SiMA-Hand: Boosting 3D Hand-Mesh Reconstruction by Single-to-Multi-View Adaptation

Yinqiao Wang[1,2], **Hao Xu**[1,2], **Pheng-Ann Heng**[1,2], **Chi-Wing Fu**[1,2]
[1]Department of Computer Science and Engineering, CUHK
[2]Institute of Medical Intelligence and XR, CUHK
{yqwang, xuhao, pheng, cwfu}@cse.cuhk.edu.hk

Abstract

Estimating 3D hand mesh from RGB images is a long-standing track, in which occlusion is one of the most challenging problems. Existing attempts towards this task often fail when the occlusion dominates the image space. In this paper, we propose SiMA-Hand, aiming to boost the mesh reconstruction performance by **Si**ngle-to-**M**ulti-view **A**daptation. First, we design a multi-view hand reconstructor to fuse information across multiple views by holistically adopting feature fusion at image, joint, and vertex levels. Then, we introduce a single-view hand reconstructor equipped with SiMA. Though taking only one view as input at inference, the shape and orientation features in the single-view reconstructor can be enriched by learning non-occluded knowledge from the extra views at training, enhancing the reconstruction precision on the occluded regions. We conduct experiments on the Dex-YCB and HanCo benchmarks with challenging object- and self-caused occlusion cases, manifesting that SiMA-Hand consistently achieves superior performance over the state of the arts. Code will be released on https://github.com/JoyboyWang/SiMA-Hand_Pytorch.

Introduction

Hand-mesh reconstruction from monocular RGB image is a fundamental and significant task. It has great potential for many applications, *e.g.*, augmented reality, human-computer interactions, *etc*. Yet, to obtain high-quality reconstruction, a common but very challenging scenario is that self- or object-caused occlusions often occur, making it hard to infer the shape and pose of the hand in the occluded parts.

Recently, several methods have been proposed to alleviate the ambiguity of occlusion. Some (Chu et al. 2017; Zhu et al. 2019; Zhou et al. 2020a) attempt to extract occlusion-robust features by adopting the spatial attention mechanism. However, their performance degrades significantly when the occluded regions dominate the image. Others (Hasson et al. 2020; Yang et al. 2020a; Xu et al. 2023) leverage non-occluded information from multi-frame inputs. Yet, they often fail when the occlusion persists over nearby frames; also, cross-frame processing is typically more time-consuming.

To address this issue, another approach is to reconstruct the hand from multi-view images. Leveraging a neural network, we can effectively build a more complete hand geometry by utilizing 3D cues from different views. POEM (Yang et al. 2023) attempts to reconstruct the hand mesh based on the interaction between mesh vertices and 3D points in the intersection area of different camera frustums. Though it helps improve the performance, the computational cost is inevitably largely increased. Besides, the approach requires a multi-camera set up, which is tedious and expensive to prepare, compared with using commodity hand-held cameras.

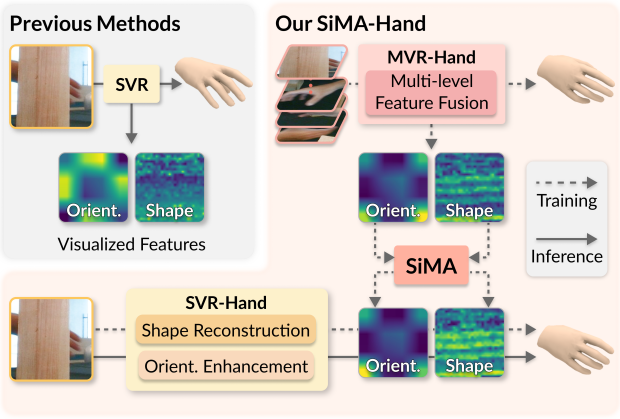

Figure 1: Framework comparison between our SiMA-Hand and previous methods. We exploit non-occluded information from the multi-view reconstructor (MVR) and carefully adopt it to the single-view reconstructor (SVR) through the proposed single-to-multi-view adaptation techniques. So, our SVR can learn to extract orientation and shape features, following the MVR's, and produce higher quality meshes.

In this paper, we present **SiMA-Hand**, a new framework formulated with **Si**ngle-to-**M**ulti-view **A**daptation to boost the hand-mesh reconstruction quality from single-view input. Fig. 1 overviews the method design. Our key insight is that human hands have limited feasible poses; given a single view, even the hand is partially occluded, we may still try to infer its 3D structure in some other views. With this preliminary, we leverage the knowledge of different views from the multi-view reconstructor (MVR) during the training, to enrich the features extracted by the single-view reconstructor (SVR) from a single-view input; see again Fig. 1. Using this

design, during the inference, the trained SVR can perform 3D hand-mesh reconstruction with only a single view as its input, yet appearing to have a multi-view input.

Technical-wise, we first design a multi-view 3D hand reconstructor, named MVR-Hand. Since all views of one hand possess different global orientations while sharing the same shape, to enable better information utilization from different views, we decouple the estimation of view-dependent orientation and view-independent shape into two separate branches, following (Xu et al. 2023). After that, we holistically design the multi-view fusion at different levels to exploit and fuse information across views. Specifically, considering that the hand orientation is related to the global picture of the entire hand, we formulate the image- and joint-level 2D feature fusion modules: the former enables knowledge exchange among views whereas the latter propagates joint information from the shape to the orientation branch. For the hand shape, as it is view-independent at the canonical pose, we design vertex-level 3D feature fusion to adaptively fuse the features of the coarse hand vertices.

Second, we introduce our single-view reconstructor, named SVR-Hand, equipped with the single-to-multi-view adaptation techniques to leverage intermediate information of the MVR-Hand to improve the single-view reconstruction quality. As regressing the hand shape at the canonical pose makes the 3D shape features less related to the camera pose, directly using the multi-view 3D vertex features as guidance is highly desirable. However, it is non-trivial to leverage the multi-view knowledge to improve the view-dependent hand-orientation precision of the target view. To this end, we propose the hand-orientation feature enhancement module in our SVR-Hand. It helps the network to effectively exploit cues from the hand-shape branch for the hand-orientation features to simulate the absent information provided by the multi-view inputs during the training.

Our main contributions are summarized below:

- We introduce SiMA-Hand, a novel framework for boosting the 3D hand-mesh reconstruction with a single-to-multi-view adaptation. To our best knowledge, this is the first work that leverages multi-view information during training to improve single-view hand reconstruction.
- We design variant feature fusion and enhancement modules. The former is distributed at image/joint/vertex levels, aggregating useful information from multiple views to promote multi-view reconstruction, whereas the latter aims to enrich single-view orientation and shape features to perform better adaptation.
- We demonstrate the effectiveness of SiMA-Hand, both quantitatively and qualitatively, on two widely-used benchmarks, Dex-YCB and HanCo, showing the state-of-the-art performance of our method.

Related Work

Single-view hand-mesh reconstruction. Single-view methods can be divided into two main categories, *i.e.*, depth-based and RGB-based, according to the input type.

Early depth-based methods (Tan et al. 2016; Taylor et al. 2014; Khamis et al. 2015) reconstruct the 3D hand mesh by iteratively fitting it to the depth image. With their strong learning ability, deep neural networks are often adopted by recent works (Malik et al. 2020; Mueller et al. 2019; Wan et al. 2020) as feature extractors.

For RGB-based methods, most works formulate hand reconstruction as a problem of regressing the MANO coefficients (Zhang et al. 2019; Zhou et al. 2020b; Yang et al. 2020b; Hasson et al. 2019; Zhang et al. 2021b; Boukhayma, Bem, and Torr 2019; Baek, Kim, and Kim 2019; Zimmermann et al. 2019; Chen et al. 2021c; Zhao, Zhao, and Wang 2021; Cao et al. 2021; Baek, Kim, and Kim 2020; Jiang et al. 2021; Liu et al. 2021; Zhang et al. 2021a; Tse et al. 2022). Other widely-used representations include voxel-based (Iqbal et al. 2018; Moon and Lee 2020; Moon et al. 2020; Yang, Chen, and Yao 2021), implicit-function-based (Mescheder et al. 2019), and vertex-based (Ge et al. 2019; Kulon et al. 2020; Chen et al. 2021b; Lin, Wang, and Liu 2021a,b). However, they rarely consider occlusions and often fail in challenging cases. HandOccNet (Park et al. 2022) implicitly handles occlusion for hand-object interaction scenes by employing self- and cross-attention. Yet, hand reconstruction from single-view input becomes severely ill-posed when the occlusions dominate the image space.

In this work, we propose to address the occlusion by leveraging useful information in multi-view inputs as guidance when training the single-view reconstructor, such that we can boost its reconstruction quality during the inference.

Multi-frame hand-mesh reconstruction. As single-view images offer limited content, some works attempt to exploit multi-frame inputs to improve the robustness and alleviate the ill-posed problem. SeqHAND (Yang et al. 2020a) utilizes Convolutional LSTM (Shi et al. 2015) to improve the temporal coherence of 3D hand-pose estimation. (Chen et al. 2021a) adopts a self-supervised approach to predict hand meshes from a video input. To address occlusion-caused ambiguity, (Wen et al. 2023) design a hierarchical network to utilize features of sequential frames. Very recently, H2ONet (Xu et al. 2023) exploits non-occluded information simultaneously at finger and hand levels from neighboring frames. Yet, the performance may degrade when a portion of the hand is consistently occluded, as the extra frames give only limited additional knowledge.

Multi-view hand reconstruction. Though multi-view reconstruction for general objects is a long-standing research topic, few methods are designed for hand. Some recent ones (Corona et al. 2022; Guo et al. 2023) use neural fields for recovering both the shape and appearance of hands. Yet, the heavy computation in rendering limits the usage in real-time tasks. POEM (Yang et al. 2023) explores multi-view mesh reconstruction on hand datasets by modeling the hand's vertices as a point cloud embedded in the intersection volume of different camera frustums. Nevertheless, at testing, it requires multiple synchronized views as input, which is expensive and tedious to obtain. SiMA-Hand is a new approach, adapting multi-view hand information to boost the performance of single-view hand reconstruction, utilizing multi-view information only at training but not at testing.

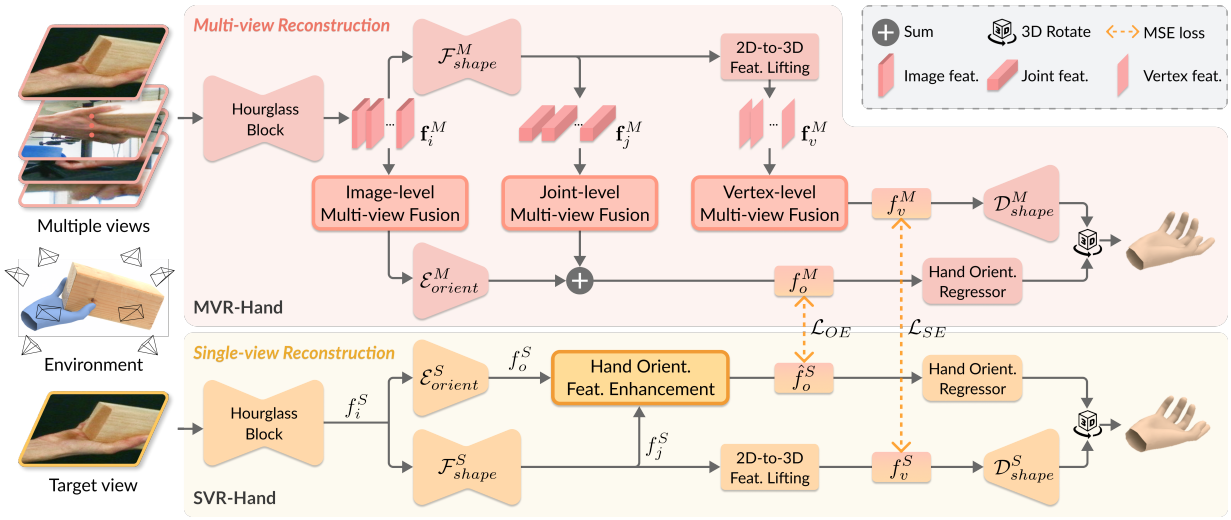

Figure 2: The architecture of SiMA-Hand: (i) the MVR-Hand takes multiple views of the hand as input for 3D hand-mesh reconstruction by fusing multi-view features at image, joint, and vertex levels; and (ii) the SVR-Hand takes only one view as input and learns to output a high-quality 3D hand mesh, even under a severely-occluded situation (see the target view at lower left), with both shape and orientation feature enhancement from the MVR-Hand. In the mathematical notations, SVR and MVR are denoted by superscripts S and M, respectively. Also, f denotes single-view or fused features, $\mathbf{f}$ denotes multi-view features, whereas subscripts i, j, v, and o denote image, joint, vertex, and orientation, respectively.

Domain alignment. It has been exploited to mitigate the disparity between the source and target domains in various tasks, *e.g.*, image classification (Yim et al. 2017; Tung and Mori 2019; Heo et al. 2019) and segmentation (Liu et al. 2019; Yang et al. 2022), 2D object detection (Guo et al. 2021; Qi et al. 2021), 3D object detection (Wang et al. 2022; Chen et al. 2023), *etc*. For RGB-based 3D hand-pose estimation, (Yang et al. 2019; Yuan, Stenger, and Kim 2018; Lin, Yang, and Yao 2023) improve network performance by distilling multi-modal information such as depth maps or point clouds. Yet, domain alignment is still under-explored in adopting multi-view knowledge to improve single-view reconstruction. To our best knowledge, SiMA-Hand is the first attempt at single-view 3D hand reconstruction.

Method

Overview

Fig. 2 shows our SiMA-Hand framework for 3D hand-mesh reconstruction. The input to the MVR-Hand is a set of multi-view images, each of these images is taken separately as the single-view input to the SVR-Hand.

- MVR-Hand effectively leverages cross-view information by fusing the image-, joint-, and vertex-level features to reconstruct a high-quality hand mesh.
- SiMA helps to reduce the domain disparity between the multi-view and single-view features via the hand-shape and orientation feature enhancement modules to boost the performance of the SVR-Hand.

Dual-branch structure. Given the multi-view images of one hand, cross-view information can be better fused by disentangling the view-independent hand shape and view-dependent hand orientation. To this end, we adopt a dual-branch structure, following (Xu et al. 2023). To be concise, we only describe the dual-branch architecture in SVR-Hand and the MVR-Hand shares the same structure.

Specifically, given a single-view image I, an hourglass block is employed to extract the image-level feature f_i^S. Then, it is fed into the dual-branch network: (i) the shape encoder $\mathcal{F}_{shape}^S$ produces the joint-level features f_j^S, then a 2D-to-3D feature lifting module (Chen et al. 2022) generates the vertex-level feature f_v^S, followed by a SpiralConv-based (Lim et al. 2018) decoder $\mathcal{D}_{shape}^S$ to predict the 3D vertices coordinates at the canonical pose; and (ii) the orientation encoder $\mathcal{E}_{orient}^S$ and an MLP-based orientation regressor are employed to estimate the hand orientation.

Multi-view Reconstruction

In the multi-view reconstruction, our MVR-Hand aims to utilize information from multiple views to obtain more accurate hand shape and orientation. So, we design three multi-view feature fusion modules at different levels: the image- and joint-level feature fusions for the hand orientation and the vertex-level feature fusion for the hand shape.

Image-level feature fusion. Fig. 3(a-1) gives the structure of this module. In detail, $\{I_k\}_{k=1}^N$ denotes N perspective images of one hand. We obtain the fused feature f_i^M by feeding the concatenated image-level features $\mathbf{f}_i^M$ into the Multi-Layer Perception (MLP) $g_i(\cdot)$, *i.e.*,

$$f_i^M = g_i(\mathrm{Cat}[\mathbf{f}_i^M]), \tag{1}$$

where $\mathrm{Cat}[\cdot]$ denotes the concatenation operation along the feature dimension.

Joint-level feature fusion. As a complementary, the joint-level feature $\mathbf{f}_j^M$ provides a global picture of the entire hand, which is a high-level guidance, for estimating the hand orientation. To effectively fuse the information in $\mathbf{f}_j^M$, inspired by (Qi et al. 2017), we concatenate the statistics of the features pooled across different views for each joint and employ an MLP, *i.e.*, g_j, to produce the correlated joint-level feature f_j^M, as shown in Fig. 3(a-2). Formally, we have

$$f_j^M = g_j(\mathrm{Cat}[\mathrm{Max}[\mathbf{f}_j^M], \mathrm{Avg}[\mathbf{f}_j^M], \mathrm{Std}[\mathbf{f}_j^M]]), \tag{2}$$

where $\mathrm{Max}[\cdot]$, $\mathrm{Avg}[\cdot]$, and $\mathrm{Std}[\cdot]$ denote the operations of maximum, average, and standard deviation pooling along the view dimension, respectively.

Vertex-level feature fusion. For the hand shape, we fuse the sparse vertex-level features $\mathbf{f}_v^M$ after the 2D-3D feature lifting. Then, we can upsample and refine the features to predict the final 3D vertex coordinates through a series of spiral convolutions. Considering the fact that the vertex-level features in the initial phases contain rich 3D information for recovering the hand shape at the canonical pose, the information derived from the single-view input could be heavily affected due to occlusion. To fuse the multi-view vertex-level features, we adopt an attention-based mechanism (Vaswani et al. 2017) by adaptively weighing the features from different views, such that the vertices within the non-occluded regions in certain views could be emphasized in the fused feature f_v^M. The procedure is illustrated in Fig. 3(a-3) and can be formulated as

$$\mathbf{W}^M = \mathrm{Softmax}(\frac{\mathbf{f}_v^M \mathbf{f}_v^{M\,T}}{\sqrt{d_v}})\mathbf{f}_v^M, \quad f_v^M = \sum_i^N \mathbf{W}_i^M \mathbf{f}_{v_i}^M, \tag{3}$$

where $\mathbf{W}_i^M$ and $\mathbf{f}_{v_i}^M$ denote the weighted score and the vertex-level feature of the i-th view, respectively. $\mathrm{Softmax}(\cdot)$ denotes the softmax operation applied on the view dimension, and d_v is the feature dimension of $\mathbf{f}_v^M$.

Single-to-Multi-view Adaptation

To avoid requiring additional views at inference, we train the single-view reconstructor, *i.e.*, the SVR-Hand, by distilling multi-view knowledge from the MVR-Hand.

Hand-shape feature enhancement. For the view-independent hand shape, since we directly regress the 3D vertex coordinates from f_v^M at the canonical pose, directly constraining the shape feature is a simple yet effective solution. We formulate the hand-shape feature enhancement as the following distillation loss between the multi-view fused feature f_v^M and the single-view feature f_v^S, *i.e.*,

$$\mathcal{L}_{SE} = ||f_v^M - f_v^S||_2. \tag{4}$$

Hand-orientation feature enhancement. Regarding view-dependent hand orientation, though the estimated result from multi-view features is notably precise, the 2D information utilized is highly related to the input image content, which intensely varies across different views. So, it is non-trivial to simulate the multi-view orientation feature directly by minimizing the distance between each pair of image- and joint-level features. To encourage the SVR-Hand to simulate the multi-view orientation features in the latent space, we design the hand-orientation feature enhancement module that learns to enrich the orientation feature f_o^S by the joint feature f_j^S. Fig. 3(b) shows its architecture, in which we first adopt a self-attention block on flattened f_j^S and f_o^S, separately, followed by a cross-attention block with the former as a query and the latter as the key and value, *i.e.*,

$$\begin{aligned} h_j^S &= \mathrm{Softmax}(\frac{f_j^S f_j^{S\,T}}{\sqrt{d_j}})f_j^S, \quad h_o^S = \mathrm{Softmax}(\frac{f_o^S f_o^{S\,T}}{\sqrt{d_o}})f_o^S, \\ \hat{f}_o^S &= \mathrm{Softmax}(\frac{h_j^S h_o^{S\,T}}{\sqrt{d_o}})h_o^S, \end{aligned} \tag{5}$$

where d_j and d_o are the feature dimensions of f_j^S and f_o^S, respectively. The enhanced feature $\hat{f}_o^S$ is then constrained by its counterpart f_o^M from the MVR-Hand, *i.e.*,

$$\mathcal{L}_{OE} = ||f_o^M - \hat{f}_o^S||_2. \tag{6}$$

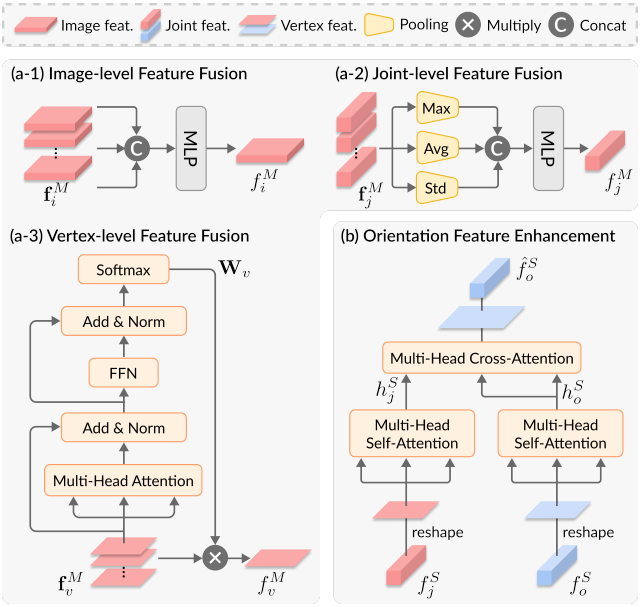

Figure 3: The designed modules in SiMA-Hand. (a) The structures of image-, joint-, and vertex-level feature fusion modules; and (b) the architecture of the orientation feature enhancement module.

Loss Functions

Our framework is trained in two stages: pre-train the MVR-Hand, and train the SVR-Hand equipped with SiMA. First, we use the same loss function $\mathcal{L}_{Recon}$ to train the SVR-Hand and MVR-Hand for 3D hand-mesh reconstruction, which consists of several 2D and 3D loss terms. Specifically, the 3D mesh loss $\mathcal{L}_M^c$ and 3D joint loss $\mathcal{L}_{J_{3D}}^c$ at the canonical pose are defined as

$$\mathcal{L}_M^c = ||\mathbf{M}^c - \hat{\mathbf{M}}^c||_1 \quad \text{and} \quad \mathcal{L}_{J_{3D}}^c = ||\mathbf{J}_c^{3D} - \hat{\mathbf{J}}_c^{3D}||_1, \tag{7}$$

Methods	Venue	PA-JPE↓	PA-JAUC↑	PA-VPE↓	PA-VAUC↑	PA-F@5↑	PA-F@15↑	JPE↓	JAUC↑	VPE↓	VAUC↑	F@5↑	F@15↑
Spurr *et al.*	*ECCV'20*	6.83	86.40	-	-	-	-	17.34	69.80	-	-	-	-
METRO	*CVPR'21*	6.99	-	-	-	-	-	15.24	-	-	-	-	-
Liu *et al.*	*CVPR'21*	6.58	-	-	-	-	-	15.28	-	-	-	-	-
Tse *et al.*	*CVPR'22*	-	-	-	-	-	-	16.05	72.20	-	-	-	-
HandOccNet	*CVPR'22*	5.80	88.40	5.50	89.00	77.95	98.99	14.04	74.82	13.09	<u>76.58</u>	<u>51.49</u>	<u>92.37</u>
MobRecon	*CVPR'22*	6.36	87.31	5.59	88.85	78.48	98.81	14.20	73.74	13.05	76.14	50.83	92.11
H2ONet	*CVPR'23*	<u>5.65</u>	<u>88.92</u>	<u>5.45</u>	<u>89.09</u>	<u>80.14</u>	<u>99.03</u>	<u>14.02</u>	<u>74.55</u>	<u>13.03</u>	76.22	51.29	92.05
SiMA-Hand	-	**5.16**	**89.37**	**4.98**	**90.04**	**81.44**	**99.24**	**13.25**	**74.99**	**12.78**	**77.12**	**53.52**	**92.48**

Table 1: Results on the Dex-YCB dataset. The best and second-best results are marked in bold and underlined for better comparison. Our method achieves the best performance on all metrics.

where $\mathbf{M}$ and $\mathbf{J}^{3D}$ mean the 3D mesh and 3D joint coordinates, respectively; superscript c denotes the canonical pose; and the hat superscript denotes the ground truth. The same losses are computed over the predicted 3D mesh and joints after rotating by the estimated hand orientation, denoted as $\mathcal{L}_M^r$ and $\mathcal{L}_{J_{3D}}^r$ respectively. The 2D joint coordinates $\mathbf{J}^{2D}$ are supervised by the normalized ground truth:

$$\mathcal{L}_{J_{2D}} = ||\mathbf{J}^{2D} - \hat{\mathbf{J}}^{2D}||_1. \tag{8}$$

For the predicted hand mesh at the canonical pose, the normal and edge-length losses are employed to penalize the outlier vertices:

$$\begin{aligned} \mathcal{L}_N^c &= \sum_{\mathbf{F}\in\mathbf{M}^c}\sum_{\mathbf{e}\in\mathbf{F}} ||\langle \mathbf{e},\ \hat{\mathbf{n}}\rangle||_1, \\ \text{and } \mathcal{L}_E^c &= \sum_{\mathbf{F}\in\mathbf{M}^c}\sum_{\mathbf{e}\in\mathbf{F}} |||\mathbf{e}| - |\hat{\mathbf{e}}|||_1, \end{aligned} \tag{9}$$

where $\mathbf{F}$ is a triangle face from the hand mesh, and $\mathbf{e}$ and $\mathbf{n}$ denote the face's edge and normal vectors, respectively.

For the hand-orientation regression, we use the L2 loss between the output rotation matrix and the ground truth:

$$\mathcal{L}_R = ||\hat{\mathbf{R}}\mathbf{R}^T - \mathbb{1}||_2, \tag{10}$$

where $\mathbf{R}$ and $\hat{\mathbf{R}}$ are the predicted and ground-truth rotation matrix, respectively, and $\mathbb{1}$ is an identity matrix.

Overall, the loss for training the MVR-Hand is

$$\begin{aligned} \mathcal{L}_{Recon} =& \mathcal{L}_M^c + \mathcal{L}_{J_{3D}}^c + \mathcal{L}_M^r + \mathcal{L}_{J_{3D}}^r \\ &+ \mathcal{L}_{J_{2D}} + \mathcal{L}_N^c + \mathcal{L}_E^c + \mathcal{L}_R. \end{aligned} \tag{11}$$

Second, we train the SVR-Hand after using the single-to-multi-view adaptation with the additional supervisions based on the above-mentioned loss functions, *i.e.*,

$$\mathcal{L}_{Total} = \mathcal{L}_{Recon} + \beta\mathcal{L}_{SE} + \gamma\mathcal{L}_{OE}, \tag{12}$$

where β and γ are parameters to balance the loss terms.

Experiments

Experimental Settings

Datasets. We conduct experiments on Dex-YCB (Chao et al. 2021) and HanCo (Zimmermann, Argus, and Brox 2021). Dex-YCB (Chao et al. 2021) is a challenging large-scale dataset for 3D hand-object reconstruction. It provides 1,000 sequences (over 582,000 frames) of 10 subjects grasping 20 different objects from 8 independent views. We adopt the default "S0" train/test split with 406,888/78,768 samples for training/testing. Evaluation on this dataset can reflect the effectiveness and robustness of different methods. HanCo (Zimmermann, Argus, and Brox 2021) is a multi-view extension of the widely-used FreiHAND dataset (Zimmermann et al. 2019). It contains more complicated hand gestures and more diverse objects compared to Dex-YCB. The original training set contains 63,864 samples recorded against a green screen, whereas the testing set contains 8,576 samples from 8 different views. We also follow (Zimmermann et al. 2019) to use the official training set augmented by randomly replacing the background with 2,195 images from Flickr, yielding another 63,864 samples for training.

Evaluation metrics. For quantitative evaluation, we adopt the commonly-used metrics following (Moon and Lee 2020; Park et al. 2022; Chen et al. 2022; Xu et al. 2023). JPE/VPE denotes the joint/vertex position error in millimeters (mm) calculated by the average Euclidean distance between the estimated and ground-truth 3D hand joint/vertex coordinates. JAUC/VAUC denotes the area under the curve of the percentage of correct keypoints (PCK) w.r.t. different error thresholds for joint/vertex. F-score is the harmonic mean of the recall and precision between the estimated and ground-truth hand-mesh vertices; we adopt F@5mm and F@15mm, which are computed with the threshold of 5mm and 15mm, respectively. Importantly, following existing work, we also report these metrics after Procrustes Alignment (PA), which directly reflects the reconstruction quality of the hand shape.

Implementation details. We follow (Chen et al. 2022) to pre-train the feature encoder network and adopt a two-stage strategy as (Xu et al. 2023) to stabilize the training. We train SiMA-Hand on four NVidia Titan V GPUs, and the Adam optimizer (Kingma and Ba 2014) is adopted. The batch size in training is set to 64 for MVR-Hand and 128/32 for SVR-Hand. The input image is resized to 128×128 and augmented by random scaling, rotating, and color jittering. All $N = 8$ views are used for training the MVR-Hand. For more details, please refer to our supplementary material and code.

Comparison with State-of-the-art Methods

Evaluation on Dex-YCB. We first present quantitative comparison results with the state-of-the-art methods on the

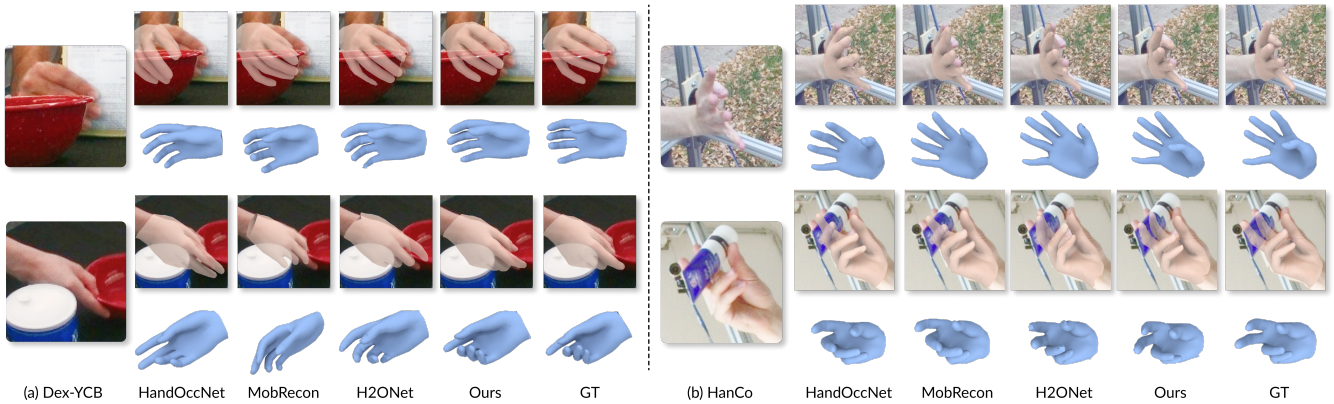

Figure 4: Qualitative comparison of our method and state-of-the-art 3D hand-mesh reconstruction methods on different datasets. The first and second rows in each example denote the normal view and another view, respectively, for better comparison.

Methods	Venue	JPE↓	JAUC↑	VPE↓	VAUC↑	F@5↑	F@15↑
I2L-MeshNet	*ECCV'20*	7.71	84.63	7.47	85.10	65.16	97.96
METRO	*CVPR'21*	7.51	85.01	6.66	86.71	71.78	98.35
Liu *et al.*	*CVPR'21*	8.44	83.19	8.17	83.71	62.87	97.07
HandOccNet	*CVPR'22*	8.09	83.87	7.85	84.34	64.53	97.31
MobRecon	*CVPR'22*	7.95	84.14	7.12	85.79	68.45	98.26
H2ONet	*CVPR'23*	6.78	86.46	6.89	86.25	70.61	98.30
SiMA-Hand	-	**6.55**	**86.93**	**6.48**	**87.07**	**73.16**	**98.48**

Table 2: Results on the HanCo dataset (after PA). Our method achieves the best performance on all metrics.

	Methods	Occluded			Non-occluded		
		VPE	PA-VPE	PA-F@5	VPE	PA-VPE	PA-F@5
Dex-YCB	HandOccNet	19.00	6.72	68.92	13.58	5.65	76.71
	MobRecon	18.48	6.76	69.47	12.49	5.53	78.71
	H2ONet	18.47	6.42	71.73	12.71	5.16	80.76
	SiMA-Hand	**18.29**	**6.18**	**73.09**	**12.40**	**4.90**	**82.01**
HanCo	HandOccNet	21.28	8.92	60.15	18.90	7.82	64.65
	MobRecon	20.08	8.85	60.68	17.66	7.07	68.66
	H2ONet	20.49	8.23	63.72	18.26	6.85	70.80
	SiMA-Hand	**18.52**	**7.67**	**67.29**	**17.27**	**6.44**	**73.32**

Table 3: Quantitative comparison between our method and recent works on occluding and non-occluding scenarios.

Dex-YCB dataset; see Tab. 1 and the left plot of Fig. 5. For fairness, we only report the performance of the single-frame version of H2ONet. From the results, we can see that our SiMA-Hand achieves *the best results consistently* on all metrics before and after PA, showing its effectiveness both in hand-shape reconstruction and hand-orientation estimation. Besides, though SiMA-Hand is proposed mainly for single-view reconstruction, our MVR-Hand (PA-J/VPE=3.57/3.87) outperforms the very recent multi-view hand reconstructor (Yang et al. 2023) (PA-J/VPE=3.93/4.00), reflecting the effectiveness of our design. Some qualitative results are shown in Fig. 4(a). Our method successfully yields plausible

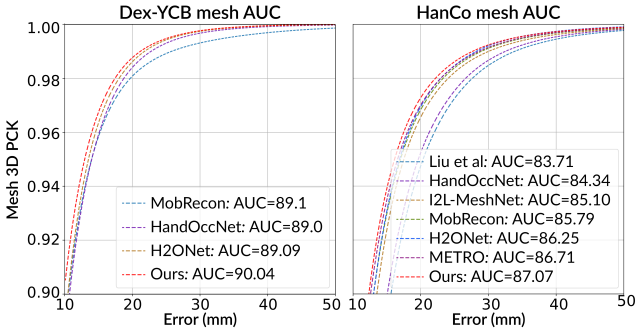

Figure 5: The mesh AUC comparison under different thresholds. Our method performs better than others, consistently.

results, while other methods often fail in heavy occlusions.

Evaluation on HanCo. To further evaluate the generalizability of our method, we conduct the same experiments on the HanCo dataset. The quantitative results after PA are shown in Tab. 2. Our SiMA-Hand consistently outperforms others on all metrics. We also provide the mesh AUC comparison under different thresholds in the right plot of Fig. 5. The visual results in Fig. 4(b) show that our method outperforms existing works by accurately reconstructing 3D hand mesh even for challenging gestures and severe self-caused occlusions. More comparisons are in our supp. material.

Detailed evaluation on occluding and non-occluding scenarios. To demonstrate the robustness of our method on challenging occluded scenes, we split the testing set of the Dex-YCB and HanCo into occluded and non-occluded parts according to the hand-level visibility as described in (Xu et al. 2023). As shown in Tab. 3, we obtain the best performance on both scenarios, manifesting the effectiveness of our idea. Please see the supp. material for full metrics.

Efficiency

Tab. 4 reports the inference speed (FPS), FLOPs, and the number of parameters of various methods. The FPS is tested

Methods	Liu *et al*.	HandOccNet	MobRecon	H2ONet	Ours
FPS	32	30	66	52	48
FLOPs (G)	3.66	15.48	0.46	0.74	1.65
#Param. (M)	32.75	37.22	8.23	25.88	78.42

Table 4: Efficiency comparison with other methods.

	Cases	Models	JPE↓	PA-JPE↓	VPE↓	PA-VPE↓
MVR	(i)	IFF w/o	8.92	3.56	8.88	3.86
	(ii)	JFF w/o	9.48	3.57	9.43	3.87
	(iii)	JFF cat.	8.55	3.57	8.54	3.86
	(iv)	VFF cat.	9.12	4.08	8.97	4.21
	(v)	VFF pool.	8.88	3.70	8.72	**3.77**
	(vi)	MVR-Hand	**8.49**	**3.57**	**8.48**	3.87
SVR	(vii)	OFE w/o	13.66	5.23	13.38	5.36
	(viii)	OFE cat.	13.59	**5.22**	13.33	**5.35**
	(ix)	SVR-Hand	**13.37**	5.24	**13.11**	5.38
SiMA	(x)	VFF cat.+SiMA	13.73	5.39	13.41	5.49
	(xi)	VFF pool.+SiMA	13.59	5.29	13.22	5.28
	(xii)	SiMA-Hand	**13.25**	**5.16**	**12.78**	**4.98**

Table 5: Ablation study on major components.

on an NVidia RTX 2080Ti. SiMA-Hand is able to run in real time, though slightly slower than MobRecon and H2ONet, it outperforms them considerably on both benchmarks.

Ablation Studies

We perform ablation studies on SiMA-Hand and its major components using **Dex-YCB**. As Tab. 5 shows, **IFF**, **JFF**, and **VFF** denote the image-, joint- and vertex-level feature fusions, respectively. **OFE** denotes the hand-orientation feature enhancement. **cat.** and **pool.** mean replacing feature fusion with concatenation and pooling, respectively. **SiMA** means adaptation between MVR- and SVR-Hand.

Image/joint/vertex-level feature fusion. The variant-level feature fusion modules are important components in our MVR-Hand. We first analyze their impact by removing or replacing one of them. Comparing Rows (i) and (vi), we can see that removing the IFF and directly using the target-view feature lead to degraded performance in orientation, thereby revealing the effectiveness of our IFF design. For the JFF, comparing Rows (ii)-(iii) and (vi), either removing it or replacing the pooling-based strategy with direct concatenation will cause negative effects on JPE and VPE, demonstrating the improvement brought by our JFF. Similarly, we replace the attention in the VFF with pooling and concatenating-based methods to study its impact on hand-shape estimation. Comparing Rows (iv)-(v) and (vi), an obvious performance drop is observed, if simply using the concatenate operation to fuse features. Though replacing it with the pooling-based strategy obtains better PA-VPE, it is not suitable for the SiMA framework due to the inevitable information loss, as demonstrated by experiments in Rows (x)-(xii). Comparing Rows (x)-(xi) and (xii), the proposed attention-based fusion in our SiMA-Hand achieves better performance than the other two feature fusion strategies, manifesting the efficacy of our full design.

	$\tilde{f}_v^S$	f_v^S	f_j^S	f_o^S	$\hat{f}_o^S$	JPE↓	PA-JPE↓	VPE↓	PA-VPE↓
(i)	✓					13.90	5.30	13.39	5.13
(ii)		✓				13.79	5.17	13.29	4.98
(iii)		✓	✓			13.74	5.16	13.24	4.98
(iv)		✓		✓		13.57	5.18	13.09	4.99
(v)		✓			✓	**13.25**	**5.16**	**12.78**	**4.98**

Table 6: Ablation study on supervision in SiMA.

Hand-shape/orientation feature enhancement. To assess the effectiveness of the orientation feature enhancement module, we conduct experiments in Rows (vii)-(ix) of Tab. 5. Comparing Rows (vii)-(viii) and Row (ix) separately, we can find that directly removing the OFF or simply concatenating the joint- and orientation feature shows subpar metrics in terms of orientation than our current design, indicating its benefit in hand-orientation estimation. Besides, Rows (ix) and (xii) share the same network structures but without and with the feature enhancement constraints, respectively. Comparing Rows (ix) and (xii), introducing such a constraint brings a large improvement in both VPE and PA-VPE, thereby showing its necessity.

Distillation in SiMA. We also explore different stages of distillation on the hand-shape and -orientation features in SiMA, as shown in Tab. 6. We perform supervision on the last-layer vertex-level features in $\mathcal{D}^S_{shape}$, denoted as $\tilde{f}_v^S$. Comparing Rows (i) and (ii), the distillation on the vertex-level feature at the initial stage brings more improvements in hand-shape estimation than the later one, indicating its heightened potential in simulating multi-view shape features. Further, by fixing the hand-shape constraint on f_v^S, we investigate the effects of various constraints on the efficacy of hand-orientation estimation in Rows (iii)-(v). The result shows that distillation on f_j^S has an adverse impact on the performance of hand-orientation estimation, while the one on f_o^S only yields a marginal improvement. The obvious boost in JPE and VPE comes from the distillation on the enhanced hand-orientation feature $\hat{f}_o^S$, indicating that the global information of the hand-shape features facilitates hand-orientation estimation after the fusion of f_j^S and f_o^S.

Conclusion

We introduce SiMA-Hand, a new framework for boosting the performance of 3D hand-mesh reconstruction by single-to-multi-view adaptation. To fully exploit cross-view information, we design image-, joint-, and vertex-level feature fusion. Besides, we propose feature enhancement on hand shape and orientation to leverage multi-view knowledge for tackling the occlusion issue in single-view reconstruction. Experimental results also confirm the state-of-the-art performance of SiMA-Hand on two widely-used benchmarks.

Acknowledgments

This work was supported in part by the grants from the Research Grants Council of the Hong Kong SAR, China (Project No. T45-401/22-N) and Project No. MHP/086/21 of the Hong Kong Innovation and Technology Fund.

References

Baek, S.; Kim, K. I.; and Kim, T.-K. 2019. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *CVPR*, 1067–1076.

Baek, S.; Kim, K. I.; and Kim, T.-K. 2020. Weakly-supervised domain adaptation via GAN and mesh model for estimating 3D hand poses interacting objects. In *CVPR*, 6121–6131.

Boukhayma, A.; Bem, R. d.; and Torr, P. H. 2019. 3D hand shape and pose from images in the wild. In *CVPR*, 10843–10852.

Cao, Z.; Radosavovic, I.; Kanazawa, A.; and Malik, J. 2021. Reconstructing hand-object interactions in the wild. In *ICCV*, 12417–12426.

Chao, Y.-W.; Yang, W.; Xiang, Y.; Molchanov, P.; Handa, A.; Tremblay, J.; Narang, Y. S.; Van Wyk, K.; Iqbal, U.; Birchfield, S.; et al. 2021. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 9044–9053.

Chen, L.; Lin, S.-Y.; Xie, Y.; Lin, Y.-Y.; and Xie, X. 2021a. Temporal-aware self-supervised learning for 3D hand pose and mesh estimation in videos. In *WACV*, 1050–1059.

Chen, X.; Liu, Y.; Dong, Y.; Zhang, X.; Ma, C.; Xiong, Y.; Zhang, Y.; and Guo, X. 2022. MobRecon: Mobile-Friendly Hand Mesh Reconstruction from Monocular Image. In *CVPR*, 20544–20554.

Chen, X.; Liu, Y.; Ma, C.; Chang, J.; Wang, H.; Chen, T.; Guo, X.; Wan, P.; and Zheng, W. 2021b. Camera-space hand mesh recovery via semantic aggregation and adaptive 2D-1D registration. In *CVPR*, 13274–13283.

Chen, Y.; Tu, Z.; Kang, D.; Bao, L.; Zhang, Y.; Zhe, X.; Chen, R.; and Yuan, J. 2021c. Model-based 3D hand reconstruction via self-supervised learning. In *CVPR*, 10451–10460.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2023. BEVDistill: Cross-Modal BEV Distillation for Multi-View 3D Object Detection. In *ICLR*.

Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A. L.; and Wang, X. 2017. Multi-context attention for human pose estimation. In *CVPR*, 1831–1840.

Corona, E.; Hodan, T.; Vo, M.; Moreno-Noguer, F.; Sweeney, C.; Newcombe, R.; and Ma, L. 2022. Lisa: Learning implicit shape and appearance of hands. In *CVPR*, 20533–20543.

Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; and Yuan, J. 2019. 3D hand shape and pose estimation from a single RGB image. In *CVPR*, 10833–10842.

Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; and Xu, C. 2021. Distilling object detectors via decoupled features. In *CVPR*, 2154–2164.

Guo, Z.; Zhou, W.; Wang, M.; Li, L.; and Li, H. 2023. HandNeRF: Neural Radiance Fields for Animatable Interacting Hands. In *CVPR*, 21078–21087.

Hasson, Y.; Tekin, B.; Bogo, F.; Laptev, I.; Pollefeys, M.; and Schmid, C. 2020. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 571–580.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 11807–11816.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019. A comprehensive overhaul of feature distillation. In *ICCV*, 1921–1930.

Iqbal, U.; Molchanov, P.; Gall, T. B. J.; and Kautz, J. 2018. Hand pose estimation via latent 2.5D heatmap regression. In *ECCV*, 118–134.

Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, 11107–11116.

Khamis, S.; Taylor, J.; Shotton, J.; Keskin, C.; Izadi, S.; and Fitzgibbon, A. 2015. Learning an efficient model of hand shape variation from depth images. In *CVPR*, 2540–2548.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kulon, D.; Guler, R. A.; Kokkinos, I.; Bronstein, M. M.; and Zafeiriou, S. 2020. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 4990–5000.

Lim, I.; Dielen, A.; Campen, M.; and Kobbelt, L. 2018. A simple approach to intrinsic correspondence learning on unstructured 3D meshes. In *ECCVW*, 349–362.

Lin, K.; Wang, L.; and Liu, Z. 2021a. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 1954–1963.

Lin, K.; Wang, L.; and Liu, Z. 2021b. Mesh graphormer. In *ICCV*, 12939–12948.

Lin, Q.; Yang, L.; and Yao, A. 2023. Cross-Domain 3D Hand Pose Estimation With Dual Modalities. In *CVPR*, 17184–17193.

Liu, S.; Jiang, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*, 14687–14697.

Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; and Wang, J. 2019. Structured knowledge distillation for semantic segmentation. In *CVPR*, 2604–2613.

Malik, J.; Abdelaziz, I.; Elhayek, A.; Shimada, S.; Ali, S. A.; Golyanik, V.; Theobalt, C.; and Stricker, D. 2020. HandVoxNet: Deep voxel-based network for 3D hand shape and pose estimation from a single depth map. In *CVPR*, 7113–7122.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, 4460–4470.

Moon, G.; and Lee, K. M. 2020. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and

mesh estimation from a single RGB image. In *ECCV*, 752–768.

Moon, G.; Yu, S.-I.; Wen, H.; Shiratori, T.; and Lee, K. M. 2020. Interhand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, 548–564.

Mueller, F.; Davis, M.; Bernard, F.; Sotnychenko, O.; Verschoor, M.; Otaduy, M. A.; Casas, D.; and Theobalt, C. 2019. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM TOG*, 38(4): 1–13.

Park, J.; Oh, Y.; Moon, G.; Choi, H.; and Lee, K. M. 2022. HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network. In *CVPR*, 1496–1505.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.

Qi, L.; Kuen, J.; Gu, J.; Lin, Z.; Wang, Y.; Chen, Y.; Li, Y.; and Jia, J. 2021. Multi-scale aligned distillation for low-resolution detection. In *CVPR*, 14443–14453.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-k.; and Woo, W.-c. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *NeurIPS*, volume 1, 802–810.

Tan, D. J.; Cashman, T.; Taylor, J.; Fitzgibbon, A.; Tarlow, D.; Khamis, S.; Izadi, S.; and Shotton, J. 2016. Fits like a glove: Rapid and reliable hand shape personalization. In *CVPR*, 5610–5619.

Taylor, J.; Stebbing, R.; Ramakrishna, V.; Keskin, C.; Shotton, J.; Izadi, S.; Hertzmann, A.; and Fitzgibbon, A. 2014. User-specific hand modeling from monocular depth sequences. In *CVPR*, 644–651.

Tse, T. H. E.; Kim, K. I.; Leonardis, A.; and Chang, H. J. 2022. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *CVPR*, 1664–1674.

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *ICCV*, 1365–1374.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 6000–6010.

Wan, C.; Probst, T.; Gool, L. V.; and Yao, A. 2020. Dual Grid Net: Hand mesh vertex regression from single depth maps. In *ECCV*, 442–459.

Wang, T.; Hu, X.; Liu, Z.; and Fu, C.-W. 2022. Sparse2Dense: Learning to densify 3d features for 3d object detection. *NeurIPS*, 35: 38533–38545.

Wen, Y.; Pan, H.; Yang, L.; Pan, J.; Komura, T.; and Wang, W. 2023. Hierarchical temporal transformer for 3D hand pose estimation and action recognition from egocentric RGB videos. In *CVPR*, 21243–21253.

Xu, H.; Wang, T.; Tang, X.; and Fu, C.-W. 2023. H2ONet: Hand-Occlusion-and-Orientation-Aware Network for Real-Time 3D Hand Mesh Reconstruction. In *CVPR*, 17048–17058.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022. Cross-image relational knowledge distillation for semantic segmentation. In *CVPR*, 12319–12328.

Yang, J.; Chang, H. J.; Lee, S.; and Kwak, N. 2020a. Seqhand: RGB-sequence-based 3D hand pose and shape estimation. In *ECCV*, 122–139.

Yang, L.; Chen, S.; and Yao, A. 2021. SemiHand: Semi-supervised hand pose estimation with consistency. In *ICCV*, 11364–11373.

Yang, L.; Li, J.; Xu, W.; Diao, Y.; and Lu, C. 2020b. BiHand: Recovering Hand Mesh with Multi-stage Bisected Hourglass Networks. In *BMVC*.

Yang, L.; Li, S.; Lee, D.; and Yao, A. 2019. Aligning latent spaces for 3d hand pose estimation. In *ICCV*, 2335–2343.

Yang, L.; Xu, J.; Zhong, L.; Zhan, X.; Wang, Z.; Wu, K.; and Lu, C. 2023. POEM: Reconstructing Hand in a Point Embedded Multi-View Stereo. In *CVPR*, 21108–21117.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 4133–4141.

Yuan, S.; Stenger, B.; and Kim, T.-K. 2018. RGB-based 3D hand pose estimation via privileged learning with depth images. *arXiv preprint arXiv:1811.07376*.

Zhang, B.; Wang, Y.; Deng, X.; Zhang, Y.; Tan, P.; Ma, C.; and Wang, H. 2021a. Interacting two-hand 3D pose and shape reconstruction from single color image. In *ICCV*, 11354–11363.

Zhang, X.; Huang, H.; Tan, J.; Xu, H.; Yang, C.; Peng, G.; Wang, L.; and Liu, J. 2021b. Hand image understanding via deep multi-task learning. In *ICCV*, 11281–11292.

Zhang, X.; Li, Q.; Mo, H.; Zhang, W.; and Zheng, W. 2019. End-to-end hand mesh recovery from a monocular RGB image. In *ICCV*, 2354–2364.

Zhao, Z.; Zhao, X.; and Wang, Y. 2021. TravelNet: Self-supervised physically plausible hand motion learning from monocular color images. In *ICCV*, 11666–11676.

Zhou, L.; Chen, Y.; Gao, Y.; Wang, J.; and Lu, H. 2020a. Occlusion-aware siamese network for human pose estimation. In *ECCV*, 396–412.

Zhou, Y.; Habermann, M.; Xu, W.; Habibie, I.; Theobalt, C.; and Xu, F. 2020b. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 5346–5355.

Zhu, M.; Shi, D.; Zheng, M.; and Sadiq, M. 2019. Robust facial landmark detection via occlusion-adaptive deep networks. In *CVPR*, 3486–3496.

Zimmermann, C.; Argus, M.; and Brox, T. 2021. Contrastive representation learning for hand shape estimation. In *DAGM German Conference on Pattern Recognition*, 250–264. Springer.

Zimmermann, C.; Ceylan, D.; Yang, J.; Russell, B.; Argus, M.; and Brox, T. 2019. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 813–822.