# A$^3$-Net: Calibration-Free Multi-View 3D Hand Reconstruction for Enhanced Musical Instrument Learning

**Geng Chen**, **Xufeng Jian**, **Yuchen Chen**, **Pengfei Ren***, **Jingyu Wang***, **Haifeng Sun**, **Qi Qi**, **Jing Wang** and **Jianxin Liao**
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{chengeng, jianxf, cyc99, rpf, wangjingyu, hfsun, qiqi8266, wangjing, liaojx}@bupt.edu.cn

## Abstract

Precise 3D hand posture is essential for learning musical instruments. Reconstructing highly precise 3D hand gestures enables learners to correct and master proper techniques through 3D simulation and Extended Reality. However, exsiting methods typically rely on precisely calibrated multi-camera systems, which are not easily deployable in everyday environments. In this paper, we focus on calibration-free multi-view 3D hand reconstruction in unconstrained scenarios. Establishing correspondences between multi-view images is particularly challenging without camera extrinsics. To address this, we propose **A**$^3$-Net, a multi-level alignment framework that utilizes 3D structural representations with hierarchical geometric and explicit semantic information as alignment proxies, facilitating multi-view feature interaction in both 3D geometric space and 2D visual space. Specifically, we first perfrom global geometric alignment to map multi-view features into a canonical space. Subsequently, we aggregate information into predefined sparse and dense proxies to further integrate cross-view semantics through mutual interaction. Finnaly, we perfrom 2D alignment to align projected 2D visual features with 2D observations. Our method achieves state-of-the-art results in task of multi-view 3D hand reconstruction, demonstrating the effectiveness of the proposed framework.

## 1 Introduction

While accurate note-playing forms the fundamental basis of musical instrument performance, the mastery of proper hand techniques—particularly seamless fingering transitions and refined hand control—ultimately plays a pivotal role in transforming musical expression into fluid artistry and advancing toward professional level of sophistication. Online videos are popular for learning these hand techniques, but is limited by fixed view points. Exsiting Mixed Reality (MR) applications enable immersive instrumental learning in 3D virtual scenes

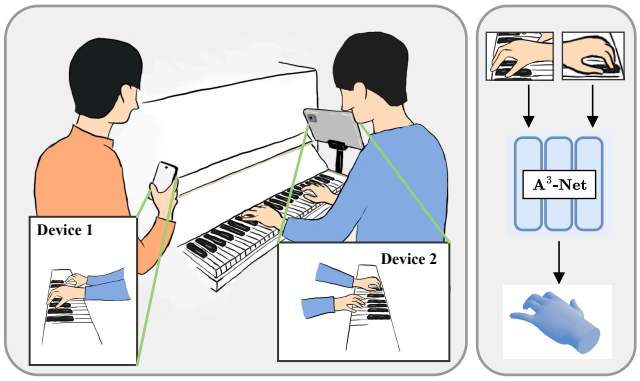


Figure 1: Extrinsic-calibration-free multi-view 3D hand reconstruction for musical instrument playing in everyday scenario. Players can capture multi-view data using their daily devices, which can be either fixed or handheld.

(such as *PianoVision*[1]), but only provide highlighted indicators on the keyboards. Recent studies [Labrou *et al.*, 2023; Liu *et al.*, 2023] found it effective to learn and correct hand movements under MR environment through following the 3D hand postures reconstructed from the teacher. Consequently, reconstructing presice 3D hand postures during instrument performance is significant for enhancing learning outcomes.

Single-view 3D hand reconstruction [Zimmermann and Brox, 2017; Park *et al.*, 2022; Boukhayma *et al.*, 2019; Chen *et al.*, 2021; Ge *et al.*, 2019] is convenient to use, but provides a limited accuracy due to depth and scale ambiguities as well as prevalent occlusions. Multi-view hand reconstruction methods [Iskakov *et al.*, 2019; Ma *et al.*, 2021; Qiu *et al.*, 2019; Tu *et al.*, 2020; He *et al.*, 2020] are more typically used in such scenarios, facilitating more robust and accurate results with explicit disambiguation clues provided by multi-view images. However, these methods are significantly influenced by the precision of extrinsic calibration, necessitating the complex calibration of multiple cameras, which are not easily deployable in everday environments. Furthermore, these methods are not applicable in uncontrolled mobile camera environments where extrinsics are unavailable.

Previous methods for multi-view pose estimation can be

*Corresponding Author.

[1]https://www.pianovision.com/

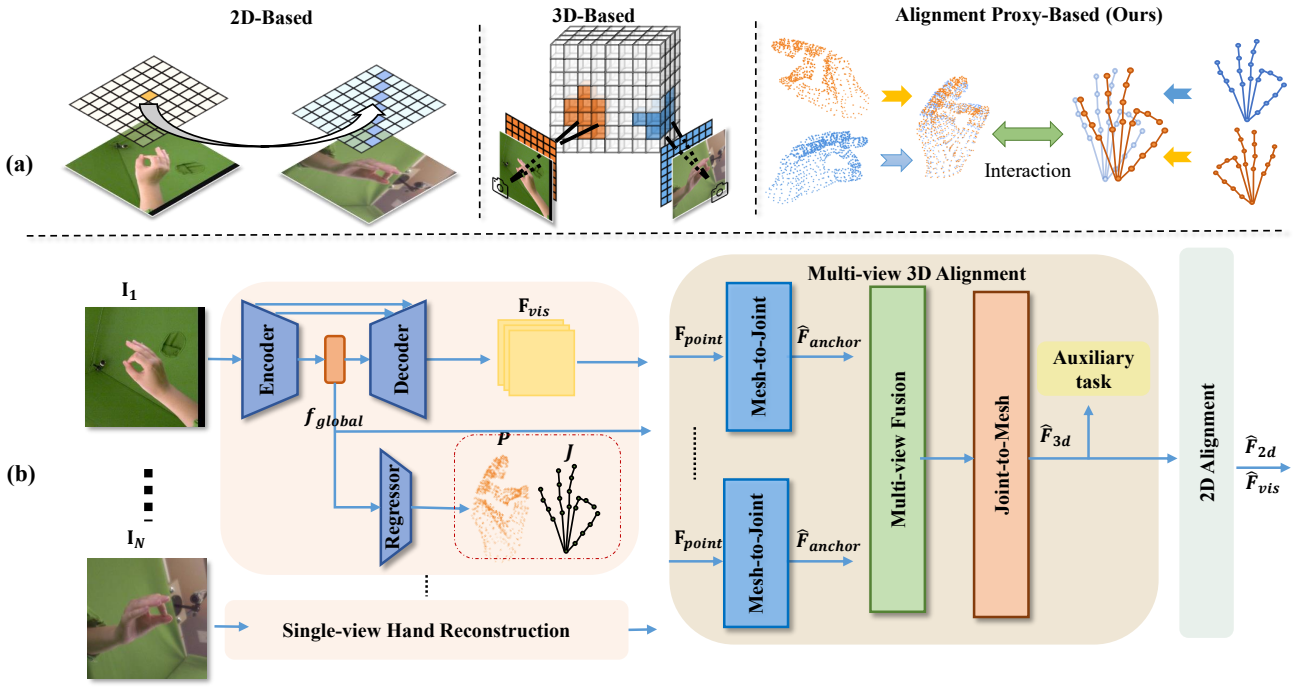


Figure 2: Comparison with other methods (a) and overview of the proposed alignment-proxy based $\mathbf{A}^3$-Net (b), which leverages predefined alignment proxies (meshes and joints) to build multi-view feature correspondence semantically and geometrically. It first leverages an encoder-decoder model for initial hand pose estimation and 2D visual feature extraction within each view. Geometric alignment are applied on meshes and joints to project them into the canonical space. Finally, it iteratively adopts multi-view feature alignment in both 3D point cloud space and 2D image space. For simplicity, only one iteration is shown.

broadly categorized into two categories: 2D-based methods and 3D-based methods. 2D-based methods perform cross view fusion by epipolar geometry [He *et al.*, 2020; Zhang *et al.*, 2021c; Ma *et al.*, 2021] or learned fusion weight [Qiu *et al.*, 2019]. Although epipolar geometry reduces the solution space for multi-view feature matching, this method does not fully utilize 3D spatial structure information. 3D-based methods involve the process of reprojecting features into a 3D voxel space and extracting features through a 3D convolution [Iskakov *et al.*, 2019; Tu *et al.*, 2020] or 3D pictorial structure models [Burenius *et al.*, 2013; Qiu *et al.*, 2019]. This approach enables the explicit interaction of features from different views in 3D space, which can effectively extract 3D geometric structure information. However, converting the feature into voxel representation inevitably introduces quantization artifacts and requires a large amount memory, which hinders the extraction of high-resolution features. Both 2D-based and 3D-based methods rely heavily on manually calibrated camera extrinsics, which can be problematic in real-world scenarios due to the complexity of camera calibration and the unsatisfactory performance with unreliable extrinsics. This reliance on accurate calibration leads to difficulties in establishing multi-view correspondences and results in degraded performance when extrinsics are unreliable or missing.

Consequently, calibration-free multi-view hand pose estimation has emerged as a significant trend, as it avoids errors associated with unreliable or missing extrinsics, resulting in better generalization in real-world applications. However, it is hard to build multi-view feature correspondences without extrinsic calibration. To solve this issue, we propose using 3D structural representations with hierarchical geometric and explicit semantic information as alignment proxies for robust multi-view correspondences. As shown in Figure 2(a), unlike existing 2D-based and 3D-based methods that first extract features and then establish multi-view correspondences dynamically, our approach utilizes predefined alignment proxies and infuses them with the extracted features, which avoids the computational burden of dynamically matching features across different views and simplifies the optimization process by focusing on refining fixed proxy correspondences rather than exhaustively searching for feature alignments. Moreover, this approach does not require maintaining the entire voxel space, thereby reducing overall complexity. Joints and meshes not only carry strong semantic significance, but also exhibit robust geometric correlations that enable efficient interaction between high-level semantic structures and detailed geometric information. Using joints and meshes as alignment proxies offers greater cross-view consistency and, by incorporating these semantically defined proxies as priors, significantly reduces the model's optimization space, reducing the risk of overfitting and improving robustness.

Effectively utilizing alignment proxies to aggregate multi-view information and addressing challenges such as self-occlusion remains complex, due to the need to encoding complementary information across different views at varying lev-

els (such as 2D and 3D) within a framework. To tackle these challenges, as illustrated in Figure 2(b), we introduce $\mathbf{A}^3$-Net, which employs a multi-level optimization strategy and contains three key modules: geometric **A**lignment at the global level, 3D feature **A**lignment in point cloud space, and 2D feature **A**lignment in image space. This layered design allows for a more hierarchical learning process by addressing alignment at different levels of abstraction, thus reducing overall complexity and improving accuracy. Specifically, we first predict the 3D hand mesh from each view and construct the 3D point cloud. To reduce the difficulty of finding correspondences between different views and facilitate multi-view feature interaction, we begin with the **geometric alignment** stage, where the 3D point cloud of each view is transformed into a canonical space. Next, in the **3D alignment** stage, we achieve finer-grained alignment by leveraging the spatial relationships between joints and meshes. This involves interactions from "mesh-to-joint" and "joint-to-mesh". For multi-view fusion, we connect features based on their anatomical semantics and fuse them using a Graph Convolutional Network (GCN). Finally, in the **2D alignment** stage, to mitigate image-mesh misalignment, we project the spatially-aware multi-view features onto the 2D image space, aligning them with the 2D visual features and performing local feature refinement. As shown in Figure 1, our method facilitates convenient uncalibrated multi-view hand reconstruction using multiple everyday portable devices, thereby offering detailed hand information to enhance music instruction.

In summary, our main contributions are threefold:

- We propose a new uncalibrated multi-view hand reconstruction framework that eliminates the dependency on camera extrinsics by aggregating information on predefined alignment proxies such as joints and meshes.
- We propose a multi-level alignment approach that incorporates geometric alignment of hand meshes, 3D alignment with predefined proxies, and 2D alignment for local feature refinement, leveraging both geometric and semantic information enhance overall performance.
- Our method achieves state-of-the-art (SOTA) results on two challenging multi-view hand-object interaction datasets, DexYCB [Chao *et al.*, 2021] and HanCo [Zimmermann *et al.*, 2022].

## 2 Related Work

### 2.1 Multi-View 3D Pose Estimation With Camera Extrinsics

Methods with known camera extrinsics can be further divided into 3D-based and 2D-based methods. 3D-based methods use camera parameters to re-project features of different view into 3D voxel representation [Iskakov *et al.*, 2019], and use 3D CNN [Tu *et al.*, 2020] or Pictorial Structure Model (PSM) [Qiu *et al.*, 2019] to fuse the multi-view features. 2D-based methods [He *et al.*, 2020; Ma *et al.*, 2021] typically fuse the multi-view features in 2D space according to epipolar geometry. Additionally, POEM [Yang *et al.*, 2023] directly operates on 3D points embedded in multi-view stereo for hand mesh reconstruction, effectively leveraging the 3D geometrical information. Despite the effectiveness of these methods, they rely heavily on the accuracy of the camera extrinsics, resulting in poor performance in real-world scenarios where camera extrinsics are unavailable or unreliable.

### 2.2 Multi-View 3D Pose Estimation Without Camera Extrinsics

Camera-extrinsics-free methods typically use body semantic prior to aggregate features from different views. For dense alignment proxy, these methods generally employ mesh or pixel-level alignment. Although they can capture fine-grained details, they often suffer from high computational costs and complexity, which limits their scalability. PaFF [Jia *et al.*, 2023] fuses pixel-aligned features on mesh vertices, allowing the regressor to iteratively align the body mesh with each input view. In [Yu *et al.*, 2022], they also use vertices of the human model as a semantic template for multi-view alignment, mapping visual features to the model and employing self-attention for pose estimation, which captures detailed information but increases computational load.

For sparse alignment proxy, these methods typically use joints as the alignment proxy. They are more computationally efficient but may lack the granularity of dense alignment methods. FLEX [Gordon *et al.*, 2022] utilizes the view-invariance characteristic of bone lengths and rotation angles between skeletal parts to reconstruct human pose. MTF-Transformer [Shuai *et al.*, 2022] uses Transformer to model the relative positions between multiple views based on joints. FusionFormer [Cai *et al.*, 2024] uses joints as the alignment proxy and models both spatial and temporal relationships, even outperforming most methods that require camera parameters. However, for tasks requiring full mesh reconstruction, these methods cannot directly use meshes as the alignment proxy due to high computational demands, thereby limiting their applications. In contrast to these methods, we use both joints and meshes as alignment proxies. To reduce computational load, we enable interaction between these two alignment proxies in 3D space, allowing for efficient exchange and aggregation of information, balancing the granularity and computational efficiency.

## 3 Method

In this paper, we propose an alignment proxies-based (*i.e.*, joints and meshes) framework to fuse multi-view features geometrically and semantically without requiring camera extrinsic calibration. As shown in Figure 2(b), to better interact and aggregate information into the proxies, we divide the entire framework into three stages to reduce initial spatial discrepancies, facilitate information flow between different proxies, and mitigate the image-mesh gap in geometric alignment stage, 3D alignment stage and 2D alignment stage, respectively. Given RGB images as input, $\mathbf{A}^3$-Net first predicts hand mesh of each view and construct the point cloud base on the surface of hand mesh, incorporating hand priors from the MANO [Romero *et al.*, 2017] model. After constructing geometry-aware initial 3D point feature, we geometrically align the point cloud from different views to the canonical space by predicting transformation matrices. In 3D

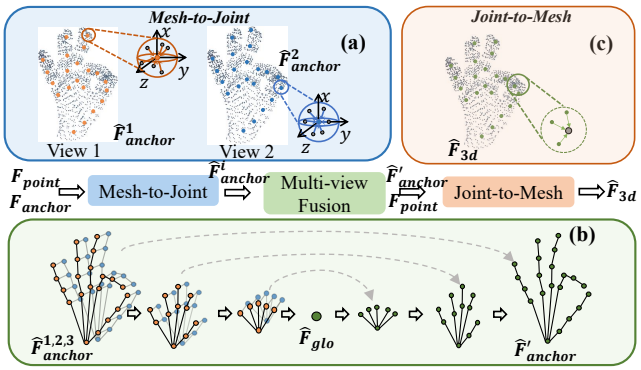


Figure 3: The process of feature alignment in 3D point cloud space, including (a) mesh-to-joint interaction, (b) multi-view fusion and (c) joint-to-mesh interaction.

point cloud space, we use an efficient local feature extraction module to extract spatially-aware feature and fuse multi-view feature of the key regions, achieving feature alignment in 3D point cloud space. Subsequently, the multi-view features are projected to each view to promote the refinement of local features in 2D image space.

### 3.1 Single-View Hand Reconstruction

To obtain fine-grained visual features, we use an encoder-decoder network [Park *et al.*, 2022; Ren *et al.*, 2022] as our backbone. The encoder extracts a global feature $\mathbf{f}_{global}$, which is used to predict the pose parameters $\theta$ and shape parameters $\beta$ of the MANO model. For $N$ views, we regress the initial mesh $\mathbf{V}$ independently. Following [Ren *et al.*, 2022], we predict pixel-wise representations $\mathbf{O}$ to obtain hand joints $\mathbf{J}$ and visual features $\mathbf{F}_{vis}$ that are strongly correlated with hand regions. The 3D point cloud $\mathbf{P}$ is constructed from the surface vertices of the hand mesh $\mathbf{V}$.

### 3.2 Geometric Alignment

To simplify interaction and facilitate multi-view feature fusion, we propose to geometrically align proxies from different views. However, this process is not straightforward due to the lack of camera extrinsics. To overcome this challenge, we designate the space of the first view as the "canonical space". Proxies of the other views are geometrically aligned to this space using predicted transformation matrices $\mathbf{T}_{i,j}$ between view $i$ and view $j$ by Multi-Layer Perceptron(MLP) according to the difference of features $\mathbf{f}_{global}$.

$$\mathbf{T}_{i,j} = \text{MLP}(\mathbf{f}^i_{global} - \mathbf{f}^j_{global}) \tag{1}$$

Finally, we apply the predicted $\mathbf{T}_{i,j}$ to project the joints $J$ and mesh point clouds $\mathbf{P}$ into the canonical space.

### 3.3 Feature Alignment in 3D Point Cloud Space

To achieve spatially and semantically-aware feature alignment in 3D point cloud space, it is crucial to model 3D spatial structure. However, previous methods either interact with multi-view information in 2D image space or in large voxel-based 3D space, which suffered from quantization errors and expensive computational costs. Furthermore, hands are suffered from complex articulated structures, hand-object occlusions and self-occlusions, making direct one-step 3D alignment challenging. Joints provide sparse, high-level structural information, while meshes offer dense, detailed geometric features. To efficiently harness these complementary strengths, we can only keep lightweight 3D features and implement dual-proxy interaction, which breaks down the 3D alignment process into several steps. This approach reduces the complexity of learning 3D spatially-aware features for key regions, optimizes memory usage, and preserves crucial spatial and semantic details for robust hand pose estimation.

**Single-View Geometry-Aware Initial Feature Construction.** Constructing effective initial point feature $\mathbf{F}_{point}$ is essential. Previous methods [Ma *et al.*, 2021; He *et al.*, 2020; Yu *et al.*, 2022] focused solely on visual information, neglecting 3D geometric details. To gather 3D spatially-aware 2D visual features $\mathbf{F}_{2d}$, we project the 3D point cloud $\mathbf{P}$ onto the 2D image plane and sample the $K_1$ closest elements from $\mathbf{F}_{vis}$ similar to [Ren *et al.*, 2023]. To incorporate 3D geometric information, we use linear transformation with batch normalization to encode the coordinates of points $\mathbf{P}$ as position information $\mathbf{F}_{3d}$ and embed $\mathbf{f}_{global}$ into global features $\mathbf{F}_{glo}$. The final point feature is then computed as $\mathbf{F}_{point} = \text{ReLU}(\mathbf{F}_{2d} + \mathbf{F}_{3d} + \mathbf{F}_{glo})$. Similarly, by replacing $\mathbf{P}$ with $\mathbf{J}$, we obtain anchor features $\mathbf{F}_{anchor}$. For geometric alignment, as $\mathbf{f}_{global}$ encodes MANO parameters, including root wrist positions and axis angles which offers crucial information for predicting inter-view rotations, we leverage it to estimate transformation matrices in Eq. 1.

**Mesh-to-Joint Interaction.** As illustrated in Figure 3(a), we first aggregate the local spatial information of neighboring points into joints of each view geometrically and semantically using ball query [Qi *et al.*, 2017]. Subsequently, we obtain the regional spatially-aware features through traditional MLP and max pooling:

$$\mathbf{F}^i_{region} = \text{Pool}(\phi_1(\text{Concat}(d_{i,j}, \mathbf{F}^{i,j}_{diff})), |j = 1, ..., K_2) \tag{2}$$

where $\mathbf{F}^i_{region}$ represents the features of region $i$, with each region being a collection of points centered around anchor $i$. Here, $\phi_1$ and $\phi_2$ are Fully-Connected (FC) layers followed by batch normalization and activation. $d_{i,j}$ is the relative coordinate between anchor $i$ and point $j$, and $\mathbf{F}^{i,j}_{diff}$ is the feature difference between them. Finally, the anchor point features $\mathbf{F}^i_{anchor}$ are updated using the regional features:

$$\hat{\mathbf{F}}^i_{anchor} = \phi_2(\text{Concat}(\mathbf{F}^i_{anchor}, \mathbf{F}^i_{region})) \tag{3}$$

**Multi-View Fusion.** As illustrated in Figure 3(b), we achieve 3D feature alignment through multi-view feature interaction and fusion based on anchors(joints). To leverage the hand prior for enhancing spatial structure information, we connect anchor points across views according to the anatomical structure of the hands. These multi-view joint features are then fused along both the joint and view dimensions using a hierarchical GCN with pooling layers. Similar to [Cai *et al.*, 2019], a U-shaped GCN is employed to construct a multi-view graph based on the skeletal and semantic correspondences among hand nodes.

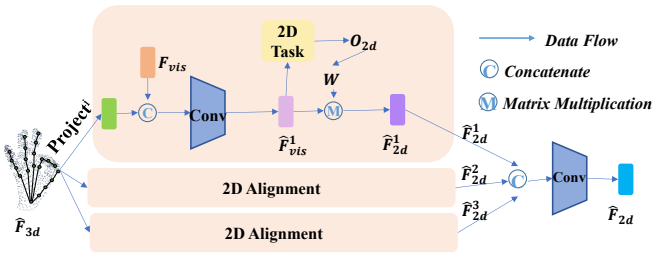


Figure 4: Local feature alignment in 2D image space.

**Joint-to-Mesh Interaction.** As illustrated in Figure 3(c), the updated anchor features are propagated to the original point cloud similar to [Qi *et al.*, 2017]. Specifically, we first select $K_3$ anchor points closest to each point and carry out feature interpolation according to the distance. The interpolated features are then concatenated with the original point features $\mathbf{F}_{point}$ and pass through point-wise MLP [Qi *et al.*, 2017] to obtain 3D spatially-aware point cloud features $\hat{\mathbf{F}}_{3d}$.

Finally, we regress point-wise representation $\mathbf{P}^p$ from the updated point cloud features $\hat{\mathbf{F}}_{3d}$. We obtain the 3D hand pose from the point-wise representation using a weighted average algorithm, and regress the MANO parameters by passing the global features $\hat{\mathbf{F}}_{glo}$ through the MLP.

### 3.4 Feature Alignment in 2D Image Space

For multi-view hand reconstruction, aligning the reprojections of hand mesh with observations from each view is crucial, as it provides strong regularization for accurate 3D hand reconstruction. Although initial point cloud features are derived from 2D visual feature maps, relying solely on 3D alignment is insufficient. This is because interacting and refining features in 3D space alone can lead to misalignment between the predicted hand meshes and the 2D images due to the lack of explicit constraints. To address this, we propose incorporating 2D image space alignment to ensure consistent feature alignment across both 3D and 2D domains.

Specifically, as illustrated in Figure 4, we first project the refined 3D features $\hat{\mathbf{F}}_{3d}$ onto the 2D image plane for each view and concatenate them with $\mathbf{F}_{2d}$ to obtain updated visual features $\hat{\mathbf{F}}^i_{vis}$ for view $i$. These updated features are then refined locally using 2D CNNs:

$$\hat{\mathbf{F}}^i_{vis} = \phi^i_3(\text{Concat}(\text{Project}^i(\hat{\mathbf{F}}_{3d}), \mathbf{F}_{vis})) \tag{4}$$

To further enhance joint-image alignment, we adopt pixel-wise regression as an auxiliary task which encodes $\hat{\mathbf{F}}^i_{vis}$ into pixel-level representations $\mathbf{O}_{2d}$ and generates 2D and 2.5D heatmaps to introduce hand anatomy priors for better alignment. The aligned 2D image features $\hat{\mathbf{F}}^i_{2d}$ are then obtained by performing a weighted sum on the refined 2D feature map $\hat{\mathbf{F}}^i_{vis}$, using a weight map $\mathbf{W}$ derived from $\mathbf{O}_{2d}$:

$$\hat{\mathbf{F}}^i_{2d} = \sum(\text{Softmax}(\mathbf{W}) \cdot \hat{\mathbf{F}}^i_{vis}) \tag{5}$$

Finally, 2D features $\hat{\mathbf{F}}^i_{2d}$ of each view are concatenated and fused by convolution to get updated 2D features $\hat{\mathbf{F}}_{2d}$.

Unlike PyMAF [Zhang *et al.*, 2021a], which continuously samples features from the static feature map, $\mathbf{A}^3$-Net project the updated 3D multi-view features $\hat{\mathbf{F}}_{3d}$ to each view and refine the 2D features using 2D CNN. By incorporating 3D spatially-aware multi-view features, $\mathbf{A}^3$-Net mitigates visual feature degradation caused by occlusion, leading to more robust local visual features.

### 3.5 Iterative Feature Alignment and Loss

To fully capture both 3D geometric structures and 2D visual information for more accurate 3D pose and hand reconstruction results, we propose to repeat the above alignment stages several times. After aligning the features in both 3D point cloud space and 2D image space, we iteratively update the point cloud features by $\hat{\mathbf{F}}_{point} = \text{ReLU}(\text{BN}(W_3\hat{\mathbf{F}}_{3d}) + \text{BN}(W_4\hat{\mathbf{F}}_{2d}) + \text{BN}(W_5\hat{\mathbf{F}}_{glo}))$, where $W_3$, $W_4$ and $W_5$ are the learnable parameters matrices. The updated point features are then passed to next 2D and 3D alignment stage.

## 4 Experiments

### 4.1 Datasets and Experimental Settings

**DexYCB** [Chao *et al.*, 2021] is a large-scale dataset that records the pose of the hand grasping on objects, which contains 582K images with 20 objects selected from the YCB-Video dataset. It provides 4 ways to divide the dataset, and we evaluat our method using default "S0" split. To construct a multi-view scene with a total of $n = 3$ viewpoints, we selected one primary viewpoint and randomly chose $n - 1$ auxiliary viewpoints from the remaining viewpoints. The random selection of $n$ cameras is intended to simulate a dynamic camera scenario, where it is impossible to determine the extrinsic parameters of each camera. However, the intrinsic parameters of the cameras are easily obtained.

**HanCo** [Zimmermann *et al.*, 2022] consists of 1517 videos with multiple views and camera calibration. As it does not provide official partition of the training and testing sets, we divide it in an 8:2 ratio. According to recent methods [Zheng *et al.*, 2023], the first 1200 sequences are used for training, while the remaining 317 sequences are used for testing. Viewpoints are selected similar to DexYCB.

### 4.2 Training Details

We implement our method by PyTorch framework. All experiments are conducted on an NVIDIA RTX 4090 GPU. We train our network using AdamW optimizer with the initial learning rate of 1e-4 and divided by 10 every 10 epochs. The whole training process takes 30 epochs with batch size of 32. We adopt random translation, random rotation, and random scaling for training augmentation. We crop the hands from the input images by 2D keypoints and resize them to resolution $256 \times 256$ during training and testing. During the training phase, each monocular image is treated as a primary viewpoint, and n-1 auxiliary viewpoints are randomly selected from other viewpoints to ensure uniform dataset sampling.

### 4.3 Comparisons with State-of-the-Arts

To the best of our knowledge, there are few previous works for multi-view 3D hand reconstruction. Following the multi-

| Dataset | DexYCB | | | | HanCo | | | |
|---|---|---|---|---|---|---|---|---|
| Method | MPJPE↓ | P-MPJPE↓ | MPVPE↓ | P-MPVPE↓ | MPJPE↓ | P-MPJPE↓ | MPVPE↓ | P-MPVPE↓ |
| PPT [Ma *et al.*, 2022] | 9.22 | 4.97 | / | / | 6.61 | 6.10 | / | / |
| Voxel [Iskakov *et al.*, 2019] | 7.81 | 4.34 | / | / | 4.92 | 3.63 | / | / |
| MMI [Ren *et al.*, 2022] | 7.93 | 4.71 | 8.32 | 5.15 | 5.54 | 3.90 | 6.24 | 4.78 |
| MVP [Zhang *et al.*, 2021b] | **6.23** | 4.26 | 9.77 | 8.14 | / | / | / | / |
| Ours * | 6.62 | **3.93** | **6.81** | **4.12** | **4.02** | **3.01** | **4.35** | **3.24** |

Table 1: Comparison with SOTA methods of MPJPE(mm), P-MPJPE(mm), MPVPE(mm) and P-MPVPE(mm) on the DexYCB & HanCo datasets. '*' denotes calibration-free method.

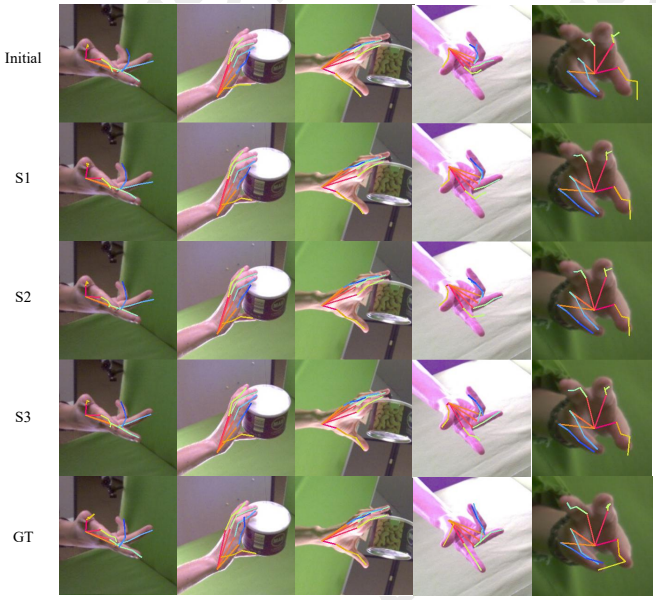


Figure 5: Qualitative visualization of the iterative alignment. We present results from the initial single view estimation, alignment stage(S1-S3) and the ground truth.

| Method | MPJPE↓ | MPVPE↓ | P-MPJPE↓ | P-MPVPE↓ |
|---|---|---|---|---|
| Visual | 7.95 | 4.63 | 8.02 | 4.94 |
| +Position | 7.02 | 4.23 | 7.41 | 4.57 |
| +Global | **6.62** | **3.93** | **6.81** | **4.12** |

Table 2: Comparison of the effects of different initial features on the DexYCB dataset.

view 3D human pose estimation method, we construct several baselines for fair comparison, including PPT [Ma *et al.*, 2022], MVP [Zhang *et al.*, 2021b], Volumetric Triangulation [Iskakov *et al.*, 2019] and MMI [Ren *et al.*, 2022]. We conducted experiments on DexYCB and HanCo datasets respectively, as shown in Table 1. For each methods, we randomly select 3 views to simulate in the wild scenario. Even without camera extrinsics, our method can achieve comparable results with MVP [Zhang *et al.*, 2021b] and outperform the other methods.

## 4.4 Ablation Study

**Effectiveness of the Initial Features Construction**

In order to illustrate the importance of spatial information, we use different combinations of visual, positional, and global features to construct initial features and carry out multi-view feature interaction. As shown in Table 2, adding position information significantly enhances the accuracy of 3D pose estimation by almost 1mm MPJPE. In addition, incorporating global features can further improve the performance.

| | ID | Method | MPJPE | P-MPJPE | MPVPE | P-MPVPE |
|---|---|---|---|---|---|---|
| Geo. | 1 | Procrustes | 7.00 | 4.20 | 7.32 | 4.50 |
| | 2 | Learning | 6.62 | 3.93 | 6.81 | 4.12 |
| | 3 | Extrinsics | **6.31** | **3.83** | **6.52** | **4.01** |
| 3D | 4 | MLP | 7.11 | 4.22 | 7.33 | 4.55 |
| | 5 | Transformer | 6.80 | 4.08 | 7.02 | 4.31 |
| | 6 | Pool-GCN | **6.62** | **3.93** | **6.81** | **4.12** |
| 2D | 7 | W/o Align | 7.52 | 4.53 | 7.95 | 4.82 |
| | 8 | W/o Sup. | 7.01 | 4.22 | 7.34 | 4.38 |
| | 9 | Seg. | 6.83 | 4.08 | 7.12 | 4.10 |
| | 10 | 2.5D Hmp. | 7.22 | 4.31 | 7.53 | 4.42 |
| | 11 | 2D Hmp. | **6.62** | **3.93** | **6.81** | **4.12** |

Table 3: Quantitative comparison of the different alignment methods on the DexYCB dataset. 'Sup' is short for supervision. 'Seg' and 'Hmp' represent supervision by segmentation and heatmap, respectively.

| Iteration | MPJPE↓ | MPVPE↓ | Params(M) | FLOPs(G) |
|---|---|---|---|---|
| 1 | 6.93 | 7.08 | 25.18 | 8.93 |
| 2 | 6.71 | 6.90 | 27.88 | 11.03 |
| 3 | **6.62** | **6.81** | 30.58 | 13.11 |
| 4 | 6.70 | 6.83 | 33.28 | 15.21 |

Table 4: Comparison of different iterations of alignments on the DexYCB dataset.

**Effectiveness of the Geometric Alignment**

We aims to reduce spatial differences between views in the geometric alignment stage. In $\mathbf{A}^3$-Net, we use an MLP to predict transformation matrices between views. Additionally, Procrustes alignment or camera extrinsics can also be used for alignment. We analyze the impact of different geometric alignment strategies on 3D hand reconstruction, as shown in Table 3 from ID1 to ID3. Among them, the extrinsic-based alignment (ID3) achieves the highest accuracy. However, our proposed learning-based method (ID2), even without camera extrinsics, performs comparably to ID3 and significantly outperforms the Procrustes-based method (ID1).

**Effectiveness of Fusion Methods in 3D Feature Alignment**

We analyze different multi-view anchor (joint) feature fusion methods, as shown in Table 3, from ID4 to ID6. Semantically-aware and geometrically-aware 3D feature alignment rely significantly on effective multi-view feature

fusion. Therefore, we propose to use pool-GCN (ID6) for multi-view feature interaction. This interaction module can also be replaced with commonly used MLP (ID4) or Transformer (ID5). MLP-based fusion (ID4) concatenates features of anchors sharing the same semantic information and directly fuses multi-view features at the channel level. However, this approach lacks the capability to facilitate interactions between distant regions in the point cloud, leading to the worst performance. Transformer-based fusion (ID5) improves upon this by performing self-attention between anchor features, enabling interaction across remote regions. Despite incorporating global information, the absence of prior structural information during feature interaction results in less accurate hand mesh predictions. In contrast, pool-GCN (ID6) not only efficiently fuses multi-view features but also effectively incorporates global context, leading to more accurate hand mesh reconstruction, with MPJPE reduced by 0.5 mm and 0.2 mm compared to ID4 and ID5, respectively. Therefore, pool-GCN is adopted in our proposed method to achieve superior multi-view feature fusion.

**Effectiveness of 2D Feature Alignment**

To demonstrate the effectiveness of our 2D feature alignment, we conduct ablation experiments to investigate the necessity of employing 2D alignment and the selection of auxiliary tasks. as illustrated in Table 3 from ID7 to ID11. First, we observe that simply applying 2D features alignment without supervision (ID8) significantly improves the performance compared to that without aligning 2D features (ID7). More importantly, when 2D feature alignment is carried out, the performance can be further improved if we use auxiliary tasks (ID9-ID11) to guide the process of feature alignment. Among these tasks, using 2D heatmap as supervision yields the best performance. Therefore, $\mathbf{A}^3$-Net adopts 2D heatmap as the supervision of 2D alignment to facilitate the local feature refinement.

**Effectiveness of the Number of Iteration Stages**

To reduce the difficulty of hand pose reconstruction, we apply 3D and 2D alignment multiple times to better capture the 3D structural information and 2D visual cues across multiple views for iterative correction. We explore the impact of different iteration stages on model performance as shown in Table 4, and observe that increasing the number of iterations generally improves performance of the whole network. To balance computational cost and performance, we adopt three iterations in the final model. Figure 5 demonstrates the gradual reduction in error between predicted joints and ground truth as alignment stages progress.

# 5 Discussion

## 5.1 Real-World Pipeline

We build a full pipeline to preliminarily validate the effectiveness of our method under everyday scenario (Figure 6). The teacher records two videos from both sides of the piano, using a smartphone and a tablet without professional extrinsic calibration. We first use a finetuned YOLOv7 [Wang *et al.*, 2023] to track 2D bounding boxes of both hands. Then, we apply $\mathbf{A}^3$-Net for calibration-free multi-view hand reconstruction. Finally, we reproduce the entire gestures in a Unity application and perform virtual projection on a Mixed Reality headset. The learner can observe the 3D virtual hands from any viewpoints and follow the hand movements as instructional guidance for skill acquisition.

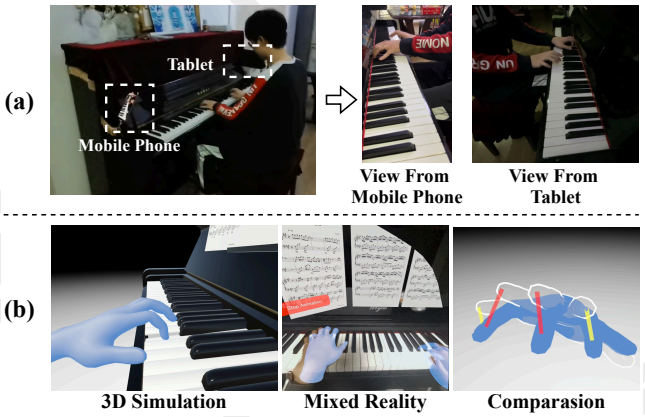


Figure 6: Visualization results for in-the-wild experiment. The player use a mobile phone and a tablet to capture multi-view images (a). The reconstruction results are then utilized for virtual 3D simulation, Mixed Reality follow-along learning and hand posture evaluation (b).

## 5.2 Use Cases

**Shared 3D Instrument Learning.** For instrumental performance, everyone can be a teacher or a learner. Using our method, users can easily model accurate 3D hand postures during preformance with two (or more) of their daily devices without extrinsic calibration. They can share their processed 3D performance data, which others can access and use for 3D immersive follow-along learning.

**Hand Performance Review.** Our method allows performers to comprehensively record and reconstruct their 3D performance process with multiple daily devices, enabling more detailed analysis and review of their hand gestures, fingering, and other hand techniques. They can identify issues such as incorrect key presses or unattractive hand postures.

# 6 Conclusion

Accurate 3D hand reconstruction is significant for musical instrument learning. In this paper, we propose $\mathbf{A}^3$-Net for multi-view 3D hand pose estimation in unconstrained scenarios without camera extrinsics. To address the challenge of searching for the correspondences between different views, we introduce a multi-level alignment strategy that utilizes hand meshes and joints as alignment proxies and performs geometric alignment, 3D alignment and 2D alignment, respectively. Our method can achieve efficient interaction of multi-view features without camera extrinsics. To evaluate the effectiveness of our method, we conduct extensive experiments on challenging hand-object interaction datasets. The state-of-the-art performance demonstrates the superiority of our method.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants (62406039, 62321001, 62471055, U23B2001, 62171057, 62201072, 62071067), the High-Quality Development Project of the MIIT(2440STCZB2584), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), the Fundamental Research Funds for the Central Universities (2024PTB-004), the Project funded by China Postdoctoral Science Foundation (2023TQ0039, 2024M750257, GZC20230320).

## Contribution Statement

This work was a collaborative effort by all contributing authors. Geng Chen and Xufeng Jian contributed equally to this research and are designated as co-first authors. Pengfei Ren and Jingyu Wang, serving as the corresponding authors, are responsible for all communications related to this manuscript.

## References

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

[Burenius *et al.*, 2013] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3618–3625, 2013.

[Cai *et al.*, 2019] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2272–2281, 2019.

[Cai *et al.*, 2024] Yanlu Cai, Weizhong Zhang, Yuan Wu, and Cheng Jin. Fusionformer: A concise unified feature fusion transformer for 3d pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 900–908, 2024.

[Chao *et al.*, 2021] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021.

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

[Ge *et al.*, 2019] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10833–10842, 2019.

[Gordon *et al.*, 2022] Brian Gordon, Sigal Raab, Guy Azov, Raja Giryes, and Daniel Cohen-Or. Flex: Extrinsic parameters-free multi-view 3d human motion reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 176–196. Springer, 2022.

[He *et al.*, 2020] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7779–7788, 2020.

[Iskakov *et al.*, 2019] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7718–7727, 2019.

[Jia *et al.*, 2023] Kai Jia, Hongwen Zhang, Liang An, and Yebin Liu. Delving deep into pixel alignment feature for accurate multi-view human mesh recovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 989–997, 2023.

[Labrou *et al.*, 2023] Katerina Labrou, Cagri Hakan Zaman, Arda Turkyasar, and Randall Davis. Following the master's hands: Capturing piano performances for mixed reality piano learning applications. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[Liu *et al.*, 2023] Ruofan Liu, Erwin Wu, Chen-Chieh Liao, Hayato Nishioka, Shinichi Furuya, and Hideki Koike. Pianohandsync: An alignment-based hand pose discrepancy visualization system for piano learning. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[Ma *et al.*, 2021] Haoyu Ma, Liangjian Chen, Deying Kong, Zhe Wang, Xingwei Liu, Hao Tang, Xiangyi Yan, Yusheng Xie, Shih-Yao Lin, and Xiaohui Xie. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. *arXiv preprint arXiv:2110.09554*, 2021.

[Ma *et al.*, 2022] Haoyu Ma, Zhe Wang, Yifei Chen, Deying Kong, Liangjian Chen, Xingwei Liu, Xiangyi Yan, Hao Tang, and Xiaohui Xie. Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 424–442. Springer, 2022.

[Park *et al.*, 2022] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1496–1505, 2022.

[Qi *et al.*, 2017] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[Qiu *et al.*, 2019] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4342–4351, 2019.

[Ren *et al.*, 2022] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20555–20565, 2022.

[Ren *et al.*, 2023] 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 *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[Romero *et al.*, 2017] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, November 2017.

[Shuai *et al.*, 2022] Hui Shuai, Lele Wu, and Qingshan Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2022.

[Tu *et al.*, 2020] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision (ECCV)*, pages 197–212. Springer, 2020.

[Wang *et al.*, 2023] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, 2023.

[Yang *et al.*, 2023] Lixin Yang, Jian Xu, Licheng Zhong, Xinyu Zhan, Zhicheng Wang, Kejian Wu, and Cewu Lu. Poem: reconstructing hand in a point embedded multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21108–21117, 2023.

[Yu *et al.*, 2022] Zhixuan Yu, Linguang Zhang, Yuanlu Xu, Chengcheng Tang, Luan Tran, Cem Keskin, and Hyun Soo Park. Multiview human body reconstruction from uncalibrated cameras. In *Advances in Neural Information Processing Systems*, 2022.

[Zhang *et al.*, 2021a] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *IEEE International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021.

[Zhang *et al.*, 2021b] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34:13153–13164, 2021.

[Zhang *et al.*, 2021c] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 129:703–718, 2021.

[Zheng *et al.*, 2023] Xiaozheng Zheng, Chao Wen, Zhou Xue, Pengfei Ren, and Jingyu Wang. Hamuco: Hand pose estimation via multiview collaborative self-supervised learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 20763–20773, 2023.

[Zimmermann and Brox, 2017] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4903–4911, 2017.

[Zimmermann *et al.*, 2022] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *Pattern Recognition (PR)*, pages 250–264. Springer, 2022.