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

Zimeng Zhao  Binghui Zuo  Zhiyu Long  Yangang Wang*
Southeast University, China

Abstract

Enormous hand images with reliable annotations are collected through marker-based MoCap. Unfortunately, degradations caused by markers limit their application in hand appearance reconstruction. A clear appearance recovery insight is an image-to-image translation trained with unpaired data. However, most frameworks fail because there exists structure inconsistency from a degraded hand to a bare one. The core of our approach is to first disentangle the bare hand structure from those degraded images and then wrap the appearance to this structure with a dual adversarial discrimination (DAD) scheme. Both modules take full advantage of the semi-supervised learning paradigm: The structure disentanglement benefits from the modeling ability of ViT, and the translator is enhanced by the dual discrimination on both translation processes and translation results. Comprehensive evaluations have been conducted to prove that our framework can robustly recover photo-realistic hand appearance from diverse marker-contained and even object-occluded datasets. It provides a novel avenue to acquire bare hand appearance data for other downstream learning problems.

1. Introduction

Both bare hand appearance and vivid hand motion are of great significance for virtual human creation. A dilemma hinders the synchronous acquisition of these two: accurate motion capture [20, 27, 68] relies on markers that degrade hand appearance, whereas detailed appearance capture [50, 59, 75] in a markerless setting makes hand motion hard to track. Is there a win-win solution that guarantees high fidelity for both?

Existing ones include markerless MoCap [26, 83, 88] and graphic rendering [16, 29, 80]. However, the former requires a pose estimator [13, 47, 90] trained with laborious annotations. And the latter often produces artifacts because it is hard to simulate photo-realistic lighting. Another insight is to “translate” the degraded appearances as bare ones end-to-end. Nevertheless, it is tough to collect paired data for its training. Moreover, most unsupervised frameworks [56, 57, 91] are only feasible when the translating target and source are consistent in structure, while our task needs to change those marker-related structures in the source. To this end, our key idea is to first disentangle the bare hand structure represented by a pixel-aligned map, and then wrap the appearance on this bare one trained with a dual adversarial discrimination (DAD) scheme.

There are two strategies to wrap the appearance from one image to another. (i) Template-based strategies learn [6, 63, 84, 96]...
Figure 2. Structure disentanglement from monocular RGBs. (Row-1) Input images. (Row-2) Mesh recovery by a template-based strategy [90]. (Row-3) Structure prediction by a template-free strategy [76]. (Row-4) Structure prediction by our sketcher w/o bare structure prior. (Row-5) Structure disentanglement by our full sketcher. Red circles indicate the artifacts in the results.

72] or optimize [2, 55] sophisticated wrappings based on parametric instance templates [59, 62]. However, the accurate estimation of those parameters is heavily influenced by the degraded appearance in the images (See Fig. 2 Row-2).

(ii) Template-free ones [40,73] excel at visible feature wrappings between structure-consistent images but are unable to selectively exclude marker-related features (See Fig. 2 Row-3 and Row-4). To address the problem, we first embed the bare hand structure prior into pixel-aligned maps. Then this prior is encoded as the token form [9], and a ViT [15] sketcher is trained to disentangle the corresponding structure tokens from partial image patches [30]. Interestingly, this ViT sketcher satisfies $S(S(X)) = S(X)$ [1], which means that when feeding its output as the input again, the two outputs should be consistent. We further utilize this elegant property to intensively train our sketcher in a semi-supervised paradigm.

Disappointingly, the recovered appearances remain unsatisfactory when a structure-assisted translator trained with existing adversarial paradigms: (i) In popular supervised paradigms [34, 76], the discriminator focuses on the quality of the translation process. (ii) In most unsupervised paradigms [5, 52, 56], the discriminator can only evaluate the translation result since there is no reliable reference for the translation process. Based on these two, we innovate the DAD scheme under a semi-supervised paradigm, which enables dual discrimination (both on the process and result) in our unpaired translation task. Initially, a partner domain is synthesized by degrading hand regions of the bare one. It possesses pairwise mapping relationships with the bare target domain, as well as similarity to the degraded source domain. During the translator training, data from the source and the partner domain are fed to the translator simultaneously. The two discriminators evaluate those translation processes and results with a clear division of labor. This scheme is more efficient than most unsupervised schemes [57, 91] because of those trustworthy pairs. It is more generalizable than a supervised scheme trained only with synthetic degradation [42, 43, 77] because of those multimodal inputs.

Our main contributions are summarized as follows.

• A semi-supervised framework that makes degraded images in marker-based MoCap regain bare appearance;
• A powerful ViT sketcher that disentangles bare hand structure without parametric model dependencies;
• An adversarial scheme that promotes the degraded-to-bare appearance wrapping effectively.

The codes will be publicly available at https://www.yangangwang.com.

2. Related Work

Hand data capture. Three procedures are widely used in hand capture: (i) Marker-based MoCap [17, 20, 27, 68,
70, 86] produces reliable motion but degraded appearance, so only the skeletal sequence is valuable for reconstruction. (ii) Skeletal appearance data [16, 29, 44, 80, 92] can be obtained from rendering digital hands [59, 62]. However, the synthetic-to-real gap still exists even with the most advanced CG technology. (iii) Markerless MoCap [51, 83, 88, 93] takes the goal to record motion without degrading appearance. It collects data in a multi-view stereo pipeline and performs learning-based pose estimation [13, 47, 79, 90] for each frame and each viewpoint. Although some weakly-supervised [3, 11, 41, 65, 66, 89] paradigms are being explored, the dependency on dataset diversity and expensive annotation are still irreplaceable to train a robust estimator [8]. Our framework recovers the marker-based MoCap data through image-to-image translation, which provides reliability for an estimator training.

**Human image synthesis.** The vision of a specific creature is highly discriminative of the authenticity of its own appearance. Thus, synthesizing photo-realistic human images is particularly challenging in both CV and CG communities. Differentiable rendering [2, 55] optimizes the whole render pipeline to minimize the difference between the result and the images. However, in our single-view task, estimating either ambient lighting or human skin texture is difficult. Generative adversarial networks (GANs) [22] yield appealing performances to synthesize or recover human face [14, 37, 60, 91]. This is primarily attributed to the fact [5, 6, 35] that the facial images could be modeled as a 2D manifold with minor spatial deformation and easily aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned. When tackling other synthesis tasks involving highly nonlinear variations, most methods synthesize appealingly aligned.

**3. Method**

There exist two unpaired domains from a translation perspective: the source domain A refers to hand images \{X_A\} with diverse appearance degradations, and the target domain B refers to \{X_B\} with bare appearance. Synthetic data is not considered in our task. Two steps are taken to recover the hand appearance in \(X_A \in \mathbb{R}^{(3,h,w)}\): First, a sketcher disentangles the bare structure map \(S(X_A)\) (Sec. 3.1). After that, a translator \(G\) learns to wrap \(S(X_A)\) with the appearance in \(X_A\) through our DAD scheme (Sec. 3.2). It is expected to generate \(Y_A\) by recovering the hand, reserving backgrounds, and removing degradations.

**3.1. Structure Disentanglement**

A desirable sketcher can disentangle the bare hand structure from images in both domains, so its input is no longer domain-distinct and generically denoted as \(X\) in this section. As shown in Fig. 4, the process is as follows:

\[
X \rightarrow t(X) \rightarrow t_{(M)}(X) \rightarrow t_s(X) \rightarrow S(X) \quad (1)
\]

**Saliency distillation.** \(X \rightarrow t(X) \in \mathbb{R}^{(n,d)}\) is executed in DINO-ViT [12] by extracting the visual tokens from \(n = (h/p) \cdot (w/p)\) non-overlapping image patches with uniform size \(p \times p\). Because it is powerful to depict visible structure [7, 73], we adopt it as a frozen backbone. A four-layer MLP behind the backbone is adopted to regress patch-wise hand saliency \(m(X) \in [0, 1]^n\). MLP is trained in a knowledge distillation scheme [19, 32] by regarding a well-trained hand saliency estimator as the teacher, who estimates \(M(X) \in [0, 1]^{(h,w)}\) (See Sup. Mat for details).

**Structure domain.** \(S(X) \in \mathbb{R}^{(c,h,w)}\) is defined under a standardized domain. Fig. 3 enumerates some candidates. The edge map is first excluded because it is too sparse. The other three can be acquired by rendering hand models [50, 62, 87], which frees us from the dependence on real images to construct the bare structure prior. Among them, values in a depth map may be ambiguous in our single-view task. The creation of an UV map additionally depends on the fixed UV unwrapping. By comparison, a normal map \(c=3\) circumvents most shortcomings. Based on it, we render a dataset \(S\) containing 100K diverse instances.

**Prior tokenization.** The selected domain that lacks an instance-specific prior is still too redundant. To make it more compact to represent a bare hand and also save the modeling capability of ViT, a discrete VAE [9, 61] \(\{T_s, F_s\}\)
is introduced to enable a map-token conversion \(S \mapsto t_s\). It is trained with the data \(S^* \in \mathbf{S}\) by the following loss:

\[
L_{\mathcal{T}_s, \mathcal{F}_s} = \lambda_8 L_{\text{MSE}}(\hat{\mathbf{t}}_s) + \|S - S^*\|_F + \lambda_9 L_{\text{ad}}[\hat{S}, S^*]
\]  

(2)

where \(\| \cdot \|_F\) is the Frobenius norm. The encoded tokens \(\hat{\mathbf{t}}_s = \mathcal{T}_s(S^*)\) are regulated as uniform distribution. The mean squared error (MSE) between the decoded map \(\hat{S} = \mathcal{F}_s(\hat{\mathbf{t}}_s)\) and \(S^*\) preserves the low-frequency details, while the adversarial term \(L_{\text{ad}}\) borrowed from [34] preserves its high-frequency details.

**Masked modeling.** Supervised by the well-trained tokenizer \(\mathcal{T}_s\), a naïve baseline is to perform \(t(X) \rightarrow t_s(X)\) with annotated data \(\{X, S^*_X\}\) in degradation-contained datasets [25, 88, 93]. However, due to their limited amount, such a process might harm the model’s generalization. We further recast it as masked image modeling (MIM) and introduce a mask-guided learning strategy. Instead of a random formulation [9, 30], we sample a fixed ratio \(\gamma\) of the tokens \([M](X)\) to be masked out according to a multinomial distribution related to the patch hand saliency \(m(X)\):

\[
[M](X) \sim \text{multinomial}[1 - m(X) + \varepsilon; \gamma]
\]  

(3)

where \(\varepsilon = 10^{-5}\) confirms the non-zero probability of all patches. As a result, \(t(X) \rightarrow t_{[M]}(X) \in \mathbb{R}^{(n,d)}\) is realized by replacing \((\gamma \cdot n)\) samples as the same learnable mask token.

After that, a ViT decoder \(\mathcal{E}_s\) learns the conversion of \(t_{[M]}(X) \rightarrow t_s(X)\). We first train it with \(\{X, S^*_X\}\) according to the following loss:

\[
L_{\mathcal{E}_s} = \|\mathcal{E}_s(t_{[M]}(X)) - t_s\|_F + \|\mathcal{E}_s(t_{[M]}(S^*_X)) - t^*_s\|_F
\]  

(4)

where \(t^*_s = \mathcal{T}_s(S^*_X)\). The second term is introduced because the standardized domain is a special image domain containing bare hand structure. This means that an expected structure map \(S(X)\) extracted from an image \(X\) should always be a fixed point [1] for our sketcher:

\[
S[S(X)] = S(X)
\]  

(5)

Based on Eqn. 5, our sketcher is further fine-tuned with those datasets \(\{X\}\) without available \(S^*_X\) as follows:

\[
L_{\mathcal{E}_s, \mathcal{F}_s} = \|S(X) - S[S(X)]\|_F
\]  

(6)

This semi-supervised paradigm significantly strengthens the collaboration between \(\mathcal{E}_s\) and \(\mathcal{F}_s\), which are always trained separately in the existing literature [9, 61]. To sum up, the whole sketcher disentangles the hand structure map by the following produce:

\[
S(X) = \mathcal{F}_s(\mathcal{E}_s[t_{[M]}(X)])
\]  

(7)

**Implementation details.** All pieces of training are optimized by Adam at a base learning rate of \(10^{-4}\) and a batch size of 16. We use DINO-ViT with \(p=16\) as the frozen backbone. We set images with size \(h=w=256\), the codebook with size 512, and \(t_s \in \mathbb{R}^{(512,16,16)}\). The backbone of \(\mathcal{T}_s\) is ResNet50, and \(\mathcal{F}\) is built symmetrically by transpose convolutions. \(L_{\text{ad}}\) is computed discriminatively by a multi-scale patchGAN [34]. The adversarial weight \(\lambda_9\) is set at 0.01 constantly. The KL weight \(\lambda_k\) is increased from 0 to \(6.6 \times 10^{-7}\) and Gumbel-SoftMax [36, 49] relaxes temperature \(\tau\) from 1.0 to 0.5 over the first 5000 updates. \(\mathcal{E}_s\) is built as 12 attention blocks with fixed sin-cos position embeddings [74]. The masking-out ratio is set to \(\gamma=0.75\).
3.2. Appearance Wrapping

The sketcher disentangles the bare structure \( S(X_A) \) containing both visible and degraded hand parts. Next, a wrapper learns to map valid appearance from \( X_A \) to \( S(X_A) \), achieving hand appearance recovery on its output \( Y_A \).

**Paradigm evolutions.** Before introducing our dual adversarial discrimination (DAD) in a semi-supervised paradigm, we progressively add related components. (i) The unsupervised paradigm [91] is shown in Fig. 5(a). Its adversarial scheme introduces result discriminator(s) \( D_B^{(r)} \), \( D_A^{(r)} \), which enables \( G \) and the inverse translator \( G^{-1} \) to form a bijection \( A \mapsto B \). (ii) The supervised one [34] is shown in Fig. 5(b). Its adversarial scheme has a process discriminator \( D_A^{(p)} \), which prompts \( G \) to learn more details based on computable MSE among paired data. Inspired by face restorations [42, 43, 77], we synthesize a partner domain \( A \) by degrading \( X_B \in B \) with diverse noise \( X_B = N(X_B) \) (See Sup. Mat for details). This paired degradation process \( B \rightarrow \bar{B} \) makes this supervised paradigm feasible.

**DAD scheme.** Our semi-supervised paradigm shown in Fig. 5(c) introduces \( D_B^{(r)}, D_B^{(p)} \) together to judge a multimodal translation \( (\bar{B}, A) \rightarrow B \) on both result qualities and process qualities. We also utilize the above-mentioned partner \( X_B = \bar{N}(X_B) \) and optimize \( G \) through the following loss:

\[
L_G = \|(Y_A-X_A) \odot (1-M[S(X_A)])\|_F + \|(Y_B-X_B) \odot (1-M[S(X_B)])\|_F + |D_B^{(r)}(Y_A) - 1| + |D_B^{(r)}(Y_B) - 1| + |D_B^{(p)}(X_A \oplus Y_A) - 1| + |D_B^{(p)}(X_B \oplus Y_B) - 1| \tag{8}
\]

where \( Y_A = G(X_A), Y_B = G(X_B) \), \( \oplus \) is channel-wise concatenation, \( \odot \) is element-wise product. The first two terms preserve the non-hand semantics outside \( M[S(X)] \). The next two encourage \( G \) to fool \( D_B^{(r)} \), and the last two to fool \( D_B^{(p)} \). Adversarially, the two discriminators are trained by:

\[
\begin{align*}
L_D^{(r)} &= |D_B^{(r)}(Y_A) - 0| + |D_B^{(r)}(Y_B) - \alpha_2| + |D_B^{(r)}(X_B) - 1| \\
L_D^{(p)} &= |D_B^{(p)}(X_A \oplus Y_A) - 0| + |D_B^{(p)}(X_B \oplus Y_B) - \alpha_1| + |D_B^{(p)}(X_B \oplus Y_B) - 1| \\
\end{align*}
\tag{9}
\]

where \( Y \) denotes stop-gradient. Two tolerances \( \alpha_1, \alpha_2 \) sampled uniformly from \( U(0.4, 0.7) \) are the scores of those plausible but synthesized translations.

**Translator Architecture.** As illustrated in Fig. 6, our translator takes \( X_A, S(X_A) \) as the inputs and separately extracts their multi-level features with a shared CNN backbone. After that, \( G \) gradually fuses the structure and appearance details at the same level with \( N \) wrappers based on the deepest features of \( S(X_A) \). The internal structure of the wrapper inherits from the synthesis layer in StyleGAN2 [38]. In each level, a max-pooling of the appearance feature provides the style to the modulated convolution, and a 1x1 convolution layer learns to filter the appearance and discard the degradation. The image mapping (“toRGB”) is added at each level, which makes the training more efficiently [38]. And the mapping at the final level outputs \( Y_A \).

**Implementation details.** The training is optimized by Adam at a base learning rate of \( 10^{-4} \) and a batch size of 16. The convolution backbone is selected as VGG-16 [64]. The image mapping at each level is supervised by the first two terms in Eqn. 8 in their resolutions. We inherit [40] to take VGG features in \( [1, 3, 5, 10, 13] \) layers. Consequently, \( N=5 \) wrappers are adopted in the following process.

4. Experiments

The baselines related to our appearance recovery task are enumerated in Sec. 4.1. The adopted data and metrics for
training and evaluations are then presented in Sec. 4.2. Different frameworks are compared in Sec. 4.3, and our key components are ablated in Sec. 4.4.

4.1. Baselines

Image translation is used to formulate our hand appearance recovery task. We make comparisons to (i) CycleGAN [91] and its successors, including GANerated [52], UAG [5] and H-GAN [56]. (ii) CUT [57] models the problem with contrastive strategies. (iii) Since an additional partner domain data \( \tilde{\mathbf{B}} \) is introduced in the DAD scheme, we also compare the translating performance trained in a full supervised paradigm [76] with paired data \( \tilde{\mathbf{B}} \leftrightarrow \mathbf{B} \). It is denoted as "Syn-Pix2pix".

Differentiable rendering optimizes parametric model [62], texture [59] and lighting in a rendering pipeline to minimize the visual differences between output and given image [2]. To make it executable, we manually make extra annotations in its testing: (i) left or right side, (ii) 2D key points for pose optimization, and (iii) silhouette for shape optimization. This pipeline is denoted as "Diff-Render".

Neural style transfer (NST) takes appearance and structure images as separated inputs [18]. For a fair comparison, we overlay \( S(X) \) to a testing image \( X \) as their structure reference (\( X \) provides extra background structure), and use the original \( X \) as their appearance reference. Three representative works are selected: WCT2 [82] is highly effective. STROTSS [40] utilizes VGG features as feedback. Splice-ViT [73] utilizes ViT features as feedback.

4.2. Datasets and Metrics

Structure data. To tokenize bare structure prior with \( \{T_s, F_s\} \), depth, IUV, or normal maps are obtained through graphic rendering. Depth map \( S_d \) and normal map \( S_n \) only need mesh data. IUV map \( S_{\text{uv}} \) uses the UV unwrapping in [59]. 100K samples for each domain are created using hand parametric models [62], pose archive [80,89] and publicly available scans [50]. The data used to train \( E_s \) contains both annotated datasets \( \{X, S^*_X\} \) (Freihand [93], HandStudio [88], HO3D [25]) and unlabeled ones \( \{X\} \) (Core50 [45], EgoDexter [54], DexterObject [67]).

Appearance data. Each of the following domains are represented by 33K image data (30K for training, 3K for evaluation). They are sampled from diverse datasets, and contain a comparable number of images in complex or monochrome backgrounds. All images are cropped to be hand-centered.
- Marker-contained domain (\( A_1 \)) covers a wide range of adhesion markers: FPHAB [17] (inertia), MHP [21] (gloves), PaintedHand [53] (paintings), and 11kHands [4] (rings).
- We further collect data containing optical markers [27].
- Object-occluded domain (\( A_2 \)) refers to the images in which the object occludes the hand, rather than any hand-object interacting images. They are sifted manually from Freihand [93], HandStudio [88], HO3D [25], Core50 [45], EgoDexter [54] and DexterObject [67].
- Bare hand domain (\( B \)) data is gathered from OneHand10k [78], DMTL [85], HandStudio [88], STB [84], InterHand2.6M [51] and GANerated [52].

Human perceptual metrics. The human perceptual survey is performed on Amazon Mechanical Turk (AMT): Each question contains degraded input and the recovering results from random frameworks in a row. Each Turker is assigned 30-40 questions. This survey is performed with 90 images.
Consequently, FID and KID based on the two features are denoted as (FID, KID) and (FID, KID), respectively. This choice also takes into account that our sketcher’s backbone is DINO-ViT-b16 (base architecture, patch size 16), so the evaluation with DINO-ViT-b8 leads to a fair comparison. In addition, all image backgrounds are filtered with $M[S(X)]$ before DNN feature extraction.

### 4.3. Comparisons

#### Qualitative comparisons
We compare with existing unsupervised frameworks in three aspects: (i) Full pipeline: The results in Fig. 1 are from their original paradigm+translator. (ii) Paradigm only: The results in Fig. 7 are from their paradigms+our $G$. (iii) Structure influence: The results in Fig. 9 are from their full pipeline w/ or w/o our $S(X)$ as an additional input. In (i) and (ii), they neither guarantee the hand bareness nor preserve the background during translation. In (iii), although their results are still not appealing enough, they do have a significant improvement in hand structure. For the sub-task of structure disentanglement, our prior-based structure sketcher is compared with a template-based pose estimator [90] and a template-free image translator [76]. The estimator needs to convert all hands in the image to left-handed as preprocessing. The translator relies on accurate hand saliencies to filter out non-hand features. More conveniently, our sketcher takes the image directly as input, and the internal ViT samples the patch adaptively. As shown in Fig. 2, our sketcher outperforms the pose estimator in hand part localization. It also ensures hand bareness to a greater extent while being robust to backgrounds and degradations. For the sub-task of appearance wrapping, we compare with Diff-Render and NSTs by feeding $X$ and $S(X)$ to them. Fig. 10 shows their performances on several testing samples. The synthesis-to-real gap is inevitable in Diff-Render results. Some skin tones and textures are aliased in NSTs.

#### Quantitative comparisons
Tab. 1 extensively illustrates the recovery qualities of the unsupervised frameworks and ours. FIDs and KIDs quantitatively evaluate results on the differences and variances to domain B in the latent space. We outperform other methods under both ViT- and CNN-based criteria, and the gap is more obvious in the criteria described by ViT. This reveals the rationality of modeling the degraded hand appearance with ViT. From Tab. 1, the performance of those mask-based methods [5, 52, 56] is underperformance as the incomplete hand structure occluded by degradations. Besides that, the methods [52, 56] used to transfer synthetic hands to real hands are also unsatisfactory for our task.

---

**Table 1. Evaluations for Translation.** The first column describes the translation framework. The next two column groups describe the performance of each framework on $A_1 \rightarrow B$ and $A_2 \rightarrow B$ respectively.
4.4. Ablation Study

We first ablate the key components in the bare structure disentanglement: (i) Which module contributes more to the overall performance? The quantitative of the two ablations are reported in Row-1 and Row-2 of Tab. 2. “w/o S(X) w/o DAD” is implemented by replacing DAD with unsupervised paradigms. The qualitative results in Fig. 7 reveal the superiority of our DAD scheme. The qualitative results in Fig. 9 show the guiding significance of S(X) to other translation process. “w/o S(X) w/ DAD” is implemented by feeding both inputs of our $G$ with duplicate X. The above experiments indicate that the presence of S(X) has a greater impact on our whole framework, and the design of DAD scheme make appearance wrapping more efficient. (ii) Is it necessary to construct a bare structure prior? As shown in the last two rows of Fig. 2, a standardized domain without prior is powerless and susceptible to the background. In addition, as illustrated in Fig. 2, CNN specializes in local details (Row-3), while ViT excels in cross-region associations (Row-4). (iii) Why choose a normal map to represent the standardized domain? We conduct variants with the other two easy-to-render candidates in Fig. 3. As shown in Row-3 and Row-4 of Tab. 2, their testing performances are weaker than a normal one. The drawbacks of the IUV map include dependency on a fixed mesh topology and dependency on a fixed mesh UV unwrapping. Consequently, compared to the normal map, the amount and diversity of available IUV data used to construct the bare structure prior become much smaller. (iii) How does the masking-out ratio affect the disentanglement performance? As shown in Fig. 11, a smaller one may introduce background distractions, while a larger one may inhibit the transmission of hand semantics. Coincidentally, $\gamma=0.75$ fits our task well. (iv) The ablations on loss terms used in the sketcher’s training are shown in Row-5 and Row-6 of Tab. 2. It can be found that semi-supervised fine-tuning significantly improves the performance of our sketcher. Row-7 and Row-8 of Tab. 2 illustrate the effectiveness of our translator architecture and DAD scheme separately.

5. Conclusion

This work pioneers a semi-supervised image-to-image translation to recover the hand appearance that was originally degraded during the marker-based MoCap process. Since this task also implies the degradation in hand structure, a prior-based sketcher is first proposed to disentangle the bare hand structure map from images. Later, an efficient adversarial scheme is devised to guide the translator to selectively wrap the appearance from the original image to the above structure map. This framework enables data from marker-based MoCap to regain complete and photorealistic hand appearance. It also provides a novel avenue to the dilemma in the simultaneous acquisition of hand appearance and motion data.

Limitations and future work. Our framework may become unstable when the input is severely degraded. Although only the single-hand case is verified, this prior-based method could also be adapted to multi-hand or body applications. In addition, improving it to tackle sequential problems could bring more benefits to the community.
References


[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 4, 8


[52] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and


[76] 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


