SCAT: Stride Consistency with Auto-regressive regressor and Transformer for hand pose estimation

Daiheng Gao[1], Bang Zhang[1], Qi Wang[1], Xindi Zhang[2], Pan Pan[1], Yinghui Xu[1],
{*daiheng.gdh, zhangbang.zb, wilson.wq, panpan.pp*}*@alibaba-inc.com, zhangxindi@gmail.com,*
renji.xyh@taobao.com
[1]DAMO Academy, Alibaba Group [2]Queen Mary University of London

Abstract

The current state-of-the-art monocular 3D hand pose estimation methods are mostly model-based. For instance, MANO is one of the most popular hand parametric models, which can depict hand shapes and poses. It is widely adopted for estimating hand poses in images and videos. However, MANO is a parametric model derived from scanned hand data with limited shapes and poses which constrains its capability in depicting in-the-wild shape and pose variations. In this paper, we propose a 3D hand pose estimation approach which does not depends on any parametric hand models yet can still accurately estimate in-the-wild hand poses. Our approach (Stride Consistency with Autoregressive regressor and Transformer, SCAT) offers a new representation for measuring hand poses. The new representation includes a mean shape hand template and its 21 hand joint offsets depicting the 3D distances between the hand template and the hand that needs to be estimated. Besides, SCAT can generate a robust and smooth linear mapping between visual feature maps and the target 3D offsets, ensuring inter-frame smoothness and removing motion jittering. We also introduce an auto-regressive refinement procedure for iteratively refining the hand pose estimation. Extensive experiments show that our SCAT can generate more accurate and smoother 3D hand pose estimation results compared with the state-of-the-art methods.

1. Introduction

As we all know, hands are the most crucial part of the human body to interact with the outside world and the embodiment of productivity. With the need for cost-saving and computational resource reduction, especially in AR, human-computer interaction (HCI) and many other scenes, 3D hand pose estimation from a single RGB image is becoming much more significant since the current trend of AR/VR (Facebook Quest2). 3D hand pose estimation is a challenging problem, due to the ambiguity of depth and exterior parameters of the camera, fast movements, occlusions, complex articulated motion and indistinguishable skin appearance. Thus the 3D hand pose estimation has attracted a lot of attention in the academia.

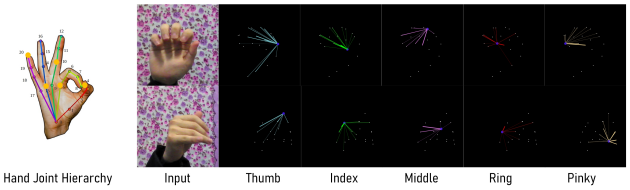

Figure 1. We present a model-free hand pose estimation approach with a single RGB image input. The left is our hand skeleton hierarchy and joint numbers, benefits from transformer, we visualize the learned relationship between joints through choosing representative joint of each finger: 4 (for thumb), 5 (for index), 10 (for middle), 13 (for ring), 20 (for pinky), where brighter color indicates stronger correlations.

To make the problem tractable, many state-of-the-art methods [50, 37, 3] incorporate prior knowledge, i.e, geometry, forward kinematics [20] either inverse kinematics [1]. The most common practice inside them is to regress the θ (pose), β (shape) of a canonical 3D parametric hand model MANO [36] and the camera parameters that serve to project the 3D coordinate to the 2D image plane, aiming to estimate a reasonable and reliable hand pose.

However, MANO [36] is a data-driven parametric model that shares the same formulation as SMPL [28]. Following the literature of MANO: *"MANO is learned from around 1000 high-resolution 3D scans of hands of 31 subjects in a wide variety of hand poses"*. Through experiments on the MANO-based model [37, 50] in complex gestures: hand heart, metal horns and etc, we found it is difficult for the MANO-based model to reconstruct a satisfactory gesture. The reason in obtaining an inferior estimation comes down to the limited and monotonous exemplars that are used to construct the parametric model. Moreover, since MANO

has left and right-hand model, which adds to the difficulty in real-world use since an extra model is needed, indicating whether an input hand image is left or right.

Here, we want to determine whether it is possible to construct a rational and expressive representation without any 3D parametric models. So we naturally turn to Model-free methods. The beginning of this genre is estimating the 3D joint coordinates directly [51]. Then researchers realized that using the heatmap (from 1D to 2.5D heatmap), which broaden the interests of a single hand joint to its corresponding neighborhood region, dramatically improves the estimation performance [30, 50, 19]. Moreover, Graph CNN [25, 7], which is adept at modeling the interactions and relationships of vertex/joint in its neighborhood, has attracted much interest. However, due to the progressive assimilation phenomenon in each cluster of graphs during training, graph-based methods are not efficient in modeling non-local vertex/joint-based interactions.

Recently, METRO [27] take advantage of the transformer [43] to propose a simple yet effective framework to model global vertex-vertex interactions, which achieves the current state-of-the-art in both body and hand pose estimation. Nonetheless, METRO still suffers from some limitations: it needs a fixed CNN network to extract the 1D feature of a single RGB image, which leads to sub-optimal predictions considering only a minority of neurons changes during the training process. Moreover, the transformer-based [43] vertex-vertex METRO requires an enormous amount of computing powers and GPU memories since it regresses the 3D coordinates of hand joints and mesh vertices in parallel. The number of template hand mesh vertices is 778, which means they need to input an extremely large feature map of $[batchsize, 778, 2051]$ into the regressor of METRO (2051 consists of 2048 feature vector extracted by the fixed pre-trained CNN and the 3D coordinates of the vertices or joints).

Inspired by the transformer encoder in constructing meaningful and mighty interrelationship, **SCAT** (**S**tride **C**onsistency with **A**uto-regressive regressor and **T**ransformer) cast off the conventional practice of using the 3D parametric model by leveraging the multi-head attention mechanism of Transformer [43]. Unlike METRO, which outputting the 3D vertex/joint coordinates directly, we obtain coarse predictions C_{coarse} through adding the transformer encoder's output (offset O) on pre-defined 21 basis coordinates (mean M) extracted from a standard hand template mesh. Through this mean-plus-offset strategy, we obtain a biomechanical-plausible prediction that is empowered by M. As shown in Fig 1, we plot five typical joints and their relationships with other joints. It demonstrates that our SCAT can establish biomechanical-rational joint-joint topologic relations, which obeys the principle as [45] points out. Additionally, we only model the joint-joint relationships, which are designated to mitigate the computational load and reduce the GPU memories.

Since the previous frame-level models are incapable of ensuring a smooth transition through time varies, we present a novel pose length regularization loss that encourages good conditioning in the mapping from feature maps $F_i, i \in 1, 2, ..., 21$ to the offsets of 3D hand pose O_i, which is the stride consistency as we proposed in title. Through the Jacobian matrix's help $J_{f,i} = \delta O_i / \delta F_i$, we then impose the penalty term to robustly produce a feature map that encourages a fixed-size step in the feature map space results in a non-zero, fixed-magnitude change in the output 3D offsets of hand pose. According to our experiments, this novel loss is crucial for maintaining the inter-frame consistency as well as reducing the jitters.

Following the successful practice of HMR [21], we develop a coarse-to-fine strategy to obtain the fine-grained 3D predictions. In short, we refined the coarse predictions, which derived from adding the 3D output offset from the transformer encoder of SCAT to our pre-defined hand-picked 21 key joint coordinates, through feed the coarse output to the auto-regressive regressor as depicted in section 3.3.

In this work, we propose SCAT, a Stride Consistency with Auto-regressive and Transformer for 3D hand pose estimation from a single RGB input image. The contributions of SCAT are summarized below:

- We introduce an end-to-end, model-free method: SCAT, which estimates an accurate yet smoother 3D hand pose on mainstream datasets, represents a competitive performance compared to the model-based method (MANO).
- SCAT learns to discover both short- and long-range interactions among 21 hand joints through multi-layer transformer encoder structure and the mean-plus-offset strategy, which yields convincing results compared to the MANO-based methods.
- We use an auto-regressive manner to regress both the 3D hand pose and camera intrinsic parameters (we use orthogonal projection) together, which solves the sub-optimal prediction problem of multi-modal distribution (cameras and poses) when using a single-mode model. We also develop a novel pose length regularization loss to enable a consistent stride between feature maps and 3D offsets, which vastly increased the stability of estimation result with time varies. It is a handy tech for real-world applications.

2. Related work

In the following, we discuss the methods that are closely related to our work.

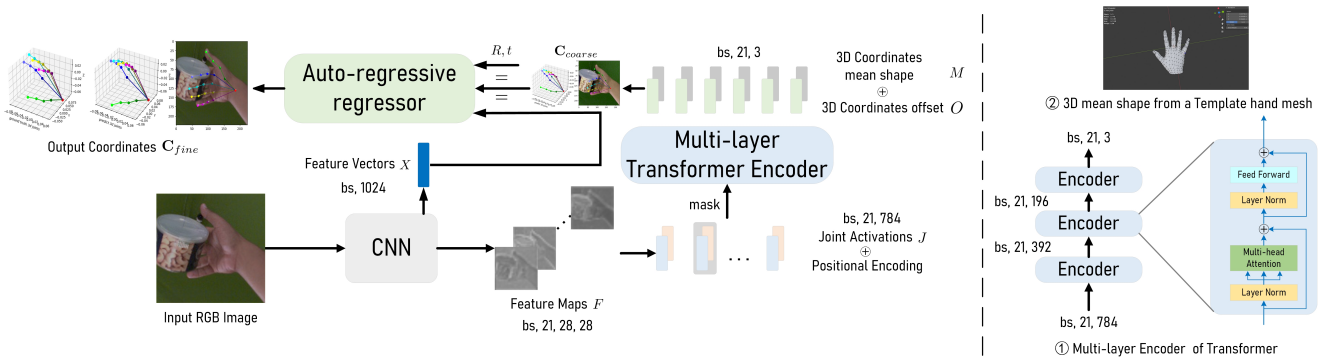

Figure 2. **Overview of the proposed framework**. SCAT receives a single RGB image and yields the feature maps **F** and feature vector **X** through a trainable convolutional neural network (CNN). Then we flatten the 2D feature maps to 1D feature vector **J** of fixed 21 joints, after performing positional encoding to **J**, a multi-layer transformer encoder is used to regress the 3D coordinates offset O = $(O_{x_i}, O_{y_i}, O_{z_i})$, the structure of which is depicted in the right. Next, we extract the mean shape M = $(M_{x_i}, M_{y_i}, M_{z_i})$ from a pre-defined hand mesh template (here we use the *hand mean* template from MANO [36] with 778 vertices, 1,538 faces and 2,315 edges, the detail is attached in the Sup. Mat.), after adding the offset O_i to the mean shape M_i, the satisfied 3D coordinates of 21 key joints are obtained by an auto-regressive manner. Here, $\bigoplus$ represents for the element-wise addition operation, $\|$ denotes the concatenation operation.

2.1. Single Image 3D Hand Pose Estimation

Although works that using depth sensors [33, 40] or inertial measurement units (IMU) [17, 44] have achieved appealing results on estimating 3D hand pose, they suffer from their own drawbacks: for depth sensors, they are incapable of working under bright sunlight and people have to be close to the sensor; for IMUs, the accumulation error is inevitable and the price of IMU is quite expensive for consumer-level usage.

To reduce costs and computational overheads, researchers recently started to research 3D hand pose estimation from monocular RGB images, which is even more challenging because of the ambiguity of depth, fast movements, occlusions, complex articulated motion and indistinguishable skin appearance.

Since it is challenging to regress the 3D pose directly from an RGB image, recent works further propose to leverage various human hand priors or segmentation maps in solving this. Zimmermann and Brox [51] have trained a CNN-based model that estimates 3D joint coordinates directly from an RGB image. However, this kind of method has been found disable to estimates a reasonable pose with partial occlusions.

For achieving a more reliable estimation result, Mueller *et al*. [32] combines CNN (heatmap-based) with kinematic 3D hand model to do skeleton fitting for RGB image input, Boukhayma *et al*. [3] uses images and 2D pose predicted from OpenPose [5] as input and regress the parameters of the MANO [36].

As there are some people who regard the parametric model as a burden, GraphCNN [8] has been used to directly regress 3D hand shape from a single RGB input. Ge *et al*. [11] directly regressed the hand by taking advantage of GraphCNN, but as a dataset with ground truth hand meshes is required for training, which restrained its appliance to the in-the-wild data. Choi *et al*. [7] proposed Pose2Mesh, which is a cascaded model using GraphCNN to reconstructs human mesh directly. While GraphCNN-based methods [11, 7, 25] are designed to model neighborhood vertex-based interactions based on a pre-specified mesh topology, it is less efficient in modeling longer range restrained by the neighbors of per-vertex due to the inherent properties of GraphCNN.

METRO, recently proposed by Lin *et al*. [27], models global interactions among joints and mesh vertices without being limited by any mesh topology, and is the first method learns with multi-head attention to model the non-local relationship between joints. Although METRO has achieved great success both in estimating hand and body, the computational complexity is exponentially growing as the number of vertex increases, which is the main bottleneck for real-word applications. As a contrast, SCAT models the sparse key joint instead of every vertex of a template mesh. A mean-plus-offset strategy is used instead of predicting output directly, implicitly ensuring a reasonable and reliable structure without the need for any sorts of pre-defined kinematic trees or explicit biomechanical constraints.

2.2. Transformer in Computer Vision

Transformer [43], first applied to the natural language processing (NLP) field, is a type of deep neural network mainly based on the multi-head self-attention mechanism. Transformer nowadays dominating the NLP area in its superior performance in language modeling at scale, BERT [9] and GPT [4] are the two milestone achievements by incorporating advantages of transformer.

In the past two years, some researchers have explored

whether similar models can learn useful representations for images. In the past few months, transformer-based network have achieved astonishing progress in image classification, such as iGPT [6], ViT [10] and DeiT [42]. There are also many paper that adopt transformer to other vision tasks: Re-Identification (ReID) [16], Multi Object Tracking (MOT) [29] and etc.

As for pose estimation, METRO [27] and TransPose [48] are the latest paper that integrates transformer into the ordinary CNN pipeline to get a better human pose estimation either fine-grained reconstruction result. However, 3D hand pose estimation has not been explored independently in this area. In this study, we estimate the 3D coordinates of the 21 hand key joints through the combination with transformer and CNN.

3. Proposed Method

The overview of SCAT is depicted in Figure 2. It takes an image of size $224 \times 224 \times 3$ as input, and predicts a set of hand joints feature vector $J_i, i \in 1, 2, ..., 21$ and the intrinsic parameter R (rescale factor), t (2D translation) for reprojecting the 3D joints coordinates to the 2D image plane. SCAT consists of three parts: Convolutional Neural Network (CNN), Multi-Layer Transformer and Auto-regressive Regressor. First, a CNN is used to extract the feature vector of desired size from an input image. Next, Multi-Layer Transformer takes in the feature vector with positional encoding and outputs the offset (3D coordinates of the hand joint) in parallel. The final stage of SCAT is to add the offset O to the 3D mean joint locations of pre-defined template hand mesh M and regress the C_{fine} through an auto-regressive manner. We will go through them in details as below.

3.1. CNN

First, we take advantage of the Convolutional Neural Network (CNN) to extract low-level features since CNN is powerful in handling high-dimensional image features. Different from the practice of METRO [27] that use the last feature vector of a pre-trained CNN model on ImageNet classification task, we use the intermediate feature map $\mathbf{F} \in \mathbb{R}^{N \times 21 \times 28 \times 28}$ (N is batch size) and input the flattened $\mathbf{F}$ of size $\mathbb{R}^{N \times 21 \times 784}$ to the multi-layer transformer. Then branch out to extract an apriori info $\mathbf{X} \in \mathbb{R}^{N \times 1024}$, which helps to regress the final 3D coordinates and camera intrinsic parameter $R \in \mathbb{R}^{N \times 1}, t \in \mathbb{R}^{N \times 2}$ iteratively.

With this design, CNN is trained to yield the most suitable representations for estimating hand pose. In SCAT, transformer benefits from the feature maps from CNN since ViT [10]-like architectures (with not even one convolution layer) is unable to regress the pose steadily and accurately according to our experiments. Moreover, we use pose length regularization to obtain a good linear mapping from feature maps $F_i, i \in 1, 2, ..., 21$ to the offsets of 3D hand pose O_i.

3.2. Multi-Layer Transformer

Inspired by METRO [27], we construct a reasonable and reliable joint-joint relationship through a similar progressive dimensionality reduction strategy. As depicted in the bottom right of Figure 2, the input to the transformer encoder are the hand joint feature vectors $\mathbf{J}$ and output the offset of 3D coordinates $\mathbf{O} \in \mathbb{R}^{N \times 21 \times 3}$. To decouple the 21 key joints, we use positional encoding to preserve the positional information explicitly. Moreover, learned from the successes [9, 41] in using the Masked Language Modeling (MLM) to the NLP field, we mask some percentages of the input joints J_i at random to predict an output that may possess good robustness for the partial occlusion problem, which frequently appears in the hand-interaction scenarios.

3.3. Auto-regressive Regressor

The last module of SCAT is the auto-regressive regressor. After a variety of combinations of intrinsic parameters R, t, feature vector X and coarse 3D coordinates predictions C_{coarse} = M + O were tested, we found that the original iterative manner proposed by Kanazawa *et al*. [21] is very effective in modeling a multi-modal distribution instead of suffering the sub-optimal prediction problem. To be specific, we concatenate X, R, t and the flattened C_{coarse} $\in \mathbb{R}^{N \times 63}$ to a total feature vector $\in \mathbb{R}^{N \times 1090}$. To balance the computational resources and the result, we set $iterations = 3$ as a trade-off. According to our Sup. Mat, the more iterations are, the better result is achieved.

Then we use the way in Algorithm 1 to regress a fine-grained C_{fine}:

Algorithm 1: Auto-regressive Regressor in estimating a fine-grained $\mathbf{C}_{fine}$

Input: R, t, $\mathbf{X}$, $\mathbf{C}_{coarse}$
Initialize: $i = 0, iterations = 3$; $\mathbf{C}_{fine}$ =$\mathbf{C}_{coarse}$
Output: $\mathbf{C}_{fine}$
while $i \leq iterations$ **do**
  $input$ = $\mathbf{X}$ || R|| t|| $\mathbf{C}_{fine}$;
  $out = regressor(input)$;
  $\mathbf{C}_{fine}$ =$\mathbf{C}_{fine}$+ out ;
  i += 1;
end

3.4. Loss

Similar to the Frankmocap [37], Minimal Hand [50], METRO [27], we consider 3D key point annotations (local root-relative coordinate system) and 2D key point annotations (image plane). Let C_{3D} denote the output 3D coordinates, and C_{2D} is the reprojection of 2D position in the

image plane. The 3D loss can be directly computed through Mean Square Error (MSE):

$$\mathcal{L}_{3D} = \left\| \mathbf{C}_{3D} - \hat{\mathbf{C}}_{3D} \right\|_2^2 \tag{1}$$

For a better alignment between 3D and 2D [21, 25], we project the 3D joints C_{3D} to the image plane using the estimated R, t, the 2D loss is calculated as below:

$$\mathcal{L}_{2D} = \left\| \mathbf{C}_{2D} - \hat{\mathbf{C}}_{2D} \right\|_1 \tag{2}$$

Since a good linear mapping between the latent feature map to the 21 key joint offsets is of great importance in obtaining a stable result through time varies, which could offer assurance to reduce the jitter between frames and maintain a smoothly inter-frame transition. Learned from the path length regularization from StyleGAN2 [23], we further migrate it to our case as pose length regularization.

We can measure the deviation from this novel ideal empirically by stepping into random directions in the output key joint offset and observing the corresponding gradients. These gradients should have close to an equal length regardless of $\mathbf{F}$, indicating that the mapping from the latent feature map space to the 21 key point space is well-conditioned.

Through the help of Jacobian matrix $J_{f,i} = \delta O_i / \delta F_i$, pose length regularization is added as below:

$$\mathcal{L}_{reg} = (\left\| \mathbf{J}_f^T \right\|_2 - a)^2 \tag{3}$$

where a is the moving average of the Jacobian $\mathbf{J}_f^T$. The overall loss $\mathcal{L}$ used to train our hand pose estimation model is defined below:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{3D} + \lambda_2 \mathcal{L}_{2D} + \lambda_3 \mathcal{L}_{reg} \tag{4}$$

$\lambda_{1,2,3}$ are weights that set to balance the loss, the weights are set to $\lambda_1 = 100,000$, $\lambda_2 = 10$, $\lambda_3 = 10$ in all our experiments.

4. Experiments

In this section, we first describe the datasets used for training. Then we describe the evaluation metrics as well as the detailed settings in our experiments. After that, we compare our SCAT with state-of-the-art approaches. Finally, we perform ablation studies to demonstrate the key components of SCAT.

4.1. Implementation Details

4.1.1 Datasets

We train SCAT on the 6 publicly widely used 3D hand datasets: Rendered Handpose Dataset (RHD) [51], FreiHAND [52], HO-3D [14], Stereo Hand Pose Tracking Benchmark (STB) [49], Multi-view Hand Pose (MHP) [12] and InterHand2.6M [31]. RHD [51] is a synthetic dataset that consists of 41,258 images with 2D and 3D annotations. FreiHAND is a real-world dataset with ground truth 3D hand joints, the 3D annotations are obtained by a multi-camera system and a semi-automated approach. HO-3D is mainly focused on the interaction between hands and objects. It is a real-world dataset and we use this dataset to prove the validity of the masked mechanism used in SCAT to the partial occlusion scenes. STB is composed of 15,000 training samples and 3,000 testing samples with both RGB images and depth images. MHP uses Leap Motion Controller to provide the 3D ground truth of the color images. InterHand2.6M is the first large-scale real-captured dataset (2,590,34) with accurate ground truth 3D interacting hand poses, of which we only use the single hand part during our experiment. To unify the definition of joints, following the practice of [11], we first re-order the joint number to the same with SMPLX [35] hand model skeleton hierarchy, then move the root joint from palm center to the wrist of the above datasets as well as rescale the coordinates according to the length between 4th and 5th key joint of the template hand. Aside from the dataset mentioned before, we use Dexter+Object (DO) [39] for validation, which owns 5 key fingertips annotation only.

4.1.2 Evaluation Metrics

For each dataset, we calculate the percentage of correct 3D keypoints (PCK) under different thresholds (range from 20mm to 50mm) and calculate the corresponding Area Under Curve (AUC) for PCK. Additionally, the Mean-Per-Joint-Position-Error (MPJPE) [18] is used to measure the Euclidean distances between the ground truth joints and the predicted joints. Each error metric is computed for the root-relative 3D pose. Finally, to validate the effectiveness of pose length regularization, we report the acceleration error followed by Kanazawa *et al*. [22]. Acceleration error (mm/s^2) is the mean difference between ground-truth and predicted 3D acceleration for every joint.

4.1.3 Detail Settings

SCAT is build on top of PyTorch [34], we input the RGB image of $224 \times 224 \times 3$ augmented with motion blur and random rotation (± 30 degree), after normalize the augmented data, we use trainable ResNet-50 [15] pretrained on ImageNet as CNN backbone as depicted in Figure 2. $\mathbf{F}$ is obtained through apply a 1×1 convolutional layer to *layer2.3.bn3*, which downgrade the feature map of *layer2.3.bn3* from $\mathbb{R}^{N \times 512 \times 28 \times 28}$ to $\mathbb{R}^{N \times 21 \times 28 \times 28}$. Feature Vector $\mathbf{X}$ is produced by the fully-connected (FC) layer of ResNet-50 leave out of the last classification FC layer. The head of all Transformer Encoder layers are 8, the head dimension of each Encoder layer in the Fig 2 are 128, 64,

32, which accompanied with the shrinkage of feature dimension from 784 to 3. We use `ResNet@head8` with positional encoding by default and report the fine-grained predictions if there is no explicit statement.

Adam optimizer [24] (lr = 5e-4) along with Warmup tactic [13] are adopted to the training process. R is initialized with 5 where t is (0, 0).

4.2. Main Results

Here, we compare SCAT with the state-of-the-art methods on STB, RHD and DO. STB and DO are used to test on real-world data, while RHD is for synthetic. We use STB for estimating hand pose under natural conditions, DO for hand-interaction scenes (partial occlusion is the typical case), RHD for test on rendered images to verify SCAT's generalization performance.

Following the mixing strategy of [37, 50], we first train SCAT on synthetic dataset RHD and then train on the rest 5 datasets to make our model generalizes well to real-world applications. Finally, we finetune our model on the specific dataset and achieve a competitive result as depicted in Table 1. Different from the methods to be compared, SCAT rank almost the best in these three datasets, which proves its stability and high performance. Besides, it is also evident that with the help of the coarse-to-fine mechanism provided by auto-regressive regressor, a pronounced lift which based on the coarse prediction C_{coarse}, is achieved as C_{fine}.

Method	AUC of PCK ↑		
	DO [39]	STB [49]	RHD [51]
Ge *et al*. [11]	-	**0.998***	0.920
Yang *et al*. [47]	-	0.996	0.943
Baek *et al*. [2]	0.650	0.995	0.926
Z & B [51]	-	0.948	0.675
Xiang *et al*. [46]	0.912	0.994	-
Zhou *et al*. [50]	0.948	0.898	0.856
Spurr *et al*. [38]	0.820	-	0.920
Rong *et al*. [37]	-	0.992	0.934
Li *et al*. [26]	0.860	0.996	**0.960***
Ours$_{coarse}$	0.892	0.977	0.915
Ours$_{fine}$	**0.951**$^{\triangle,*}$	**0.994**$^{\triangle}$	**0.954**$^{\triangle}$

Table 1. Comparison with state-of-the-art methods on three public datasets. Here, superscript $\triangle$ for our result where $*$ for best, '-' demonstrate for those who did not report the results.

In order to verify the generalization performance between our model and current methods, we qualitatively compare SCAT with Frankmocap as depicted in Fig 3. While Frankmocap is a model-based method that relies on MANO [36], SCAT better estimates the 3D hand pose without reliance on MANO and other parametric 3D models.

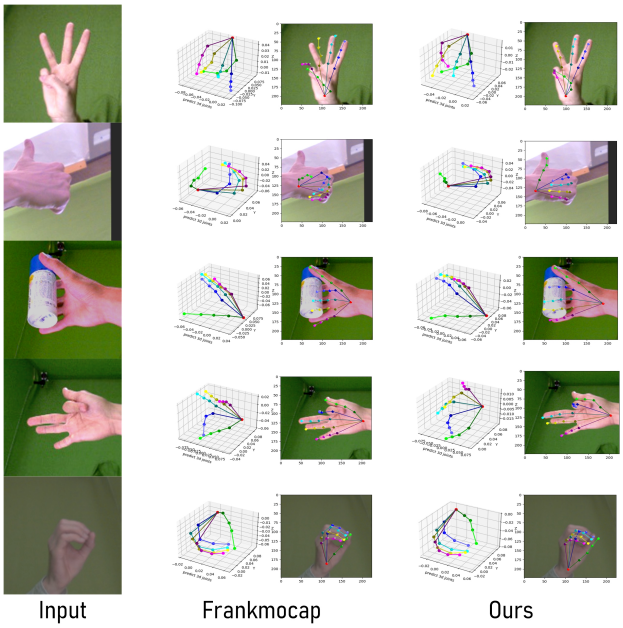

Figure 3. Comparison to the current method Frankmocap [37]. The noteworthy thing is that both models we used are trained on a mixture of the above 6 datasets rather than finetune on a specific dataset. We found SCAT generalizes well in unseen data.

4.3. Ablation study

Structure with different settings: We want to dig deep into the network structure to achieve top performance on in-the-wild data. First, we try our transformer encoder with different head numbers. Apart from transformer structure, whether the choice of CNN backbone is a central factor in obtaining a successful pose estimation process also appeals us. Here, we use ResNet-50 [15], HRNetW24 and InceptionV3 for ablation test. All backbones are pre-trained on the 1000 class image classification task of ImageNet. Furthermore, we are also interested in the positional encoding, so we conduct experiment without positional encoding for comparison. We observe SCAT achieves competitive performance on FreiHAND.

In Table 3, we found positional encoding (PE) is vital for SCAT, `ResNet@head8` with PE outperforms its no PE version by a large margin: 8.591 mm in MPJPE and 6.9% in AUC of PCK. Besides, it is evident from Table 3 that the number of head of transformer encoder is the most important factor in obtaining a satisfactory result, the larger the number of the head is, the higher precision and lower MPJPE achieved. Moreover, the choice of CNN backbone contribute to different precision level, the optimal performance was achieved by `IncepV3@head8` with PE.

Pose Length Regularization: One of our SCAT's significant contributions is the pose length regularization (PL), which ensures a smooth transition between consecutive

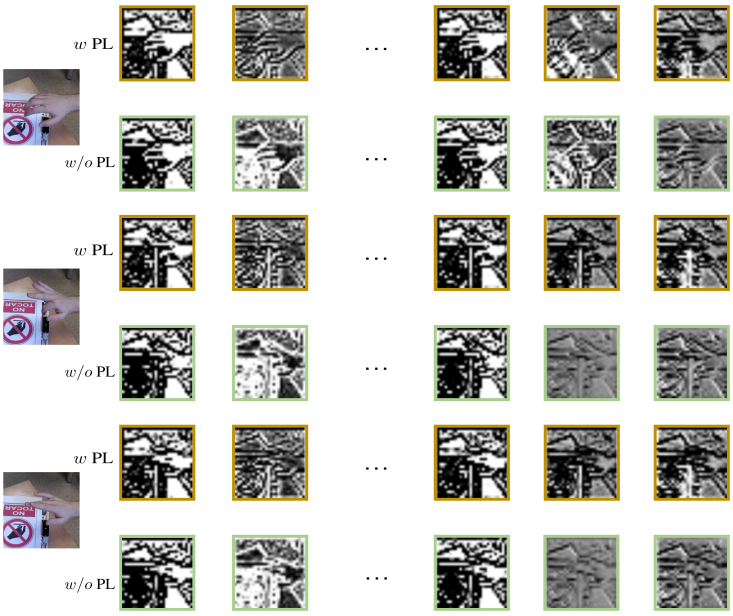

Figure 4. Ablation study of Pose Length (PL) Regularization. To qualitative validate PL, we visualize the feature maps **F** train with PL (brown) and without PL (green) under the same structure. It is rather obvious that PL empowers SCAT and produces a smooth and distinct contour of gesture in its feature maps while SCAT without PL yields the inferior result: people can easily observe the fuzzy and blurry pattern in its feature maps in the last three columns.

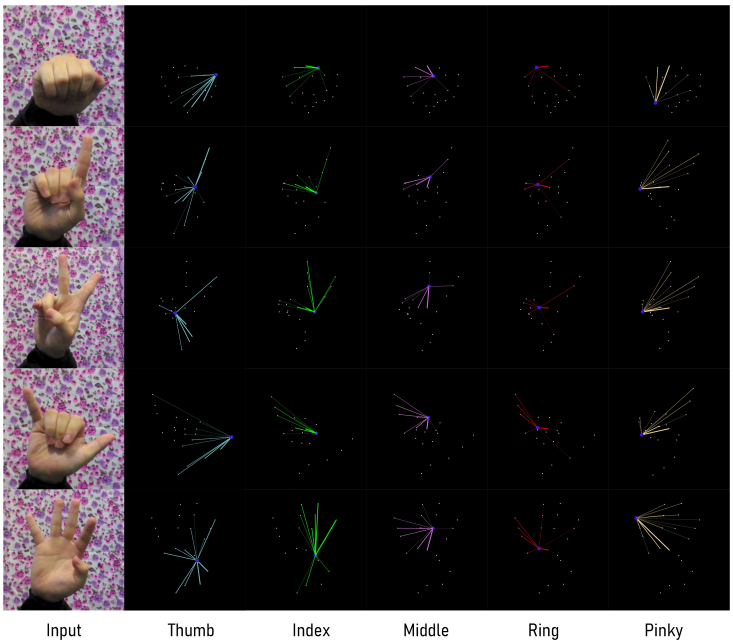

Figure 5. Qualitative results of SCAT on STB test set. As defined in Fig 1, we visualize the attention weights between the specified joint in each finger with the rest 20 joints, where brighter color indicates stronger attention.

frames through imposing a consistent stride constraint between feature maps **F** and 3D offsets **O**. To demonstrate the effectiveness of PL, we conduct experiments on the dataset with continuous image clips: MHP, FreiHand. According to Table 2, PL reduces the acceleration metric markedly, which helps yield a smooth pose estimation result without the addition of any temporal priors. Model with PL helps to reduce MPJPE and increase AUC. Furthermore, as depicted in Fig 4 without PL regularization (green contour), the estimation result is fragile and unstable with time progress.

Method	Accel ↓		MPJPE ↓		PA-MPJPE ↓		AUC ↑	
	HO-3D	MHP	HO-3D	MHP	HO-3D	MHP	HO-3D	MHP
w/o PL	6.78	11.110	3.02	**4.729**	1.50	3.611	0.794	0.918
w PL	**4.40**	**7.709**	**2.99**	4.985	**1.43**	**3.522**	**0.803**	**0.922**

Table 2. Ablation experiment on temporal-consistent datasets with pose length regularization (PL). We test HO-3D with submitting result to CodaLab while MHP is test with provided evaluation test (1,1524 images with 3D annotations).

Method	MPJPE ↓	AUC of PCK ↑
w/o PE, ResNet@head8	15.454	0.890
w PE, ResNet@head2	22.382	0.847
w PE, ResNet@head4	14.299	0.901
w PE, ResNet@head8	6.863	0.959
w PE, HRNet@head2	21.105	0.855
w PE, HRNet@head4	13.726	0.913
w PE, HRNet@head8	6.914	0.946
w PE, IncepV3@head2	18.551	0.877
w PE, IncepV3@head4	11.329	0.920
w **PE, IncepV3@head8**	**6.702**	**0.965**

Table 3. Ablation experiments on FreiHAND dataset.

Method	MPJPE (mm) ↓	AUC of PCK ↑
w/o MJM	13.559	0.920
w 10% MJM	**12.215**	0.932
w 20% MJM	12.908	**0.937**
w 30% MJM	13.721	0.922
w 40% MJM	16.230	0.876
w 50% MJM	23.355	0.809

Table 4. Ablation of the Masked Joint Modeling (MJM) objective with different percentages of masked input **J** to transformer encoder. n% indicates we mask randomly from 0% to n% of input joint **J**.

With PL regularization (brown contour), SCAT is capable of maintaining the inter-frame consistency.

Masked Joint Modeling for Occlusion scenes: It appeals to almost everyone in this field to identify whether a new objective helps solve the occlusion problem since the self-occlusion (due to different viewpoints) or partial occlusion (caused by interacting with other objects) are the two stumbling blocks in the pose estimation area. So we borrowed the masked joint modeling (MJM) from the NLP field and tried to use it to enhance our SCAT's generalization performance on the occlusion scenes. Here, we conduct our ablation study on the HO-3D dataset because HO-3D mainly focused on the interaction between hands and objects; thus, there are various type of occlusion inside. Table 4 shows that an appropriate proportion of masking joint assists in elevating metrics, though we observe a distinct drop in both MPJPE and AUC while the masked rate above 30%.

Relationship between 21 Hand Joints: To further understand SCAT's capacity in learning interactions among joints, we dive into the self-attentions in the transformer encoder. Fig 5 shows the correlations of key joints. Each row displays a specific input image and the relationship between five representative joints according to the definition by Fig 1. According to the extensive experiments on in-the-wild data and datasets mentioned before, we found the self-attention mechanism, which derives from transformer encoder, proves its powerful non-local modeling ability in relevant the remote joint to the current joint. Take the middle point (index 10, blue dot in the Middle column in Fig 5) for illustration.This joint has the most substantial ties with its parent joint (index 9) and closes up with the ring's upper joint, which follows the articulated relations. Moreover, we found each finger has an independent scope that affects only the interested related joints except for the thumb, which follows the biomechanical analysis of [38] to some extent.

5. Conclusion

We propose a simple yet effective method for 3D hand pose estimation from a single RGB image, **SCAT**: **S**tride **C**onsistency with **A**uto-regressive and **T**ransformer. By utilizing a simple mean shape of a template hand mesh and the strong correlation modeling capacity bring from the transformer encoder, a reasonable and reliable 3D hand pose is predicted. To the best of our knowledge, we are the first to come up with novel pose length regularization in the pose estimation field to ensures a smoother prediction through time went on, without any temporal priors needed, which greatly enhanced our frame-based pose estimation method. Moreover, we use an auto-regressive regressor for fine-grained prediction in a coarse-to-fine manner, a popular and effective practice to boost performance. We also propose the masked joint modeling (MJM) to enhance SCAT's robustness in self-occlusion and partial occlusion scenes. Experimental results show that our SCAT achieves competitive results compared with the state-of-the-art methods on mainstream hand datasets. Finally, without the need for any complex kinematic priors as well as inverse kinematic powered post processing methods, all the results are obtained through optimizing SCAT in an end-to-end manner, which solves 3D pose estimation in a simple and convenient way.

References

[1] Andreas Aristidou, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir. Inverse kinematics techniques in computer graphics: A survey. In *Computer Graphics Forum*, volume 37, pages 35–58. Wiley Online Library, 2018.

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019.

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[7] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787. Springer, 2020.

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint arXiv:1606.09375*, 2016.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] 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.

[11] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

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

[13] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[14] 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.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification, 2021.

[17] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018.

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[19] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018.

[20] Reza N Jazar. *Theory of applied robotics: kinematics, dynamics, and control*. Springer Science & Business Media, 2010.

[21] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[22] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[23] 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.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[26] Moran Li, Yuan Gao, and Nong Sang. Exploiting learnable joint groups for hand pose estimation. *arXiv preprint arXiv:2012.09496*, 2020.

[27] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. *arXiv preprint arXiv:2012.09760*, 2020.

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[29] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers, 2021.

[30] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *arXiv preprint arXiv:2008.03713*, 2020.

[31] 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 (ECCV)*, 2020.

[32] 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.

[33] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[36] 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.

[37] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020.

[38] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. *arXiv preprint arXiv:2003.09282*, 8, 2020.

[39] 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.

[40] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013.

[41] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[44] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[45] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297. IEEE, 2012.

[46] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019.

[47] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2335–2343, 2019.

[48] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Towards explainable human pose estimation by transformer. *arXiv preprint arXiv:2012.14214*, 2020.

[49] 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.

[50] 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.

[51] 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.

[52] 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.