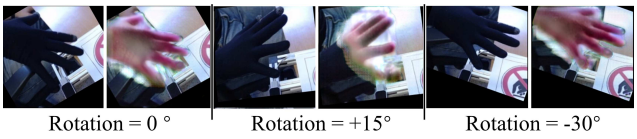

Figure 3. **Severe degradation cases**. When the input appearance is severely degraded, our model is not rotation-invariant.

$[M](X)$ and the disentangled structure map $S(X)$ are also comprehensively illustrated for each example.

An interesting phenomenon is that during the translation $\mathbf{A}_2 \rightarrow \mathbf{B}$, the object may be partially removed (the part that occludes the hand) or completely removed. This may be because we do not assign as much weight to the constrained background-consistent MSE loss in training as pix2pix [10]. In this way, the object part outside the bare hand region may also be penalized by the result discriminator $\mathcal{D}_{\mathbf{B}}^{(r)}$ in our DAD scheme, which makes the translator tend to erase them.

F. Failure Cases

As shown in Fig. 3, our framework is unstable when the input is severely degraded. In this case, hand features come entirely from the illusions of our CNN translator, which has anisotropic convolution kernels.

G. Acknowledgments

We thank Tiankai Hang for his constructive recommendations on the SOTA generation models. We thank Zesong Yang for his technical support on the maker-based hand MoCap system. We also thank Wei Xie, Wenqian Sun, Xingyou Liu, Jiaqi Qiao, and Zhi Li for their participation in our data collection and processing.

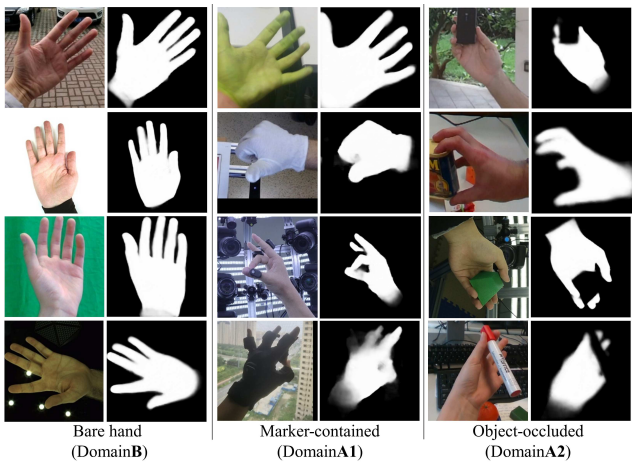

Figure 4. **Additional hand saliency results on different domains**. From left to right: hand saliency estimations on domain $\mathbf{B}$, hand saliency estimations on domain $\mathbf{A}_1$ and hand saliency estimations on domain $\mathbf{A}_2$.

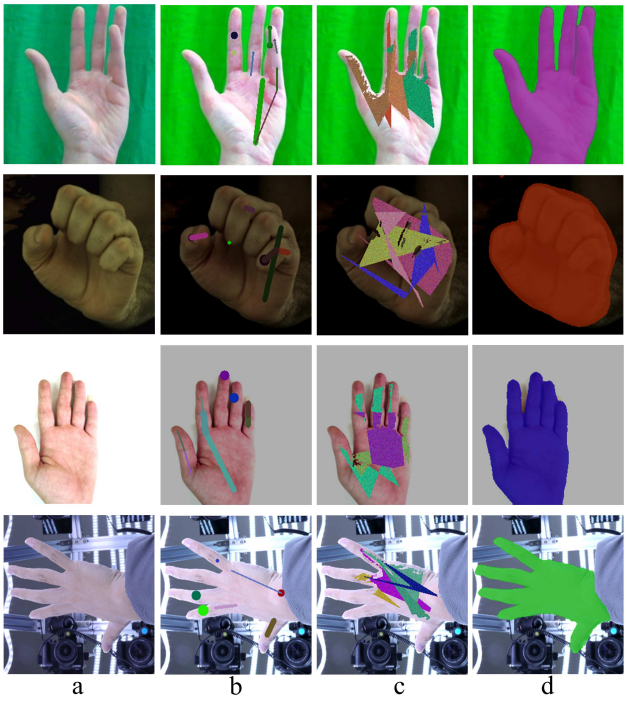

Figure 5. **Degradation process to obtain partner domain**. From left to right: **a**.Input images; **b**.Combined with spot and line degradations based on visible hand joints and affinities. **c**.Polygon-based region degradations. **d**.Mask-based whole degradations.

References

[1] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. *Advances in neural information processing systems*, 31, 2018. 2

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee.

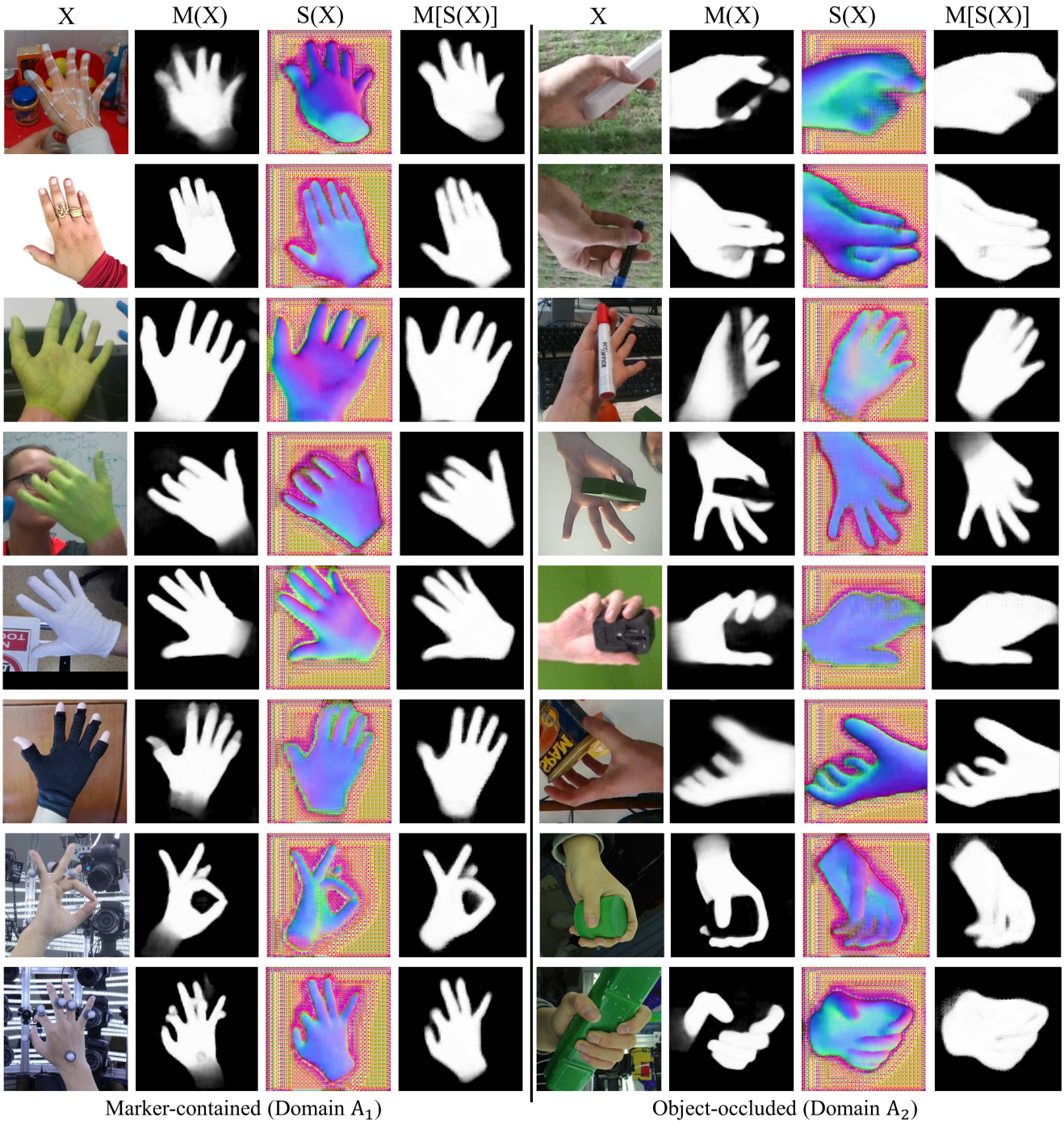

Figure 6. **Saliency Estimations from images and structure maps**. Our saliency estimator can both estimate visible hand saliency $M(X)$ from an RGB image X, or hand structure saliency $M[S(X)]$ from a structure map $S(X)$.

Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 1

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[4] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 2

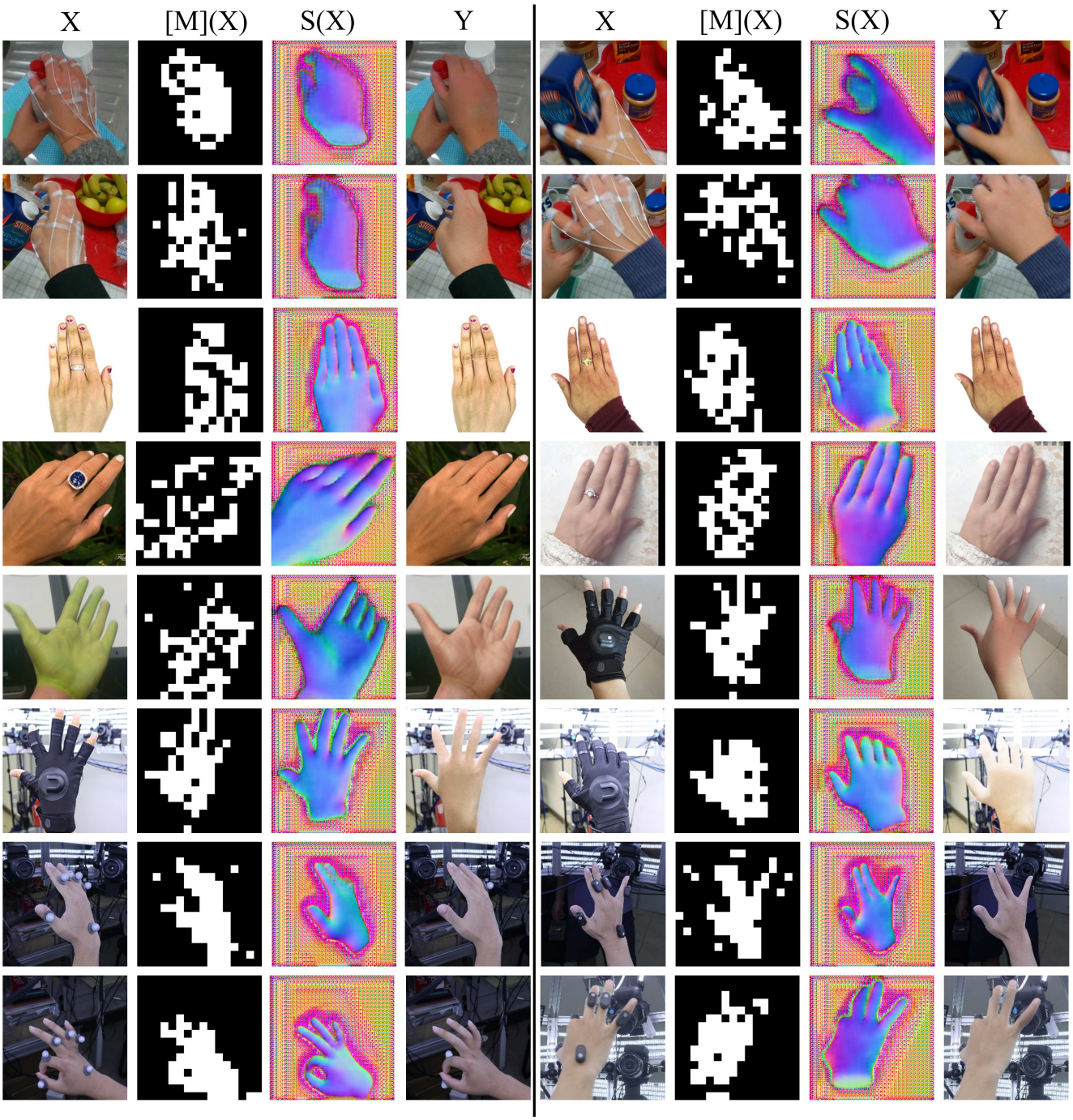

Figure 7. **Translation results of our framework on $\mathbf{A_1} \rightarrow \mathbf{B}$.** Two groups of results are presented in each row. Each group of results includes the input image X, the sampled mask $[M](X)$, the disentangled structure map $S(X)$ and the appearance wrapping result Y.

[5] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018. 2

[6] Sergiu Oprea, Giorgos Karvounas, Pablo Martinez-Gonzalez, Nikolaos Kyriazis, Sergio Orts-Escolano, Iason Oikonomidis, Alberto Garcia-Garcia, Aggeliki Tsoli, Jose Garcia-Rodriguez, and Antonis Argyros. H-gan: the power of gans in your hands. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2

[7] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-

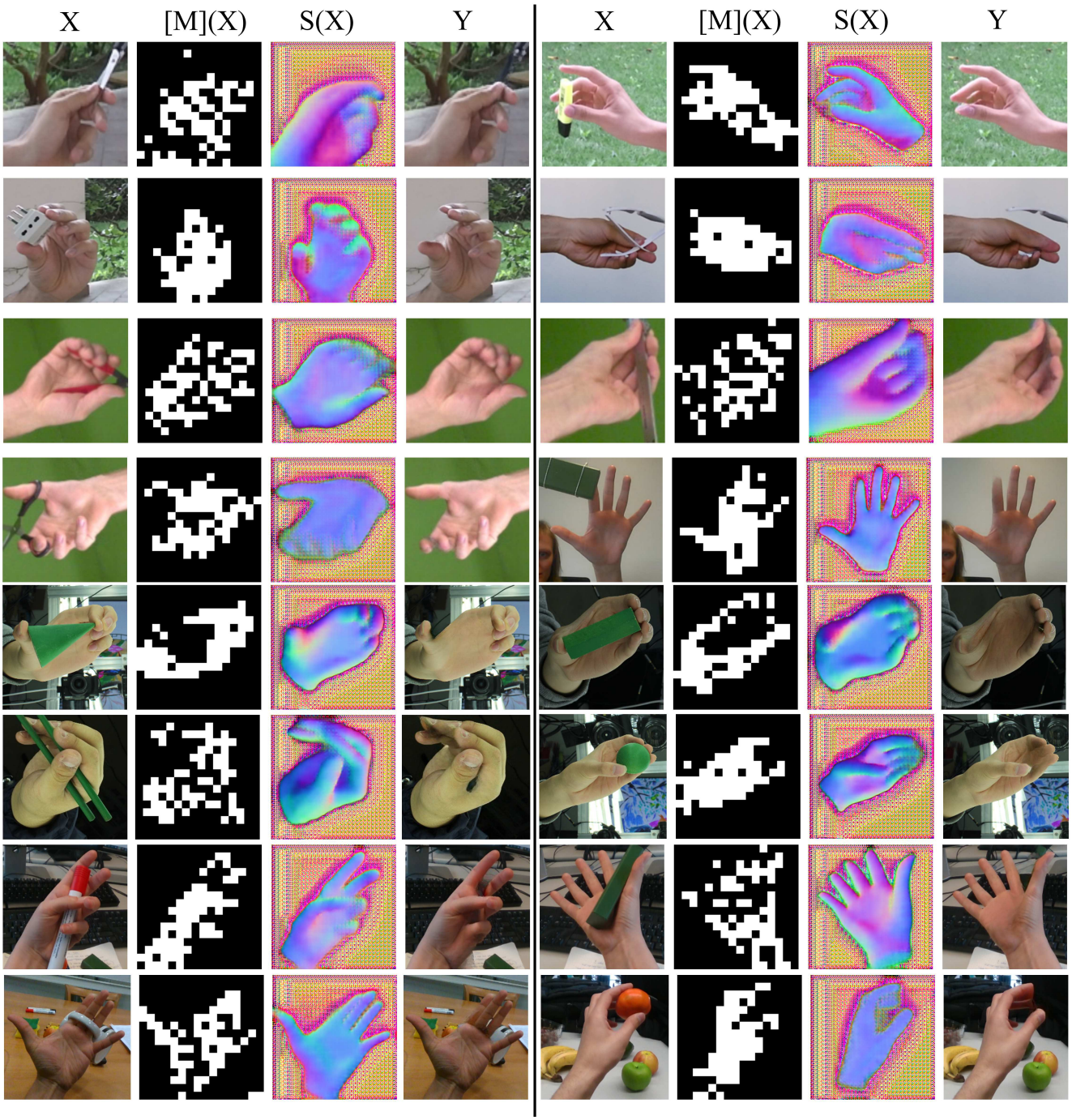

Figure 8. **Translation results of our framework on $\mathbf{A_2} \rightarrow \mathbf{B}$.** Two groups of results are presented in each row. Each group of results includes the input image X, the sampled mask $[M](X)$, the disentangled structure map $S(X)$ and the appearance wrapping result Y.

Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 2

[8] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2020. 1

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

[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao,

Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3

[11] Yangang Wang, Baowen Zhang, and Cong Peng. Srhandnet: Real-time 2d hand pose estimation with simultaneous region localization. *IEEE transactions on image processing*, 29:2977–2986, 2019. 2

[12] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[13] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2