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.

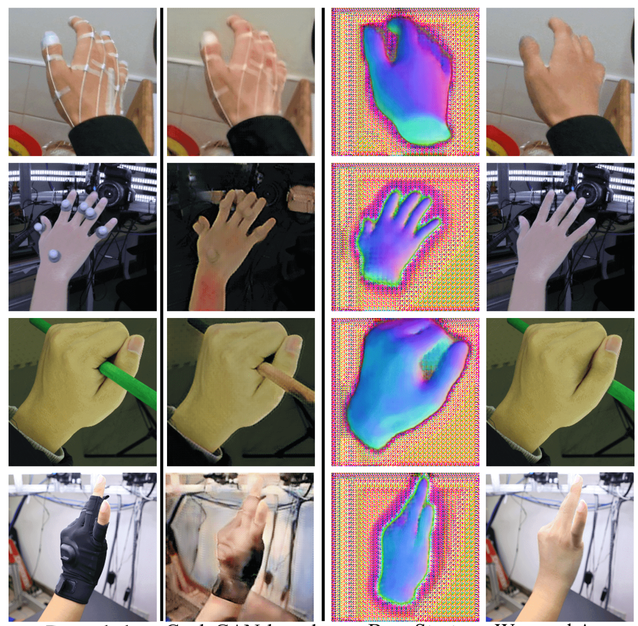

Figure 1. **Hand appearance recovery from diverse degradations**. Compared with CycleGAN-based frameworks, we recover more *bare* hand appearance while preserving more semantics.

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,

*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).

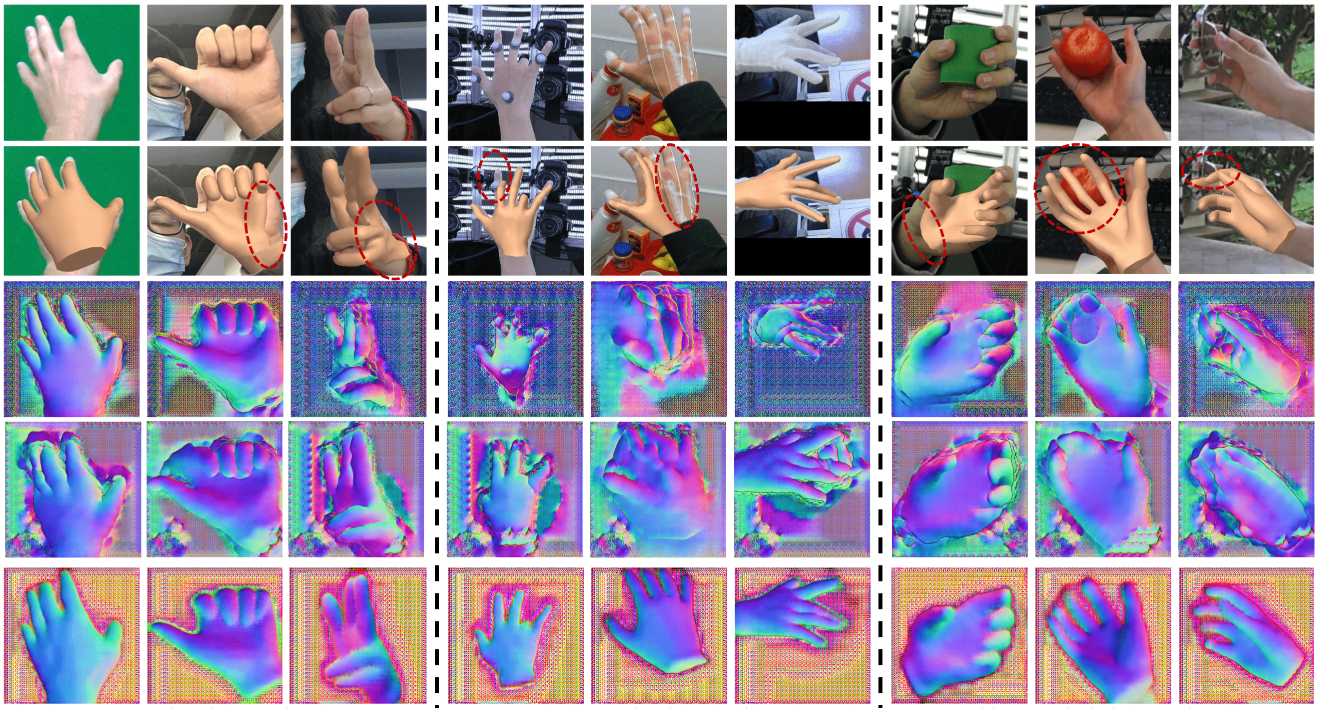

Figure 2. **Structure disentanglement from monocular RGBs**. (Row-1) Input images. (Row-2) Mesh recovery by a template-based strategy [90]. (Row-3) Structure prediction by a template-free strategy [76]. (Row-4) Structure prediction by our sketcher w/o/ bare structure prior. (Row-5) Structure disentanglement 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) Synthetic 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 a 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 given 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 appearance based on object-specific templates, *e.g.* joint hierarchy [28, 48, 63, 72], part segmentations [6, 23, 24] or surface topology [46]. Nevertheless, the presence of hand markers reduces the accuracy of estimating template parameters from images. We tackle this by embedding the bare hand structure prior in a pixel-aligned representation.

Unsupervised image-to-image translation. Our task regards appearance recovery as a pixel-aligned image translation without pixel-aligned paired data. Some pioneers [52, 56] applied cycle consistency [39, 81, 91] to synthetic-to-real adaptation. However, the cycle paradigm models entire images without the semantics distinction and is not suitable for our task requiring partial modification. Some improvements [5, 52, 71] regard adaptive saliencies as attentions to assist the generator in focusing on partial semantics. Nonetheless, since neither supervision nor penalty is adopted for these saliencies, this may harm the generator. Recently, the attention expressed by dot product in latent space [15, 74] has been widely adopted. Especially after a self-distillation pretraining [12], these attentions [15] show powerful semantics in correlations [7] and appearance modeling [73] without extra training or fine-tuning.

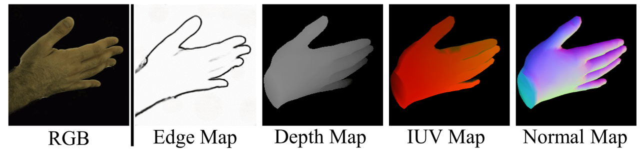

Figure 3. **Standardized domain candidates**. Edge map is estimated [58] from images. The others are rendered from meshes.

3. Method

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

$$X \rightarrow \boldsymbol{t}(X) \rightarrow \boldsymbol{t}_{[M]}(X) \rightarrow \boldsymbol{t}_s(X) \rightarrow S(X) \qquad (1)$$

Saliency distillation. $X \rightarrow \boldsymbol{t}(X) \in \mathbb{R}^{(n,d)}$ 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 $\boldsymbol{m}(X) \in [0,1]^n$. 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 IUV 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 $\mathbf{S}$ 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] $\{\mathcal{T}_s, \mathcal{F}_s\}$

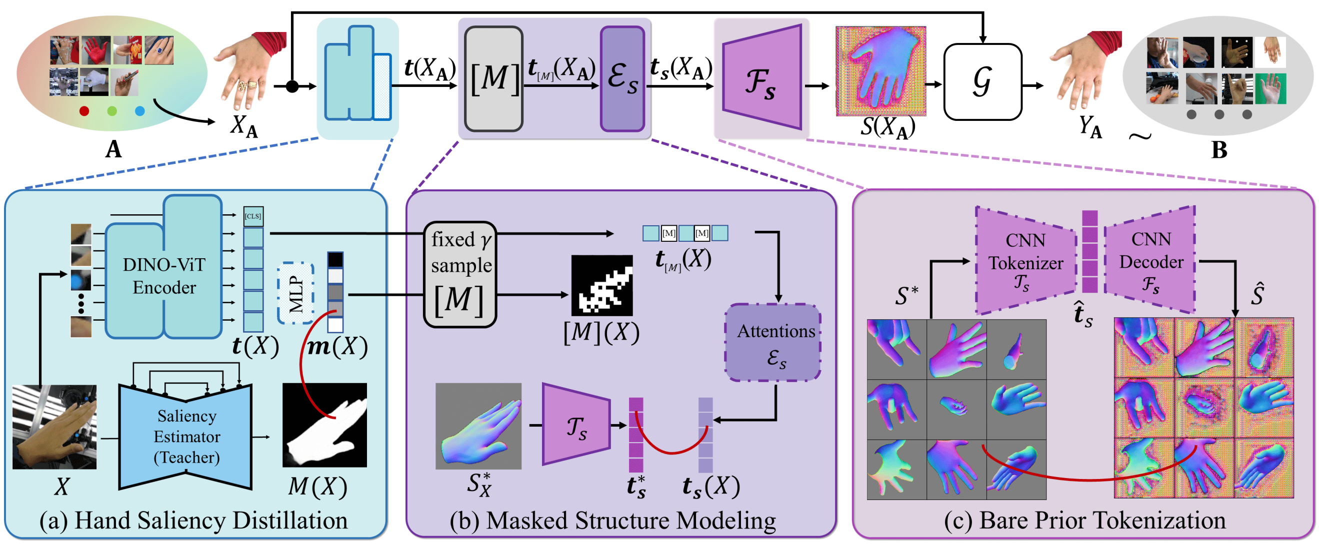

Figure 4. **Learning modules to structure disentanglement (better viewed in color)**. (a) *Hand saliency distillation* gives the sketcher backbone extra attention for the hand. (b) *Masked structure modeling* facilitates the sketcher's robustness to diverse degradations on hand structure. (c) *Bare Prior Tokenization* makes the prior of the bare hand structure more compact in representation. In the training step of each module, the blocks in dashed are with learnable weights. The red curves illustrate the supervisions of training.

is introduced to enable a map-token conversion $S \leftrightarrows \boldsymbol{t}_s$. It is trained with the data $S^\star \in \mathbf{S}$ by the following loss:

$$L_{\mathcal{T}_s,\mathcal{F}_s} = \lambda_k L_{\mathrm{kl}}(\hat{\boldsymbol{t}}_s) + \|\hat{S} - S^\star\|_F + \lambda_g L_{\mathrm{ad}}[\hat{S}, S^\star] \quad (2)$$

where $\|\cdot\|_F$ is the Frobenius norm. The encoded tokens $\hat{\boldsymbol{t}}_s$=$\mathcal{T}_s(S^\star)$ are regulated as uniform distribution. The mean squared error (MSE) between the decoded map $\hat{S}$=$\mathcal{F}_s(\hat{\boldsymbol{t}}_s)$ and $S^\star$ preserves the low-frequency details, while the adversarial term L_{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 $\boldsymbol{t}(X) \to \boldsymbol{t}_s(X)$ with annotated data $\{X, S^\star_X\}$ in degradation-contained datasets [25, 88, 93]. However, due to their limited amount, such a process may 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 γ of the tokens $[M](X)$ to be masked out according to a multinomial distribution related to the patch hand saliency $\boldsymbol{m}(X)$:

$$[M](X) \sim \mathrm{multinomial}[1 - \boldsymbol{m}(X) + \varepsilon; \gamma] \quad (3)$$

where ε=10^{-5} confirms the nor-zero probability of all patches. As a result, $\boldsymbol{t}(X) \to \boldsymbol{t}_{[M]}(X) \in \mathbb{R}^{(n,d)}$ 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 $\boldsymbol{t}_{[M]}(X) \to \boldsymbol{t}_s(X)$. We first train it with $\{X, S^\star_X\}$ according to the following loss:

$$L_{\mathcal{E}_s} = \|\mathcal{E}_s(\boldsymbol{t}_{[M]}(X)) - \boldsymbol{t}^\star_s\|_F + \|\mathcal{E}_s(\boldsymbol{t}_{[M]}(S^\star_X)) - \boldsymbol{t}^\star_s\|_F \quad (4)$$

where $\boldsymbol{t}^\star_s = \mathcal{T}_s(S^\star_X)$. 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 $\overline{S(X)}$ extracted from an image X should always be a fixed point [1] for our sketcher:

$$S[\overline{S(X)}] = \overline{S(X)} \quad (5)$$

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

$$L_{\mathcal{E}_s,\mathcal{F}_s} = \|S(X) - S[S(X)]\|_F \quad (6)$$

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:

$$S(X) = \mathcal{F}_s\big(\mathcal{E}_s[\boldsymbol{t}_{[M]}(X)]\big) \quad (7)$$

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 $\boldsymbol{t}_s \in \mathbb{R}^{(512,16,16)}$. The backbone of $\mathcal{T}_s$ is ResNet50, and $\mathcal{F}$ is built symmetrically by transpose convolutions. L_{ad} is computed discriminatively by a multi-scale patchGAN [34]. The adversarial weight λ_g is set at 0.01 constantly. The KL weight λ_k is increased from 0 to 6.6×10^{-7} and Gumbel-SoftMax [36, 49] relaxes temperature τ 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 γ=0.75.

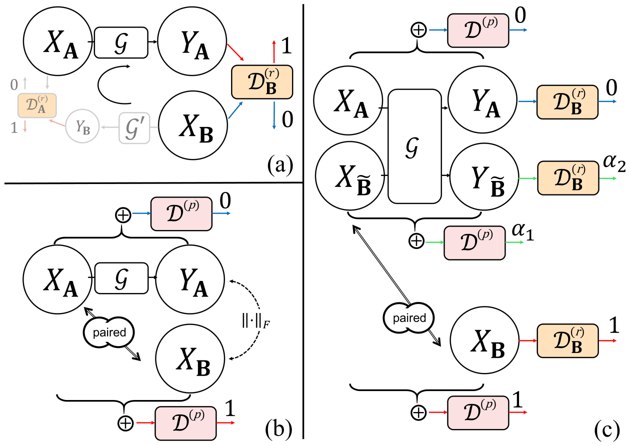

Figure 5. **Learning paradigm comparisons.** (a) CycleGAN [91] in an unsupervised paradigm. (b) Pix2pix [34] in a supervised paradigm. (c) Our DAD in a semi-supervised paradigm. In each sub-figure, blocks of the same color share weights.

3.2. Appearance Wrapping

The sketcher disentangles the bare structure $S(X_\mathbf{A})$ containing both visible and degraded hand parts. Next, a wrapper learns to map valid appearance from $X_\mathbf{A}$ to $S(X_\mathbf{A})$, achieving hand appearance recovery on its output $Y_\mathbf{A}$.

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) $\mathcal{D}^{(r)}_\mathbf{B}, \mathcal{D}^{(r)}_\mathbf{A}$, which enables $\mathcal{G}$ and the inverse translator $\mathcal{G}'$ to form a bijection$\mathbf{A} \leftrightarrows \mathbf{B}$. (ii) The supervised one [34] is shown in Fig. 5(b). Its adversarial scheme has a *process discriminator* $\mathcal{D}^{(p)}$, which prompts $\mathcal{G}$ to learn more details based on computable MSE among paired data. Inspired by face restorations [42, 43, 77], we synthesize a partner domain $\tilde{\mathbf{B}}$ by degrading $X_\mathbf{B} \in \mathbf{B}$ with diverse noise $X_{\tilde{\mathbf{B}}}$=$\tilde{N}(X_\mathbf{B})$ (See Sup. Mat for details). This paired degradation process $\mathbf{B} \to \tilde{\mathbf{B}}$ makes this supervised paradigm feasible.

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

$$
\begin{aligned}
L_\mathcal{G} = &\|(Y_\mathbf{A}\text{-}X_A) \odot (1\text{-}M[S(X_\mathbf{A})])\|_F \\
&+ \|(Y_{\tilde{\mathbf{B}}}\text{-}X_{\tilde{\mathbf{B}}}) \odot (1\text{-}M[S(X_{\tilde{\mathbf{B}}})])\|_F \\
&+ |\mathcal{D}^{(r)}_\mathbf{B}(Y_\mathbf{A})\text{-}1| + |\mathcal{D}^{(r)}_\mathbf{B}(Y_{\tilde{\mathbf{B}}})\text{-}1| \\
&+ |\mathcal{D}^{(p)}(X_\mathbf{A} \oplus Y_\mathbf{A})\text{-}1| + |\mathcal{D}^{(p)}(X_{\tilde{\mathbf{B}}} \oplus Y_{\tilde{\mathbf{B}}})\text{-}1|
\end{aligned}
\tag{8}
$$

where $Y_\mathbf{A}$=$\mathcal{G}(X_\mathbf{A}), Y_{\tilde{\mathbf{B}}}$=$\mathcal{G}(X_{\tilde{\mathbf{B}}})$, $\oplus$ 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 $\mathcal{G}$ to fool $\mathcal{D}^{(r)}_\mathbf{B}$, and the last two to fool $\mathcal{D}^{(p)}$. Adversarially, the two discriminators are trained by:

$$
\begin{cases}
L^{(r)}_\mathcal{D} = |\mathcal{D}^{(r)}_\mathbf{B}(\underline{Y_\mathbf{A}})\text{-}0| + |\mathcal{D}^{(r)}_\mathbf{B}(\underline{Y_{\tilde{\mathbf{B}}}})\text{-}\alpha_2| + |\mathcal{D}^{(r)}_\mathbf{B}(X_\mathbf{B})\text{-}1| \\
L^{(p)}_\mathcal{D} = |\mathcal{D}^{(p)}(X_\mathbf{A} \oplus \underline{Y_\mathbf{A}})\text{-}0| + |\mathcal{D}^{(p)}(X_{\tilde{\mathbf{B}}} \oplus \underline{Y_{\tilde{\mathbf{B}}}})\text{-}\alpha_1| \\
\qquad + |\mathcal{D}^{(p)}(X_{\tilde{\mathbf{B}}} \oplus X_\mathbf{B})\text{-}1|
\end{cases}
\tag{9}
$$

where $\underline{Y}$ denotes stop-gradient. Two tolerances α_1, α_2 sampled uniformly from $U(0.4, 0.7)$ are the scores of those plausible but synthesized translations.

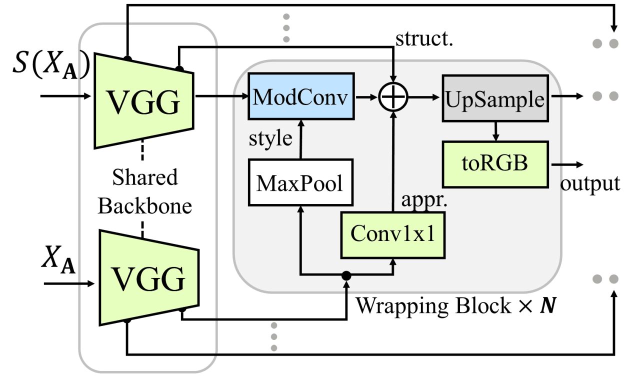

Figure 6. **Translator architecture**. $\boxed{\text{ModConv}}$ is the modulated convolution [38]. $\boxed{\text{Conv1x1}}$ is the convolution with a kernel size of 1. $\boxed{\text{toRGB}}$ is the convolution that converts the multi-channel feature to the three-channel feature with a kernel size of 1.

Translator Architecture. As illustrated in Fig. 6, our translator takes $X_\mathbf{A}, S(X_\mathbf{A})$ as the inputs and separately extracts their multi-level features with a shared CNN backbone. After that, $\mathcal{G}$ gradually fuses the structure and appearance details at the same level with N wrappers based on the deepest features of $S(X_\mathbf{A})$. 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 $Y_\mathbf{A}$.

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

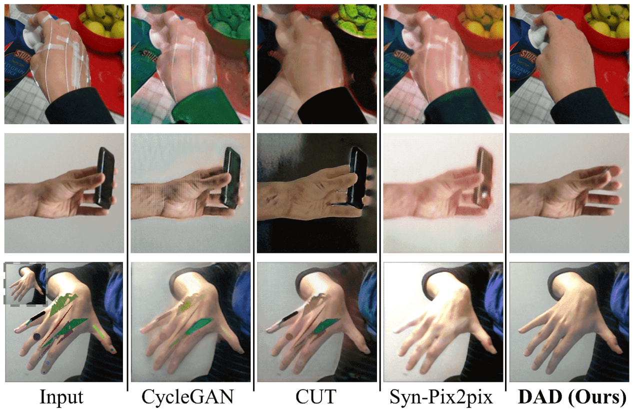

Figure 7. **Comparisons on learning paradigms**. From left to right: The input, CycleGAN [91], CUT [57], Syn-Pix2pix [76], and Ours. The testing samples are from domain $\mathbf{A}_1$, $\mathbf{A}_2$ and $\tilde{\mathbf{B}}$. The original version of the sample from $\tilde{\mathbf{B}}$ is on its top left.

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 $\tilde{\mathbf{B}} \leftrightarrows \mathbf{B}$. 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. SpliceViT [73] utilizes ViT features as feedback.

4.2. Datasets and Metrics

Structure data. To tokenize bare structure prior with $\{\mathcal{T}_s, \mathcal{F}_s\}$, 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_{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 $\mathcal{E}_s$ contains both annotated datasets $\{X, S_X^\star\}$ (Freihand [93], HandStudio [88], HO3D [25]) and unlabeled ones $\{X\}$(Core50 [45], EgoDexter [54], DexterObject [67]).

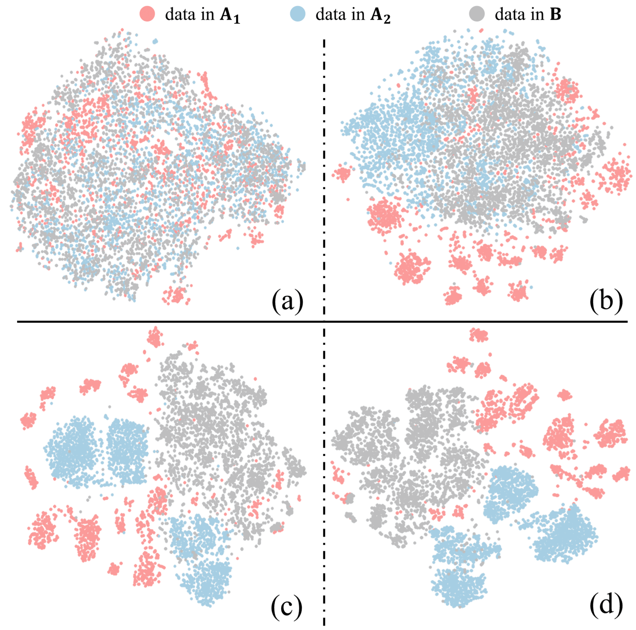

Figure 8. **Feature distributions of domain datasets (better viewed in color)**. (a) InceptionV3 pre-aux features(dim=768). (b) InceptionV3 final features (dim=2048). (c) DINO-ViT-b16 class tokens (dim=768). (d) DINO-ViT-b8 class tokens (dim=768).

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 ($\mathbf{A}_1$) 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 ($\mathbf{A}_2$) 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 ($\mathbf{B}$) 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

Tasks	$\mathbf{A}_1 \rightarrow \mathbf{B}$				$\mathbf{A}_2 \rightarrow \mathbf{B}$			
Metrics	$\text{FID}_\text{i} \downarrow$	$\text{KID}_\text{i}(*100) \downarrow$	$\text{FID}_\text{v} \downarrow$	$\text{KID}_\text{v} \downarrow$	$\text{FID}_\text{i} \downarrow$	$\text{KID}_\text{i}(*100) \downarrow$	$\text{FID}_\text{v} \downarrow$	$\text{KID}_\text{v} \downarrow$
CycleGAN [91]	76.39	4.46 ± 0.176	1266.17	32.02 ± 0.994	65.12	4.44 ± 0.196	1021.14	27.13 ± 0.948
GANerated [52]	76.97	4.72 ± 0.153	1220.53	31.72 ± 0.907	68.50	5.05 ± 0.199	985.82	26.75 ± 0.987
H-GAN [56]	93.02	6.53 ± 0.209	1488.32	37.85 ± 1.061	62.94	4.12 ± 0.197	876.94	24.61 ± 0.816
UAG [5]	87.80	5.90 ± 0.177	1375.55	35.67 ± 0.949	70.98	5.35 ± 0.210	1069.45	28.31 ± 0.926
CUT [57]	78.02	5.54 ± 0.192	1230.67	33.34 ± 1.015	58.88	3.73 ± 0.160	749.22	20.02 ± 0.826
Ours	**60.37**	$\mathbf{3.45 \pm 0.236}$	**994.67**	$\mathbf{28.67 \pm 0.916}$	**41.53**	$\mathbf{3.37 \pm 0.154}$	**673.43**	$\mathbf{15.72 \pm 1.209}$

Table 1. **Evaluations for Translation.** The first column describes the translation framework. The next two column groups describe the performance of each framework on $\mathbf{A}_1 \rightarrow \mathbf{B}$ and $\mathbf{A}_2 \rightarrow \mathbf{B}$ respectively.

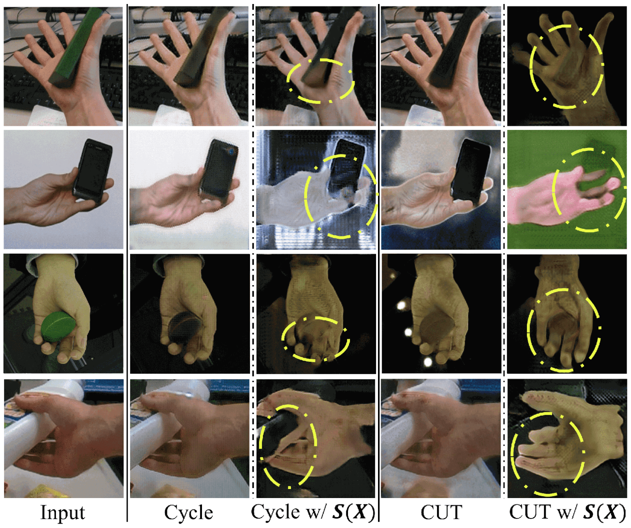

Figure 9. **Influence of the disentangled structure**. From left to right: The input, CycleGAN [91] w/ or w/o/ $S(X)$ assistance, and CUT [91] w/ or w/o/ $S(X)$ assistance. The yellow circles indicate the improvements in each group brought by our bare structure.

gathered from the testing samples in degraded domains. They are reported in our Sup. Mat .

DNN perceptual metrics. Frechet Inception Distance [31] (FID) and Kernel Inception Distance [10] (KID) measure the difference between translated results and target domain. Because they evaluate images in a latent space, an image feature extractor is required. Considering that CNN and ViT have a different emphasis, both InceptionV3 [69] and DINO-ViT are adopted. We explore their distinguishability on our data domains through t-SNE [33]. As shown in Fig. 8, InceptionV3 final features and DINO-ViT-b8 class tokens describe the cross-domain relationships more clearly. Consequently, FID and KID based on the two features are denoted as $(\text{FID}_\text{i}, \text{KID}_\text{i})$ and $(\text{FID}_\text{v}, \text{KID}_\text{v})$, 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 $\mathcal{G}$. (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 $\mathbf{B}$ 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.

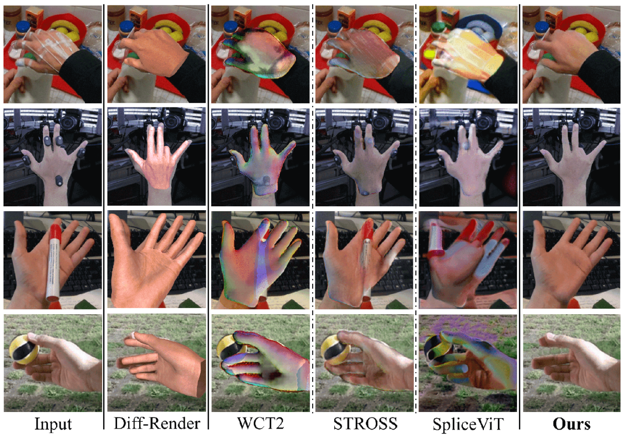

Figure 10. **Comparisons to the SOTA NSTs**. From left to right: The input, Diff-Render, WCT2 [82], STROTSS [40], Splice-ViT [73], and Ours. The first two rows are samples from $\mathbf{A}_1$ and the last two rows from $\mathbf{A}_2$.

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/ $S(X)$ 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 $\mathcal{G}$ 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, γ=0.75 [30] fits our task well. (iv) The ablations on loss terms used in the sketcher's

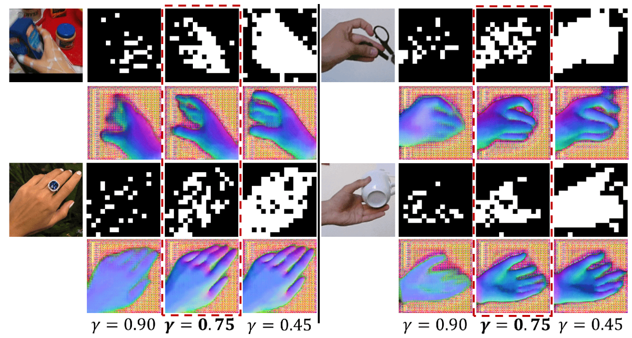

Figure 11. **Influence of γ settings**. Testing results of the sketcher trained with different sampling masking-out ratios.

Tasks	$\mathbf{A}_1 \to \mathbf{B}$		$\mathbf{A}_2 \to \mathbf{B}$	
Variants	FID$_v$	KID$_v$	FID$_v$	KID$_v$
w/ $S(X)$ w/o DAD	1213.29	30.48 ± 1.201	724.31	16.40 ± 0.751
w/o $S(X)$ w/ DAD	1259.33	35.06 ± 1.409	753.18	20.37 ± 1.016
w/ S_{uv} w/ DAD	1235.37	31.49 ± 1.378	732.90	16.68 ± 1.206
w/ S_d w/ DAD	1248.29	33.82 ± 1.392	750.95	19.01 ± 1.319
w/o L_{ad}(Eqn. 2)	1172.12	32.07 ± 1.102	712.15	16.73 ± 1.032
w/o $L_{\mathcal{E}_s,\mathcal{F}_s}$(Eqn. 6)	1161.14	30.61 ± 1.135	694.73	16.32 ± 0.893
w/ U-net $\mathcal{G}$	1174.90	29.87 ± 1.201	685.50	16.07 ± 1.034
w/ Syn-Pix2pix	1080.33	29.09 ± 1.195	680.32	15.99 ± 1.132
Full	**994.67**	$\mathbf{28.67 \pm 0.916}$	**673.43**	$\mathbf{15.72 \pm 1.209}$

Table 2. **Ablation studies on $\mathbf{A_1} \to \mathbf{B}$ and $\mathbf{A_2} \to \mathbf{B}$.** The first column describes the ablation item, and the other columns show the impact of the ablation on the final translation result.

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 photo-realistic 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

[1] Fixed point definition. https://en.wikipedia.org/wiki/Fixed_point_(mathematics). 2, 4

[2] Pytorch 3d. https://pytorch3d.org/. 2, 3, 6

[3] Masoud Abdi, Ehsan Abbasnejad, Chee Peng Lim, and Saeid Nahavandi. 3d hand pose estimation using simulation and partial-supervision with a shared latent space. *arXiv preprint arXiv:1807.05380*, 2018. 3

[4] Mahmoud Afifi. 11k hands: Gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 78(15):20835–20854, 2019. 6

[5] 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, 3, 6, 7

[6] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 1, 3

[7] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 3

[8] Anil Armagan, Guillermo Garcia-Hernando, Seungryul Baek, Shreyas Hampali, Mahdi Rad, Zhaohui Zhang, Shipeng Xie, MingXiu Chen, Boshen Zhang, Fu Xiong, et al. Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3d hand pose estimation under hand-object interaction. In *European Conference on Computer Vision*, pages 85–101. Springer, 2020. 3

[9] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 3, 4

[10] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7

[11] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018. 3

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3

[13] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13283, 2021. 1, 3

[14] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[16] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. Dart: Articulated hand model with diverse accessories and rich textures. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 1, 3

[17] 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, 6

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 6

[19] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021. 3

[20] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (ToG)*, 38(4):1–15, 2019. 1, 2

[21] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3d hand pose dataset. *Image and Vision Computing*, 81:25–33, 2019. 6

[22] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 3

[23] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2019. 3

[24] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3

[25] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 4, 6

[26] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual

reality. *ACM Transactions on Graphics (ToG)*, 39(4):87–1, 2020. 1

[27] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018. 1, 2, 6

[28] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019. 3

[29] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 1, 3

[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

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[32] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3

[33] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002. 7

[34] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 4, 5

[35] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *European Conference on Computer Vision*, pages 619–635. Springer, 2020. 3

[36] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[37] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[38] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 5

[39] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017. 3

[40] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 2, 5, 6, 8

[41] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 3

[42] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, pages 399–415. Springer, 2020. 2, 5

[43] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018. 2, 5

[44] Fanqing Lin, Connor Wilhelm, and Tony Martinez. Two-hand global 3d pose estimation using monocular rgb. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2373–2381, 2021. 3

[45] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26. PMLR, 2017. 6

[46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

[47] Jun Lv, Wenqiang Xu, Lixin Yang, Sucheng Qian, Chongzhao Mao, and Cewu Lu. Handtailor: Towards high-precision monocular 3d hand recovery. *arXiv preprint arXiv:2102.09244*, 2021. 1, 3

[48] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017. 3

[49] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 4

[50] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision*, pages 440–455. Springer, 2020. 1, 3, 6

[51] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564. Springer, 2020. 3, 6

[52] 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, 3, 6, 7

[53] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG)*, 38(4):1–13, 2019. 6

[54] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1154–1163, 2017. 6

[55] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 2, 3

[56] 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. 1, 2, 3, 6, 7

[57] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 1, 2, 6, 7

[58] Xavier Soria Poma, Edgar Riba, and Angel Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1923–1932, 2020. 3

[59] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *European Conference on Computer Vision*, pages 54–71. Springer, 2020. 1, 2, 3, 6

[60] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3

[61] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3, 4

[62] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):1–17, 2017. 2, 3, 6

[63] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 1, 3

[64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[65] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11230–11239, 2021. 3

[66] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision*, pages 211–228. Springer, 2020. 3

[67] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *European Conference on Computer Vision*, pages 294–310. Springer, 2016. 6

[68] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019. 1, 2

[69] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[70] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 2

[71] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 3

[72] Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. Gesturegan for hand gesture-to-gesture translation in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 774–782, 2018. 1, 3

[73] 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, 3, 6, 8

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[75] Bohan Wang, George Matcuk, and Jernej Barbič. Hand modeling and simulation using stabilized magnetic resonance imaging. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1

[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

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8798–8807, 2018. 2, 6, 7

[77] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 2, 5

[78] Yangang Wang, Cong Peng, and Yebin Liu. Mask-pose cascaded cnn for 2d hand pose estimation from single color image. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3258–3268, 2018. 6

[79] 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. 3

[80] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: Rgb-sequence-based 3d hand pose and shape estimation. In *European Conference on Computer Vision*, pages 122–139. Springer, 2020. 1, 3, 6

[81] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 3

[82] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 6, 8

[83] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016. 1, 3

[84] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986. IEEE, 2017. 6

[85] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. Hand image understanding via deep multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11281–11292, 2021. 6

[86] Wenping Zhao, Jinxiang Chai, and Ying-Qing Xu. Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data. In *Proceedings of the ACM SIGGRAPH/eurographics symposium on computer animation*, pages 33–42, 2012. 2

[87] Zimeng Zhao, Ruting Rao, and Yangang Wang. Supple: Extracting hand skeleton with spherical unwrapping profiles. In *2021 International Conference on 3D Vision (3DV)*, pages 899–909. IEEE, 2021. 3

[88] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2478–2482. IEEE, 2020. 1, 3, 4, 6

[89] Zimeng Zhao, Xi Zhao, and Yangang Wang. Travelnet: Self-supervised physically plausible hand motion learning from monocular color images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11666–11676, 2021. 3, 6

[90] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 1, 2, 3, 7

[91] 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. 1, 2, 3, 5, 6, 7

[92] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 3

[93] 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. 3, 4, 6