

# Pose-Guided Temporal Enhancement for Robust Low-Resolution Hand Reconstruction

Kaixin Fan, Pengfei Ren,<sup>\*</sup> Jingyu Wang,<sup>†</sup> Haifeng Sun, Qi Qi, Zirui Zhuang, Jianxin Liao  
State Key Laboratory of Networking and Switching Technology,  
Beijing University of Posts and Telecommunications

{fkxbupt, rpf, wangjingyu, hfsun, qiqi8266, zhuangzirui, liaojx}@bupt.edu.cn

## Abstract

3D hand reconstruction is essential in non-contact human-computer interaction applications, but existing methods struggle with low-resolution images, which occur in slightly distant interactive scenes. Leveraging temporal information can mitigate the limitations of individual low-resolution images that lack detailed appearance information, enhancing the robustness and accuracy of hand reconstruction. Existing temporal methods typically use joint features to represent temporal information, avoiding interference from redundant background information. However, joint features excessively disregard the spatial context of visual features, limiting hand reconstruction accuracy. We propose to integrate temporal joint features with visual features to construct a robust low-resolution visual representation. We adopt Triplane Features, a dense representation with 3D spatial awareness, to bridge the gap between the joint features and visual features that are misaligned in terms of representation form and semantics. Triplane Features are obtained by orthogonally projecting joint features, embedding hand structure information into the 3D spatial context. Furthermore, we compress the spatial information of the three planes into a 2D dense feature through Spatial-Aware Fusion to enhance the visual features. By using enhanced visual features enriched with temporal information for hand reconstruction, our method achieves competitive performance at much lower resolutions compared to state-of-the-art methods operating at high resolution on DexYCB, HanCo and H2O. Code is available at <https://github.com/NewbieFan/Temp-LowRes-hand>.

## 1. Introduction

3D hand reconstruction, a crucial component in non-contact human-computer interaction applications, has made signifi-

<sup>\*</sup>Equal contribution.

<sup>†</sup>Corresponding author.

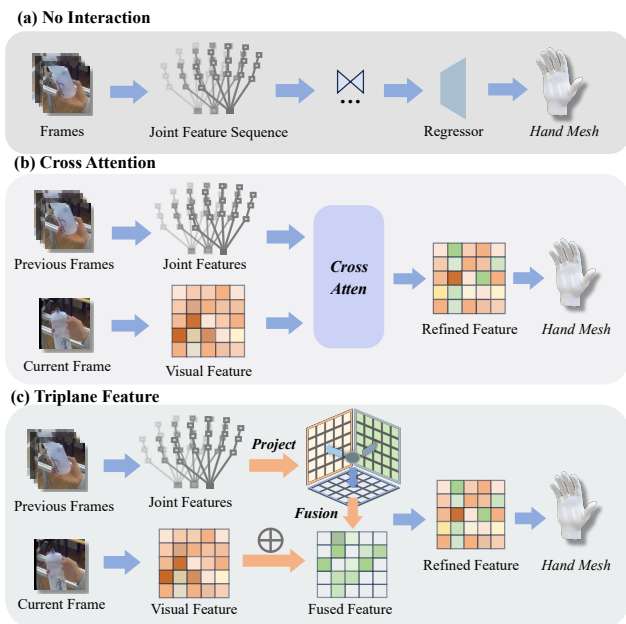


Figure 1. Different way to utilizing joint features. a) Directly reconstruct hand from joint features, relying solely on sparse features without spatial context. b) Fusing joint and visual features through cross-attention, which avoids representation discrepancies but neglects the hand’s structural priors. c) Using triplane features to bridge joint and visual features by decoding structural joint information into dense representations that enhance visual features.

cant advancements in recent years with the advent of large-scale datasets and deep learning technologies. Many studies [6, 7, 30, 31, 34, 40] have focused on reconstructing 3D hands from a single RGB image. In typical interaction scenarios, these methods require the subject to stay close to the capture camera to obtain high-resolution (e.g., 256×256) cropped hand images. However, as the interaction distance increases—such as in XR teaching activities within a classroom—the available cropped hand images are limited to lower resolutions (e.g., 64×64). Current methods

perform poorly on such low-resolution images, as these images provide fewer cues for addressing joint self-similarity and depth ambiguity, significantly increasing the difficulty of 3D hand reconstruction.

To address the low-resolution problem, one intuitive approach is to lift the resolution through super-resolution techniques [14]. However, general super-resolution methods can lead to structural distortions and artifacts in the hand regions, which brings a domain gap between super-resolved image and real high-resolution image. Low-Res Pose [45] adopted confidence-aware learning to supervise continue joint offset fields, thereby alleviating the quantization errors. RSC-Net [48] adopted cross-resolution contrastive learning to align the estimation results of low-resolution images with high-resolution images. However, the performance of these methods is still limited due to the inherent ambiguity resulting from information loss in low resolution.

A single low-resolution image lacks detailed hand appearance, leading to ambiguous visual representation. Continuous frames provide comprehensive hand appearance and motion context, offering strong disambiguation cues for inferring global hand structure. To reduce interference from redundant background information and computational costs, existing temporal methods typically compress single-frame information into joint features [51, 58], and then model the intra- and inter- frame correlation between joints through temporal interaction. However, joint features excessively discard spatial context of visual features, which limits the accuracy of pose estimation. Integrating the spatial context of visual features with the temporal context of joint features can construct robust visual representation with 3D spatial awareness. However, integrating joint features into visual features present two major challenges. Firstly, sparse joint features are not aligned with dense visual features in terms of representation form. Secondly, joint features contain prior geometric structural information in 3D space (e.g., hand scale, skeleton length), while visual features represent detailed appearance information in 2D space (e.g., texture, color), leading to a semantic misalignment.

In this paper, we integrate the temporal clue from historical frames into the current visual features to achieve robust and accurate 3D hand pose estimation under low-resolution scenarios. Specifically, we compress temporal information into joint representations, and model intra- and inter frame relationships of joints through spatio-temporal interaction. Then, we adopt triplane features[4], a 3D spatial-aware dense visual representation, to bridge the sparse 3D joint representation and the dense 2D visual representation. We orthogonally project the joint features into three individual planes, with each plane independently embedding the sparse joint features into specific spatial contexts. The orthogonal planes ensure comprehensive coverage of 3D hand structure information in a low-redundancy manner. How-

ever, they still contain redundant spatial information, directly integrating these features into the visual plane leads to the overlap of spatial context from different planes. Thus, we compress the features from two planes orthogonal to the visual plane into 1D global features, guiding each pixel in visual features to integrate distinct spatial context from the other planes, reducing redundancy and clarifying the 3D spatial context. Finally, we reconstruct the hand based on the enhanced visual features.

Experimental results on several large-scale hand pose estimation video datasets, DexYCB [5], HanCo [60] and H2O [25], show the proposed method reach competitive performance at much lower resolution compared to both single-frame based and temporal based state-of-the-art (SOTA) methods at high resolution. In summary, our contribution is list as below:

- We propose a temporal enhancement framework that integrates temporal joint features with visual features to construct a robust visual representation for low-resolution conditions, thereby improving hand reconstruction performance in low-resolution videos.
- We introduced triplane feature encoding to encode joint features into dense feature maps. Through triplane feature, visual features can perceive the prior hand structure information from the temporal joint features.
- Our method reaches competitive result at much lower resolution compared to SOTA methods at high resolution.

## 2. Relate Work

### 2.1. Single Image based Hand Pose Estimation

Early works [2, 17, 18, 36, 43, 59] focused on estimating hand pose solely from input images, which suffered from occlusion issues due to the lack of modeling the inherent structure of the hand. MANO [41] introduces a parametric hand model, which contains rich priors of human hands, and now most work focus on reconstruct this pre-defined hand from image. Some works [1, 23, 32, 37–39, 52, 56] directly predict parameters of MANO, 3D joints and mesh vertices coordinates are then retrieved using the differentiable MANO model. However, predicting MANO parameters directly from a single RGB image is a challenging job, which is difficult for network optimization and is prone to mesh-image misalignment [44], especially in a low-res input condition. To alleviate these issues present in parametric predicting, some works [13, 24, 27] regress the 3D coordinates or occupancy directly for mesh vertices. For example, [53] adopts graph convolution neural network to regress 2.5D Heatmap for every mesh vertices. [26] proposed an End to end framework to recover mesh using transformer. These works shows that using a regression-based approach can maintain flexibility while achieving comparable results to parametric methods. Thus, we adopt a regression-based

method described in [53] to reconstruct the hand mesh from enhanced visual features.

## 2.2. Low Resolution Pose Estimation

Recently, a few works have explored on low resolution pose estimation. A intuitive way is leveraging Super Resolution approach [14] to improve the image quality and pose estimation performance. However, the addition super-resolution network brings extra computation cost and may bring unpleasant artifacts to the input image. [45] find that offset-based methods can provide more reliable results in low resolution, and further proposed a confidence-aware learning to selectively weigh the learning of heatmap and offset. [48] controls the scaling of the CNN architecture model using a set of hyperparameters, enforcing it to output scale-invariant features at different resolutions through contrastive learning and self-supervised learning. Although specific learning schemas improve the poor results in low resolution, the features extracted by the network are still constrained by the lack of detailed hand appearance. Other work [33] explored on 3D pose estimation in low-resolution multi-view videos, showing that temporal and multi-view information could be a supplement to alleviate information loss. However, they only correct the prediction under certain adverse conditions, ignoring temporal information can reduce the inherent ambiguity in single low-resolution image and improve the overall performance.

## 2.3. Temporal Hand Pose Estimation

Some recent methods [11, 15, 22, 29, 50, 55] have begun to explore leveraging temporal information in videos to overcome the limitations of single-view images. Hasson et al. [15] use photometric supervision to keep consistency between the re-projected 3D hand and the optical flow. VIBE [22] use a temporal generation network to predict model parameters for each frame, and employs a recurrent network as a discriminator to supervise motion sequence. However, they ignored the spatial and feature correlations in the video. More recently, [20] has harnessed temporal information by propagating frame-wise features with a sequence-to-sequence model based on LSTM [16], but it only provides sparse joint location predictions. TCMR [9] introduces a temporally consistent system that synthesizes implicit features from past and future frames to maintain temporal consistency in the mesh of the current frame. Deformer [12] deforms the hand poses from neighborhood frames using predicted motion to avoid unreliable queries generated with local feature when leveraging self-attention mechanism.

In low resolutions, these methods are limited to the poor quality of single frame features. Additionally, they lack a focus center when processing temporal features, which may cause some estimation results to deviate from the actual image due to the influence of temporal context.

## 3. Method

The overview of our method is shown in the Fig. 2. For a given frame  $I_T$  in an image sequence  $V = \{I_t\}_{t=1}^T$ , we extract temporal cues from the previous  $T - 1$  frames to enhance the visual feature of the current frame and reconstruct the hand mesh from the enhanced visual feature. Specifically, we first compress the information of each historical frames into joint features and establish their inter- and intra-frame correlations through a spatio-temporal Transformer. Then, we encode the structural information from the enhanced joint features into 3D spatial-aware dense representations through triplane feature encoding to enhance the visual features of the current frame. Finally, we leverage the enhanced visual feature to reconstruct the hand mesh.

### 3.1. Static Visual Features

As shown in Fig. 3, we adopt a Vision Transformer (ViT) [10] to extract visual features. During the extraction process, ambiguous local features would propagate from low-level features to subsequent high-level features, leading to ambiguities in both semantic content and fine-grained details in the final visual features. Thus, we incorporate temporal information into intermediate features at different levels to gradually reduce ambiguous local representations and ultimately obtain visual features with a clear perception of hand structure. However, static visual features lack the ability to perceive dynamic temporal information represented in joint features. Therefore, we introduce pose-aware tokens as anchors for fusing dynamic temporal information into a static representation, which are used for further interaction with visual features. These tokens capture hand pose and appearance information from the current visual feature through attention mechanism. We categorize the functionalities of these tokens based on finger semantic to ensure they incorporate both fine-grained visual details (e.g., texture) and high-level semantic content (e.g., skeletal structures).

### 3.2. Dynamic Temporal Joint Features

We use compact joint features to represent dynamic temporal information from historical frames. Give a single frame, we obtain its visual feature map along with pose-aware tokens through the backbone. Each pose-aware token carries semantic information corresponding to different individual fingers. We further combine local visual features to expand each token into a complete joint feature, enabling fine-grained temporal interaction around each joints. We adopt a FC layer to predict 2D joint coordinates  $p \in \mathbb{R}^{J \times 2}$  from pose-aware tokens, and use it to sample joint-centric local features  $J^{sample} \in \mathbb{R}^{J \times C}$  from the visual feature map. The joint coordinates is encoded as pose feature  $J^{pose}$ , and then add to the sampled feature  $J^{sample}$  to obtain the initial joint feature  $J^{init} = J^{pose} + J^{region}$ . We concatenate the joint features from each frame in the upstream sequence

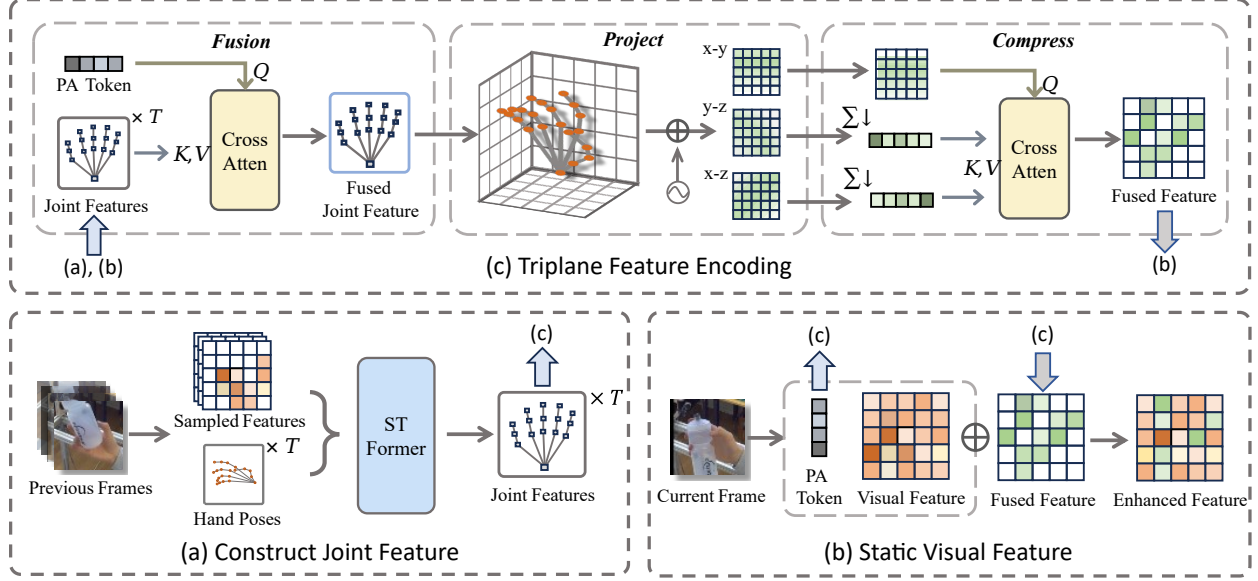


Figure 2. **Overview of Our Framework.** We first construct joint features from the sampled past frame features and 2D poses using a Spatial-Temporal Transformer (a), capturing the temporal dynamics and spatial relationships between joints. We then use these joint features to enhance the visual features of the current frame (b) by converting them into spatial-aware dense features through Triplane Feature Encoding (c). In (c), we fuse different joint features (from a) using pose-aware tokens (from b), and project them into three orthogonal planes to obtain the Triplane features. These features are then compressed into a single plane through space-aware fusion.

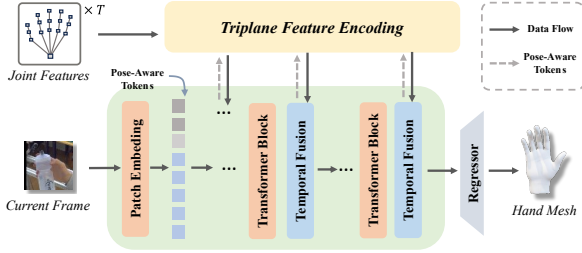


Figure 3. Backbone of our framework.

to form the initial representation of temporal information  $J^{temp} = \text{Concat}(J_t^{init}) \in \mathbb{R}^{T \times J \times C}$ . Then, we adopt a spatio-temporal Transformer to establish inter- and intra-frame correlation between these joints. The spatio-temporal Transformer is a combination of alternately stacked temporal and spatial attention modules. In the spatial attention module, attention interactions occur among different joints within the same joint feature:

$$S\text{-MHSA}(J^{temp}) = \text{MHSA}(Q_j, K_j, V_j) \quad (1)$$

$$Q_j = W_q J_j^{temp}, K_j = W_k J_j^{temp}, V_j = W_v J_j^{temp} \quad (2)$$

In the temporal attention module, attention interactions occur among the same joint within different joint features:

$$T\text{-MHSA}(J^{temp}) = \text{MHSA}(Q_t, K_t, V_t) \quad (3)$$

$$Q_t = W_q J_t^{temp}, K_t = W_k J_t^{temp}, V_t = W_v J_t^{temp} \quad (4)$$

Here,  $W_q$ ,  $W_k$ , and  $W_v$  are learnable weight matrices used to transform the input features into queries, keys, and values, respectively.  $J_j^{temp}$  represents the joint features for the  $j$ -th joint across all time steps, while  $J_t^{temp}$  represents the joint features for the  $t$ -th time step across all joints.

### 3.3. Triplane Feature Encoding

Joint features contain temporal context of the current hand pose, which can be integrated into visual features to construct low-resolution robust visual representations. However, it is difficult to fuse joint features with visual features that are not aligned in semantic content and representation form. A straightforward approach is to treat joint features as global tokens and interact them with visual features through cross attention. However, visual features struggle to effectively perceive the 3D spatial properties of joint features through attention mechanisms. Thus, we propose to use triplane features as an intermediate representation to enhance the visual features. Unlike [4, 42, 46] that constructs dense Triplane features from images for spatial indexing, we leverage sparse joint features to generate triplane features, embedding 3D geometric attributes of the skeleton into specific spatial context.

Specifically, we first copy each pose token to the corresponding finger joints as a Pose Template  $J^P \in \mathbb{R}^{J \times C}$ , and use it to fuse dynamic joint features into a static one. Then splat it onto three orthogonal planes as an intermediate representation for interaction with visual features. Fur-

thermore, we integrate the information from different planes into a single plane through spatial-aware fusion, allowing for direct fusion with the visual features.

**Triplane Projection of Joint Features** Visual features represent specific spatial information in the form of dense representations. Transferring joint features into the 2D visual feature space through 3D-2D pose projection [40] can provide effective disambiguation cues for hand poses. However, a single projection plane cannot capture the complete 3D hand structure implied in joint features, resulting in the loss of crucial structural disambiguation cues. Therefore, we propose to project joint features onto three orthogonal planes to fully preserve 3D hand structure information. We first decode the 3D hand pose from the fused joint features, and then use it to generate distance-aware heatmaps corresponding to the x-y, y-z, and z-x planes. Then, we splat the fused joint feature across the three planes through these heatmaps, and collect the features from each plane as the Triplane Feature maps  $F_{triplane} = \{F_{x,y}, F_{x,z}, F_{y,z}\}$ .

**Spatial-Aware Fusion** Triplane encoding transforms the fused joint feature into dense 2D features, eliminating the misalignment in representation form with visual features. However, the initial plane features derived from projecting sparse joint information primarily encode geometric positions while overlooking the relative spatial relationships and interactions between neighboring joints, such as occlusions or depth variations, which are crucial for resolving structural ambiguities in visual features. Additionally, these plane features are spatially non-unified: The x-y plane shares the same spatial dimension with visual features, while the y-z and x-z planes are spatially orthogonal to visual features. To embed the complete 3D hand structure information represented by the Triplane features into the visual features in 2D space, it is essential to compress the y-z and x-z plane features as supplementary depth information relative to the x-y plane into the pixel-wise semantic content. However, directly integrating the y-z and x-z plane features into the x-y plane via cross-attention is inefficient and introduces redundant information. Therefore, we compress the y-z and x-z features along distinct directions into 1D global representations. This compression allows us to guide each pixel in the x-y plane to selectively integrate spatial context from the other two planes, reducing unnecessary overlap while preserving key spatial dependencies.

Specifically, we first add 2D positional encoding to each plane feature and apply convolution layers to each plane feature to capture local spatial information (See Fig. 2 (c) *project*). Next, we perform summation along different directions for  $F_{x,z}$  and  $F_{y,z}$  to obtain the 1D features  $I_x, V_x$  and  $I_y, V_y$  (See Fig. 2 (c) *compress*), where  $I$  denotes summation along the Z direction, and  $V$  denotes summation along the X or Y direction. Finally, we integrate these

features into the pixel-wise content of  $F_{x,y}$  through cross-attention to obtain spatially compressed features.

### 3.4. Optimization

We supervise hand mesh regression through reconstruct loss, supervise temporal enhancement process through token pose loss and temporal pose loss.

#### 3.4.1. Reconstruct Loss

We used the same reconstruction loss as [53], including vertex/joint loss and mesh smoothing loss. Vertex/Joint loss supervises the prediction error of each mesh and joint point through L1 loss, i.e

$$L_{V/J} = \sum_{i=1}^N \|\nu_i - \nu_i^*\|_1 + \sum_{i=1}^K \|j_i - j_i^*\|_1 \quad (5)$$

where  $\nu_i$  is the  $i^{th}$  vertex of the mesh,  $j_i$  is the  $i^{th}$  joint of the pose, and  $*$  denotes the ground-truth.

Mesh smooth loss includes supervision of the surface normal vectors and edge lengths to obtain visually reasonable results. The loss functions for normal vector and edge length are defined as follows:

$$L_{normal} = \sum_f \sum_{\{i,j\} \subset f} \left| \left\langle \frac{\nu_i - \nu_j}{\|\nu_i - \nu_j\|_2}, n_f^* \right\rangle \right| \quad (6)$$

$$L_{edge} = \sum_f \sum_{\{i,j\} \subset f} \left| \|\nu_i - \nu_j\|_2 - \|\nu_i^* - \nu_j^*\|_2 \right| \quad (7)$$

where  $f$  and  $n_f$  indicate a mesh face and unit normal vector of face  $f$ , respectively.  $n_f^*$  is computed from ground-truth vertex sets  $V^*$ , where  $*$  denotes ground-truth.

The final hand reconstruction task loss combines the vertex/joint loss and the mesh smoothness loss as  $L_{hand}$ .

#### 3.4.2. Pose Loss

For pose-aware tokens, we enable them to perceive pose changes by supervising their predicted 2D poses. During the temporal fusion process, we use the fused joint feature to predict 2.5D hand pose coordinates for triplane project. We supervise the predicted 2.5D coordinates at each temporal fusion stage to ensure accurate decoding of hand structure information. These two loss functions are represented as follows:

$$L_{token} = \sum_i^K \|j_i^{2d} - j^{2d*}\|_1 \quad (8)$$

$$L_{temporal} = \sum_{t=1}^T \sum_{i=1}^K \|j_{t,i}^{2.5d} - j^{2.5d*}\|_1 \quad (9)$$

The final training loss of our framework, which combines reconstruction loss and pose loss, can be expressed as follows:

$$L = L_{hand} + L_{token} + L_{temporal} \quad (10)$$

| Methods            | Input     | Resolution | PA-MPJPE(↓) | PA-MPVPE(↓) | AUC(↑)      | F@5(↑)      | F@15(↑)     |
|--------------------|-----------|------------|-------------|-------------|-------------|-------------|-------------|
| METRO[26]          | Monocular | 256        | 6.99        | -           | -           | -           | -           |
| Liu et. al.[28]    | Monocular | 256        | 6.60        | -           | -           | -           | -           |
| MeshGraphormer[27] | Monocular | 256        | 6.41        | -           | -           | -           | -           |
| HandOccNet[34]     | Monocular | 256        | 5.80        | 5.50        | 88.4        | 78.0        | 99.0        |
| H2ONet[47]         | Monocular | 128        | 5.70        | 5.50        | 89.1        | 80.1        | 99.0        |
| SimpleHand[57]     | Monocular | 256        | 5.50        | 5.50        | -           | -           | -           |
| VIBE[22]           | Sequence  | 224        | 6.43        | -           | 87.1        | -           | -           |
| TCMR[9]            | Sequence  | 224        | 6.28        | -           | 87.5        | -           | -           |
| Deformer[12]       | Sequence  | 512        | 5.22        | -           | 89.6        | -           | -           |
| HMR[19] + LR-ViT   | Monocular | 64         | 6.45        | 6.24        | 87.3        | 73.0        | 98.5        |
| VIBE[22] + LR-ViT  | Sequence  | 64         | 6.22        | 6.09        | 87.5        | 74.1        | 98.7        |
| <b>Ours</b>        | Sequence  | 64         | <b>5.72</b> | <b>5.52</b> | <b>88.5</b> | <b>78.4</b> | <b>98.7</b> |
| <b>Ours</b>        | Sequence  | 32         | 6.03        | 5.82        | 88.0        | 76.5        | 98.6        |

Table 1. Comparison with the state-of-the-art on the DexYCB dataset.

| Methods       | Input     | MPJPE        | PA-MPJPE    |
|---------------|-----------|--------------|-------------|
| Cho et al.[8] | Sequence  | 17.20        | -           |
| HandOccNet    | Monocular | 16.19        | 6.13        |
| HMR + LR-ViT  | Monocular | 30.72        | 7.49        |
| VIBE + LR-ViT | Sequence  | 27.89        | 7.05        |
| <b>Ours</b>   | Sequence  | <b>18.15</b> | <b>6.11</b> |

Table 2. Comparison with the state-of-the-art on the H2O dataset.

| Methods          | Input     | J-PA        | V-PA        | F@5         | F@15        |
|------------------|-----------|-------------|-------------|-------------|-------------|
| HaMuCo[54]       | Monocular | 7.00        | -           | -           | -           |
| EpipolarPose[21] | Monocular | 6.10        | -           | -           | -           |
| MobRecon[7]      | Monocular | 5.70        | 5.81        | -           | -           |
| HaMeR[35]        | Monocular | 5.99        | 5.61        | 77.8        | 99.0        |
| HMR + LR-ViT     | Monocular | 6.88        | 6.44        | 73.4        | 98.7        |
| VIBE + LR-ViT    | Sequence  | 6.79        | 6.27        | 75.0        | 98.8        |
| <b>Ours</b>      | Sequence  | <b>5.77</b> | <b>5.74</b> | <b>77.6</b> | <b>99.0</b> |

Table 3. Comparison results on the HanCo dataset. The evaluation results for HaMuCo and EpipolarPose using monocular input are derived from [54].

| Retrained | HanOccNet | SimpleHand | HaMeR | Ours* |
|-----------|-----------|------------|-------|-------|
| ×         | 6.63      | -          | 8.70  | -     |
| ✓         | 5.98      | 5.92       | -     | 5.72  |

Table 4. SOTA performance (P-MPJPE) on interpolated low-resolution images (DexYCB). \*Using raw low-resolution inputs.

## 4. Experiments

### 4.1. Datasets

**DexYCB** contains multi-camera videos of humans grasping 20 YCB [3] objects. It consists of 1000 sequences with 582K frames captured by cameras from 8 different views. During the data collection, the hands are frequently oc-

cluded during hand-object interaction. We use the dataset split setup s0 from [5] for the training and test sets.

**HanCo** consists of 1,517 videos with multiple views and camera calibration. It has 860,304 frames in total, i.e. 107,538 time-step per view. Since HanCo does not have an official train/test split, we follow [54] use the first 1,200 sequences for training and validation and the last 317 sequences for testing in all experiments for fair comparisons. We conduct all ablation studies on this dataset.

**H2O** includes data from four participants performing 36 distinct actions involving interactions with eight different objects. It provides markerless 3D annotations for both hands. We follow the setup in [25] to divide the training and test sets for pose estimation.

### 4.2. Implementation

**Handling Low Resolution** Our experiment was conducted under two original resolution settings,  $32 \times 32$  and  $64 \times 64$ . Unless otherwise specified, all experiments use the original resolution of  $64 \times 64$ . To avoid additional noise introduced by excessive image interpolation, we set the size of the input image to  $64 \times 64$ . We have modified ViTPose-B [49] as the backbone. To adapt it to low resolution inputs, we replaced the patch embedding part with two convolution layers with zero padding, and reduce the number of tokens in ViT to 64. These 64 tokens were rearranged into an  $8 \times 8$  feature map during the fusion of temporal features and regression of hand mesh. We adopt a regression-based method in [53] to reconstruct hand from the final feature.

**Handling Temporal Input** Our method utilizes the temporal information of historical frames to enhance the current hand reconstruction. We have chosen a sequence length of  $T=6$  with a 5-frames interval based on observations of each datasets. In training, we first extract the visual features and pose-aware tokens of the first 5 frames through the ViT

backbone without temporal fusion, then compressed them into joint features. During this phase, the parameters of the ViT backbone are not updated. These joint features are subsequently processed through spatio-temporal Transformer to enhance the feature extraction process of last frame. We train all network parameters in this phrase.

**Training** We train and evaluate our method on a single server with an Nvidia GeForce RTX 4090 GPU. All experiments are implemented with PyTorch. For all datasets, the batch size is set to 64. Before formal training on each dataset, we fine-tune the backbone using the pre-trained weights of ViTPose-B. We conducted 30 epochs of fine-tuning for DexYCB and HanCo, and 20 epochs for H2O. For DexYCB, H2O and HanCo, the model is trained for 50,30,50 epochs, respectively. The model is optimized with AdamW solver with parameters ( $\beta_1$  of 0.9,  $\beta_2$  of 0.999, and weight decay of 0.1). We use an initial learning rate of  $5e-4$  and cosine schedule. We perform data augmentation including scaling ( $\pm 10\%$ ), rotation ( $\pm 180^\circ$ ), translation ( $\pm 10 pixels$ ), and color jittering ( $\pm 20\%$ ).

### 4.3. Evaluation Metrics

We employ the standard metric for hand pose estimation: Mean Per Joint Position Error (MPJPE, J-PE) and Mean Per Vertex Position Error (MPVPE, V-PE), measured after Procrustes alignment, to measure the accuracy of the pose and shape of the estimated hand meshes. In addition, we report the area under the curve of PCK (AUC) score and F-score at 5mm (F@5) and 15mm (F@15) threshold. All evaluation metrics are calculated in the root-aligned space.

### 4.4. Comparisons with State-of-the-arts

We compared the proposed low-resolution method with state-of-the-art video-based and single-view methods that use high-resolution input. Additionally, we further use our backbone network (LR-ViT) to establish a low-resolution benchmark for comparison. We trained HMR and VIBE with our backbone on raw low-resolution images to explore the effectiveness of commonly used methods. Comparison results are shown in Tab. 1, Tab. 2 and Tab. 3. In DexYCB, our method performs equally or better than other high-resolution methods at  $32 \times 32$  and  $64 \times 64$  resolution settings, except for Deformer that uses  $512 \times 512$  input resolution. In HanCo and H2O, our method also demonstrates competitive results. We also report the performance of SOTA methods on upsampled low-resolution images in Tab. 4, including cases where models were retrained with upsampled images and where they were not.

### 4.5. Ablation Study

#### 4.5.1. Backbone settings

In this section, we experiment with different backbone settings and report their performance with and without tem-

| Token Settings | Monocular |      | Temporal |      |
|----------------|-----------|------|----------|------|
|                | J-PA      | V-PA | J-PA     | V-PA |
| 0              | 6.71      | 6.75 | -        | -    |
| 1              | 6.58      | 6.64 | 6.03     | 6.09 |
| 6 (SP)         | 6.19      | 6.19 | 5.77     | 5.74 |
| 21 (JP)        | 6.07      | 6.05 | 5.80     | 5.82 |

Table 5. Ablation study on backbone and pose-aware tokens.

poral enhancement. Under monocular input, the backbone skips temporal fusion and uses only transformer blocks to process the input. Our experiment further explored the impact of Pose-Aware Tokens. We adopted different settings of introduced tokens to test their impact. Under monocular input, we still supervise these tokens to predict 2D hand poses. The experiment result is shown in Tab. 5. The introduction of temporal enhancement universally improves performance. Under monocular input, the original ViT (w/o pose-aware tokens) performs poorly at low resolutions, while other networks generally achieve performance gains. This demonstrates that supervising intermediate features through specific tokens is more effective than solely employing end-to-end training on the final regression results. Meanwhile, due to the ability of finger based pose-aware tokens (SP) to jointly carry high-level semantic information and low-level visual details, it is more effective in temporal enhancement than a single pose token and skeleton based pose tokens (JP).

#### 4.5.2. Temporal Fusion Schemes

In this section, we explored the effects of different temporal fusion schemes. We consider three ways to fuse temporal features and visual features: cross attention, single plane encoding, and triplane encoding. The cross attention approach uses joint features as K and V, and visual features as Q to interact through attention mechanisms. The single plane encoding method uses 2D Heatmap to diffuse joint features into a single plane. The single plane feature is also embedded with 2D positional encoding and processed using convolutional layers. For the triplane encoding, we also tested the effect of directly integrating the full  $F_{x,z}$  and  $F_{y,z}$  with  $F_{x,y}$ . The comparison results are presented in Table 6. Our proposed triplane encoding performs best. Compared to the other two methods, the triplane encoding effectively extracts crucial 3D structural information, providing cues for eliminating hand structure ambiguity in visual features. Additionally, using compressed 1D features for interaction is superior to using complete feature.

#### 4.5.3. Position for inserting temporal fusion

In this section, we further investigated the impact of the insertion position of temporal fusion operations in the backbone. Specifically, we compared two strategies: one where

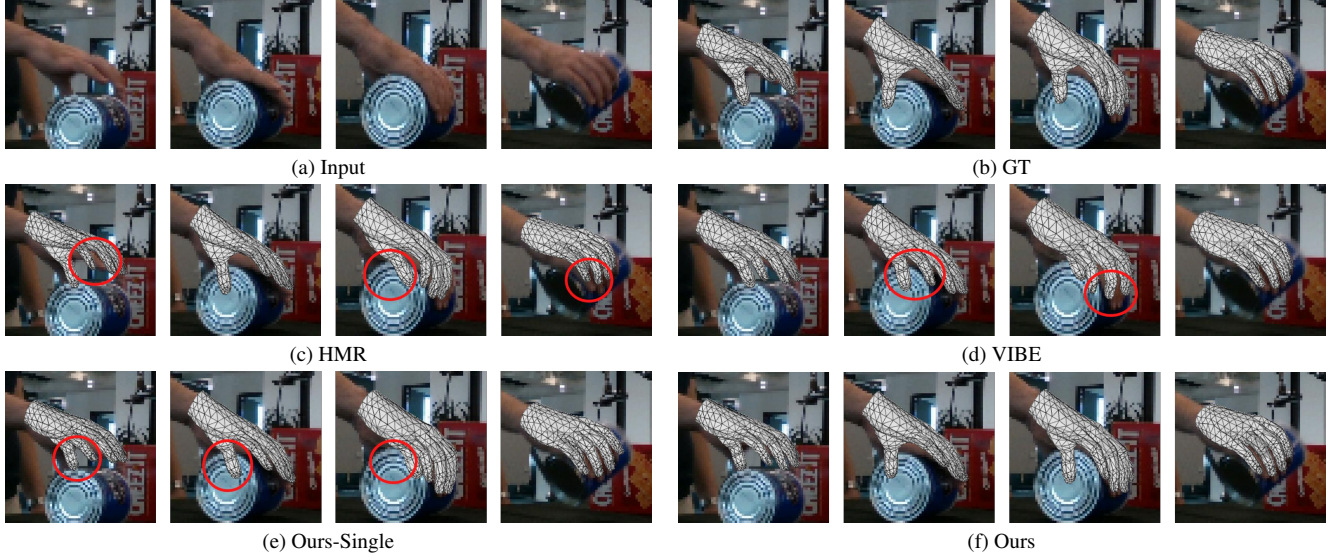


Figure 4. Qualitative Comparison Results on DexYCB. Regions with significant prediction errors are marked with red circles.

| Fusion Scheme                          | J-PA | V-PA | F@5  | F@15 |
|----------------------------------------|------|------|------|------|
| Cross-attention                        | 5.95 | 5.99 | 75.8 | 98.9 |
| Single plane Encoding                  | 5.85 | 5.90 | 76.7 | 99.0 |
| Triplane Encoding<br>(w/o 1D compress) | 5.80 | 5.85 | 76.9 | 99.0 |
| Triplane Encoding                      | 5.77 | 5.74 | 77.6 | 99.0 |

Table 6. Ablation study of different temporal fusion schemes.

| Insert Position | J-PA | V-PA | F@5  | F@15 |
|-----------------|------|------|------|------|
| Final layer     | 6.01 | 6.05 | 75.1 | 98.6 |
| Middle layer    | 5.77 | 5.74 | 77.6 | 99.0 |

Table 7. Ablation Study on the Position of Temporal Information Fusion in the Backbone Network.

temporal features are fused at multiple intermediate levels throughout the network, and another where fusion occurs solely at the final level, as shown in Tab. 7. The experimental results in Tab. 7 demonstrate that fusing temporal features at multiple intermediate levels leads to superior performance compared to only performing fusion at the final level. This suggests that incorporating temporal information at various stages of the network helps to better capture ambiguous local features, contributing to more accurate and consistent hand pose reconstruction in low resolution.

#### 4.6. Qualitative Results

We present the qualitative comparison results on DexYCB in Fig. 4. Compared with low resolution benchmarks (HMR, VIBE) that use the same LR-ViT as the backbone, our method achieves better alignment of the reconstructed

hands with low-resolution images. Additionally, by leveraging historic temporal information to supplement missing details in visual features, our approach enhances the ability to cope with local occlusions compare to solely rely on a single frame (see columns 2 and 3 in Fig. 4 (e) and (f)).

## 5. Conclusion

In this paper, we propose a temporal enhancement framework for low-resolution hand reconstruction. We represent temporal information through joint features to capture the spatiotemporal correlations between joints. Furthermore, we integrate the temporal context provided by joint features with the spatial context provided by visual features to construct a robust low-resolution representation with 3D spatial awareness. Specifically, we introduce a Triplane feature representation to address the misalignment in semantics and representation forms between joint and visual features, preserving 3D structural information during fusion. Quantitative experiments on multiple datasets have shown that our method achieves comparable performance to high-resolution SOTA methods at much lower resolutions. We also demonstrated the contribution of each design to the scheme performance through ablation study.

**Limitation and Future Work** Our work did not explicitly model long-term hand motion cues across video sequences, which may result in a lack of constraints on motion consistency. As a result, the model may struggle with maintaining consistent hand shapes over extended sequences. Future work should explore utilizing motion cues from long sequences to better constrain the reconstructed hand shape and improve temporal consistency.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China under Grants (62406039, 62101064, 62171057, 62201072, U23B2001, 62001054, 62071067), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center, BUPT Excellent Ph.D. Students Foundation (CX20241014), in part by the Project funded by China Postdoctoral Science Foundation (2023TQ0039, 2024M750257), the Postdoctoral Fellowship Program of CPSF (GZC20230320).

## References

- [1] 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. 2
- [2] Yujun Cai, Lihao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 666–682, 2018. 2
- [3] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. 6
- [4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 2, 4
- [5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2, 6
- [6] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12929–12938, 2021. 1
- [7] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022. 1, 6
- [8] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeon Lee, Elkhani Ismayilzada, and Seungryul Baek. Transformer-based unified recognition of two hands manipulating objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4769–4778, 2023. 6
- [9] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1964–1973, 2021. 3, 6
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3
- [11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 3
- [12] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23600–23611, 2023. 3, 6
- [13] Lihao 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. 2
- [14] Peter Hardy, Srinandan Dasmahapatra, and Hansung Kim. Can super resolution be used to improve human pose estimation in low resolution scenarios? *arXiv preprint arXiv:2107.02108*, 2021. 2, 3
- [15] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 3
- [16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3
- [17] Weiting Huang, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. Awr: Adaptive weighting regression for 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11061–11068, 2020. 2
- [18] 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. 2
- [19] 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. 6
- [20] Leyla Khaleghi, Alireza Sepas-Moghaddam, Josh Marshall, and Ali Etemad. Multiview video-based 3-d hand pose estimation. *IEEE Transactions on Artificial Intelligence*, 4:896–909, 2021. 3
- [21] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF conference on*

- computer vision and pattern recognition*, pages 1077–1086, 2019. 6
- [22] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3, 6
  - [23] Deying Kong, Linguang Zhang, Liangjian Chen, Haoyu Ma, Xiangyi Yan, Shanlin Sun, Xingwei Liu, Kun Han, and Xiaohui Xie. Identity-aware hand mesh estimation and personalization from rgb images. In *European Conference on Computer Vision*, pages 536–553. Springer, 2022. 2
  - [24] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 2
  - [25] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 2, 6
  - [26] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 2, 6
  - [27] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 2, 6
  - [28] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 6
  - [29] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 3
  - [30] Jingyu Liu, Pengfei Ren, Yuanyuan Gao, Jingyu Wang, Haifeng Sun, Qi Qi, Zirui Zhuang, and Jianxin Liao. Key-point fusion for rgb-d based 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3756–3764, 2024. 1
  - [31] Gyeongseok 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. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 1
  - [32] Gyeongseok Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2308–2317, 2022. 2
  - [33] Chihiro Nakatsuka and Satoshi Komorita. Stable 3d human pose estimation in low-resolution videos with a few views. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 427–433. IEEE, 2022. 3
  - [34] Joonkyu Park, Yeonguk Oh, Gyeongseok Moon, Hongsuk Choi, and Kyoung Mu Lee. Handocnet: Occlusion-robust 3d hand mesh estimation network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 6
  - [35] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 6
  - [36] Pengfei Ren, Haifeng Sun, Jiachang Hao, Qi Qi, Jingyu Wang, and Jianxin Liao. Pose-guided hierarchical graph reasoning for 3-d hand pose estimation from a single depth image. *IEEE Transactions on Cybernetics*, 53(1):315–328, 2021. 2
  - [37] Pengfei Ren, Haifeng Sun, Jiachang Hao, Qi Qi, Jingyu Wang, and Jianxin Liao. A dual-branch self-boosting framework for self-supervised 3d hand pose estimation. *IEEE Transactions on Image Processing*, 31:5052–5066, 2022. 2
  - [38] Pengfei Ren, Haifeng Sun, Jiachang Hao, Jingyu Wang, Qi Qi, and Jianxin Liao. Mining multi-view information: a strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20555–20565, 2022.
  - [39] Pengfei Ren, Yuchen Chen, Jiachang Hao, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Two heads are better than one: Image-point cloud network for depth-based 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2163–2171, 2023. 2
  - [40] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8014–8025, 2023. 1, 5
  - [41] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2
  - [42] J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20875–20886, 2023. 4
  - [43] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–98, 2018. 2
  - [44] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11698–11707, 2021. 2
  - [45] Chen Wang, Feng Zhang, Xiatian Zhu, and Shuzhi Sam Ge.

- Low-resolution human pose estimation. *Pattern Recognition*, 126:108579, 2022. [2](#), [3](#)
- [46] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [4](#)
- [47] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17048–17058, 2023. [6](#)
- [48] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre. 3d human pose, shape and texture from low-resolution images and videos. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 4490–4504, 2021. [2](#), [3](#)
- [49] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. [6](#)
- [50] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: Rgb-sequence-based 3d hand pose and shape estimation. In *European Conference on Computer Vision*, pages 122–139. Springer, 2020. [3](#)
- [51] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Jun-song Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13232–13242, 2022. [2](#)
- [52] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. [2](#)
- [53] Xiaozheng Zheng, Pengfei Ren, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Sar: Spatial-aware regression for 3d hand pose and mesh reconstruction from a monocular rgb image. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 99–108. IEEE, 2021. [2](#), [3](#), [5](#), [6](#)
- [54] Xiaozheng Zheng, Chao Wen, Zhou Xue, Pengfei Ren, and Jingyu Wang. Hamuco: Hand pose estimation via multi-view collaborative self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20763–20773, 2023. [6](#)
- [55] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. [3](#)
- [56] 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. [2](#)
- [57] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2024. [6](#)
- [58] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. [2](#)
- [59] 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. [2](#)
- [60] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *Pattern Recognition*, pages 250–264, Cham, 2021. Springer International Publishing. [2](#)