Keypoint Fusion for RGB-D Based 3D Hand Pose Estimation

Xingyu Liu*, Pengfei Ren*, Yuanyuan Gao, Jingyu Wang†, Haifeng Sun†, Qi Qi, Zirui Zhuang, Jianxin Liao

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
{liuxingyu, rpf, gaoyuanyuan, wangjingyu, hfsun, qiqi8266, zhuangzirui, liaojx}@bupt.edu.cn

Abstract

Previous 3D hand pose estimation methods primarily rely on a single modality, either RGB or depth, and the comprehensive utilization of the dual modalities has not been extensively explored. RGB and depth data provide complementary information and thus can be fused to enhance the robustness of 3D hand pose estimation. However, there exist two problems for applying existing fusion methods in 3D hand pose estimation: redundancy of dense feature fusion and ambiguity of visual features. First, pixel-wise feature interactions introduce high computational costs and ineffective calculations of invalid pixels. Second, visual features suffer from ambiguity due to color and texture similarities, as well as depth holes and noise caused by frequent hand movements, which interferes with modeling cross-modal correlations. In this paper, we propose Keypoint-Fusion for RGB-D based 3D hand pose estimation, which leverages the unique advantages of dual modalities to mutually eliminate the feature ambiguity, and performs cross-modal feature fusion in a more efficient way. Specifically, we focus cross-modal fusion on sparse yet informative spatial regions (i.e. keypoints). Meanwhile, by explicitly extracting relatively more reliable information as disambiguation evidence, depth modality provides 3D geometric information for RGB feature pixels, and RGB modality complements the precise edge information lost due to the depth noise. Keypoint-Fusion achieves state-of-the-art performance on two challenging hand datasets, significantly decreasing the error compared with previous single-modal methods.

Introduction

3D hand pose estimation is a critical technology for interactive media and human-computer interaction applications, many of which present more challenging scenarios and demand higher accuracy in 3D hand pose estimation. For instance, augmented reality applications involve hand-object interactions with frequent movements and occlusions; and more accurate poses are essential to enhance interactive realism and user experience when manipulating objects. Despite recent advancements in single RGB or depth-based approaches (Kulon et al. 2020; Ge et al. 2019; Park et al. 2022; Chen et al. 2021; Huang et al. 2020b,c), they encounter challenges due to the inherent defects of their respective modalities. RGB-based methods are prone to color similarity and lack local geometric information, while depth-based methods suffer from noise and depth holes at the edges of hands and objects, especially during motion, as shown in Fig. 1.

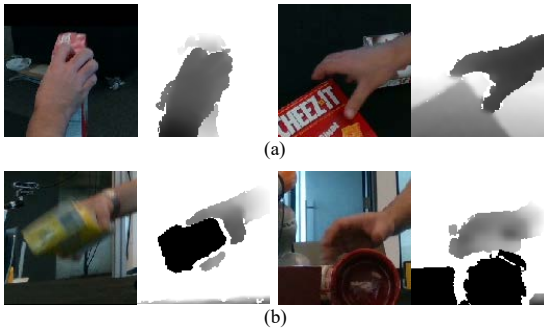


Figure 1: In the 3D hand pose estimation task, depth maps are prone to (a) noise and depth holes at the edges of hands and objects, especially during (b) motion.

Fortunately, there exist many complementary properties between RGB and depth modalities. For example, RGB data can provide rich texture and accurate edge information, while depth data contains detailed geometric structure features. By fusing their complementary advantages, the impact of individual modality defects on the performance of 3D hand pose estimation can be alleviated.

Multi-sensor modality fusion has emerged as a significant trend in visual perception technology. Although not widely explored in 3D hand pose estimation, multimodal fusion has been extensively studied in many other computer vision tasks, such as LiDAR-camera 3D object detection and RGB-D semantic segmentation. Early works (Hu et al. 2019; Chen et al. 2020; Seichter et al. 2021) perform weighted aggregation of cross-modal features through the channel and spatial attention mechanism (Woo et al. 2018). Recently, with the extensive research of vision Transformer (Dosovitskiy et al. 2020; Liu et al. 2021b; Vaswani et al. 2017), several methods utilize the attention mechanism for interactions between cross-modal features. For instance, TransFusion (Bai et al. 2022) uses object queries to obtain initial 3D bounding

*Equal contribution.
†Corresponding author.

boxes from point cloud features, and fuses the image features to boost the quality of the initial prediction. AutoAlign (Chen et al. 2022c) uses the cross-attention mechanism to aggregate the cross-modal feature between the camera-view domain and the voxel domain. CMX (Liu et al. 2022a) connects the soft attention and cross-attention mechanism in serial to calibrate and aggregate the cross-modal feature.

However, there exist two key problems for applying existing fusion methods in 3D hand pose estimation. First, these methods suffer from high computational costs introduced by the dense interactions between multimodal feature elements, which contain redundant interactions of many invalid pixels, such as empty voxels and background pixels. To reduce the computational complexity and memory usage, several works (Chen et al. 2022b; Kim et al. 2022) introduce deformable attention (Zhu et al. 2020) to generate the offsets and weights around the sampling points, and perform sparse cross-modal feature fusion at a small set of sampling locations. However, the pixel-level offsets adopted in these fusion strategies have difficulty in capturing the joint-level feature dependencies, while the correlation between long-range joints has been proven to be crucial in 3D hand pose estimation (Lin, Wang, and Liu 2021a,b).

Second, the frequent movement and high degrees of freedom of human hands, coupled with frequent interactions with objects in real-world scenes, leads to severe feature ambiguity, such as similarity in color and texture appearance, as well as depth noise. Previous fusion methods directly fuse the multimodal features through the pixel-wise weighted aggregation or cross-attention mechanism, which are inherently data-driven and thus can be interfered with by the aforementioned low-quality features.

The representation of keypoints is adopted in some 3D hand pose estimation methods (Hampali et al. 2022) to represent potential joint locations, which unveils more informative feature regions in the image. By focusing cross-modal fusion on these spatially sparse yet informative regions, it is feasible to reduce redundant interactions involving invalid pixels, thereby enhancing the efficiency of multimodal fusion. In addition, there exists more reliable information in the respective modality, which can be explicitly extracted as robust disambiguation evidence to help alleviate the interference of intra-modal ambiguity on cross-modal fusion.

Inspired by the above motivation, we propose Keypoint-Fusion, an RGB-D fusion approach for 3D hand pose estimation. We first propose a Keypoint Feature Aggregation Module (KFAM) to aggregate RGB and depth local features around the initially predicted joints. During RGB feature aggregation, the depth modality provides 3D geometric structure information. Conversely, during depth feature aggregation, the RGB modality complements the precise edge information lost due to depth holes and noise. Then, based on disambiguated and sparse keypoint features, we employ cross-modal feature interaction to model long-range feature correlations efficiently. Code is available at https://github.com/ru1ven/KeypointFusion.

The main contributions of our work are threefold:

1) We propose a sparse RGB-D fusion approach for 3D hand pose estimation, which performs cross-modal feature interaction and fusion in a more efficient way.

2) We introduce a cross-modal information guidance strategy, which can excavate the unique advantages of complementary modalities to clarify intra-modal ambiguous information before cross-modal feature interaction.

3) Experiments show that our method achieves leading performance on two challenging datasets, DexYCB (Chao et al. 2021) and HO-3D (Hampali et al. 2020), significantly outperforming previous state-of-the-art (SOTA) methods.

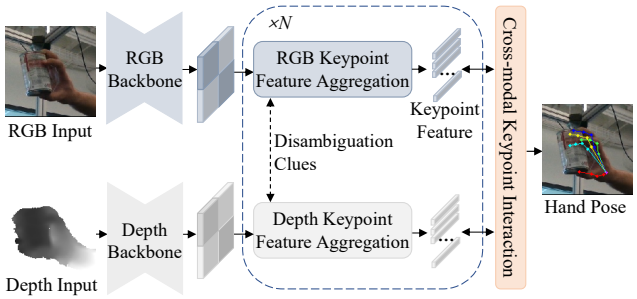


Figure 2: Overview of Keypoint-Fusion.

Related Work

Depth-based 3D Hand Pose Estimation

Existing depth-based 3D hand pose estimation methods can be categorized into 2D image-based and 3D data-based methods according to the input data. Early methods (Huang et al. 2020c; Ren et al. 2019; Fang et al. 2020; Du et al. 2019) use the 2D convolutional neural network (CNN) to extract visual features and estimate the hand pose from single-channel depth images. To overcome the lack of geometric information in 2D images, some works (Huang et al. 2020a; Ge, Ren, and Yuan 2018; Malik et al. 2020, 2021) use 3D CNN and point cloud networks to process 3D point cloud or volumetric representation converted by depth data. Recently, IPNet (Ren et al. 2023) estimates the initial hand pose from depth images and refines the hand pose through the point cloud. However, due to the inherent defects of depth sensors, hands are prone to depth holes and noise, which reduces the performance of depth-based methods.

RGB-based 3D Hand Pose Estimation

RGB-based methods present more challenges due to the lack of depth information. Some methods (Zheng et al. 2021b; Moon and Lee 2020; Iqbal et al. 2018) introduce novel per-pixel representations such as 2.5D heatmap to resolve scale and depth ambiguities. In addition, some works (Lin, Wang, and Liu 2021a,b; Chen et al. 2021; Ge et al. 2019) regress 3D joint and hand mesh from RGB feature using Graph Convolution Network (GCN) or Transformer. Recently, several works (Hampali et al. 2022; Park et al. 2022) use Transformers to model the interaction of non-local image features to enhance the robustness against occlusion. However, RGB-based methods face challenges due to the appearance similarity and geometric ambiguities of visual features.

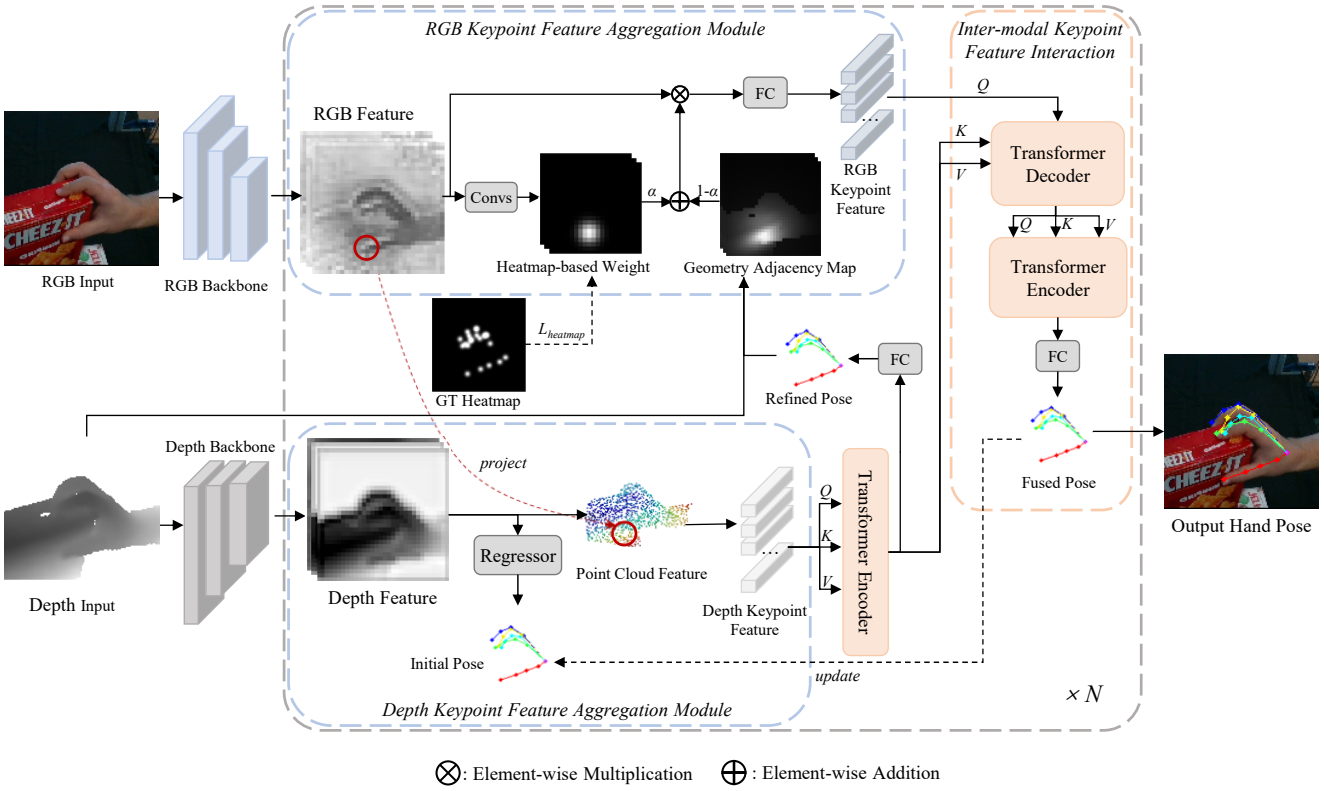


Figure 3: The details of Keypoint-Fusion. Keypoint-Fusion first extracts RGB-D visual features and predicts the initial hand pose. Then, the proposed KFAM aggregates RGB and depth local features around the joints, during which the unique advantages of complementary modalities are leveraged to clarify intra-modal ambiguous information. Finally, Keypoint-Fusion performs sparse cross-modal interaction between the aggregated keypoint feature.

Multi-Sensor Fusion

Recently, multi-sensor fusion methods have been widely studied in many computer vision tasks, such as RGB-D semantic segmentation and 3D object detection. Early works (Hu et al. 2019; Sun et al. 2020; Chen et al. 2020; Seichter et al. 2021) construct two parallel backbones for RGB and depth images separately, employing attention mechanisms to fuse RGB-D visual features. Inspired by the recent superior success of Transformers in vision tasks, many works employ Transformer to fuse features of homogeneous modalities (e.g. RGB-D images) and heterogeneous modalities (e.g. camera and LiDAR point clouds). CMX (Liu et al. 2022a) comprehensively uses soft attention and cross-attention for RGB-X semantic segmentation. TransFusion (Bai et al. 2022) fuses the object queries from LiDAR point cloud features and the image features with cross-attention. FUTR3D (Chen et al. 2022a) uses 3D object queries to sample and aggregate multimodal features from multi-sensors. AutoAlign (Chen et al. 2022c) models the mapping relationship between the image and point clouds for pixel-level and instance-level feature fusion. DeMF (Yang et al. 2022), AutoAlignv2 (Chen et al. 2022b) and 3D Dual-Fusion (Kim et al. 2022) introduce deformable attention (Zhu et al. 2020) to perform sparse LiDAR-camera feature fusion.

Method

Fig. 2 illustrates the overall pipeline of Keypoint-Fusion. Given RGB and depth input images, we first use two parallel 2D CNN backbones to extract RGB and depth visual features, and predict initial 3D hand pose. Then, the proposed Keypoint Feature Aggregation Module individually aggregates the RGB and depth features around the joints, and iteratively updates the hand pose. During feature aggregation, we use the precise geometric structure information of depth modality to eliminate color feature ambiguity, and use the fine-grained edge information of RGB modality to supplement depth holes and noise. Finally, we perform long-range cross-modal feature interaction on sparse keypoint features.

Initial Hand Pose Estimation

Firstly, we extract the RGB and depth visual features and estimate the initial 3D hand pose. Specifically, given RGB and depth images, we first adopt two ResNet-18 (He et al. 2016) as the RGB and depth CNN backbones to extract the RGB visual features $F_{2d}^{RGB} \in \mathbb{R}^{H \times W \times C}$ and depth visual features $F_{2d}^{D} \in \mathbb{R}^{H \times W \times C}$ respectively. Then, we predict 2D heatmap $H_{2d} \in \mathbb{R}^{H \times W \times J}$ and initial 3D hand pose $J_{init} \in \mathbb{R}^{3 \times J}$ from the depth feature map through the weighted average regression (Huang et al. 2020c).

Keypoint Feature Aggregation Module

As shown in Fig. 3, we aggregate RGB and depth local features of each joint complying with the initial hand pose. During aggregation, RGB-D features mutually resolve intra-modal ambiguous information by leveraging reliable disambiguation clues from complementary modalities.

RGB Keypoint Feature Aggregation The ambiguity of similar colors and the lack of local geometric information are challenging problems of RGB local feature aggregation. For example, the depth ordering of coupled fingers is indistinguishable due to the color similarity. Additionally, the weights of the identical RGB feature pixels are ambiguous due to the possibility of visual alignment but varying depths among them. These ambiguities can be resolved by utilizing the precise local geometric information of each joint from depth data. Therefore, we introduce a per-joint local geometric structure representation by calculating the real-world distance between each predicted joint and feature pixels.

Specifically, we propose a learnable weighted aggregation operation for RGB feature aggregation. This operation jointly utilizes the 2D heatmaps and geometry structure information to generate the weights. Firstly, we concatenate 2D heatmap H_{2d} and RGB features F_{2d}^{RGB} along the channel dimension, and then apply N 1×1 convolution kernels to generate the heatmap-based weight map $W_{hm} \in \mathbb{R}^{H\times W\times J}$:

$$W_{hm} = \phi(F_{2d}^{RGB} \| H_{2d}), \tag{1}$$

where N, ϕ, and $\|$ represent the number of hand joints, the convolution layer, and the concatenation operation, respectively. In practice, during the first few training epochs, we employ the ground truth 2D heatmap to supervise the generation of heatmap-based weight map W_{hm}.

Then, by downsampling the depth image to the same size as the RGB feature map and converting it to the real-world coordinate system through the camera intrinsic parameters, we can obtain the real-world coordinates (x_p, y_p, z_p) of each pixel point p of the RGB visual feature map. According to the 3D coordinates (x_j, y_j, z_j) of each joint J of the initial hand pose J_{init}, the geometry adjacency map $W_{geo} \in \mathbb{R}^{H\times W\times J}$ can be generated by calculating the 3D Euclidean distance from RGB visual feature pixels to each hand joint:

$$W_{geo}^{J}(i,j) = \frac{1}{\gamma((x_p - x_j)^2 + (y_p - y_j)^2 + (z_p - z_j)^2)}, \tag{2}$$

where γ reflects the speed at which the correlation between joints and feature pixels decreases as their distance increases, and we set it to 10. The geometry adjacency map W_{geo} presents rich local geometric structure information of each joint neighborhood and effectively reflects the correlation between joints and feature pixels. Finally, the RGB keypoint feature $K_{RGB} \in \mathbb{R}^{J\times C}$ is aggregated according to the heatmap-based weight map W_{hm} and geometry adjacency map W_{geo}, which can be generalized as:

$$K_{RGB} = FC((\alpha W_{hm} + (1-\alpha) W_{geo}) F_{2d}^{RGB}), \tag{3}$$

where FC represents Fully-Connected (FC) layer, and α is a learnable parameter for adjusting the contribution of the heatmap-based weight map and geometry adjacency map.

Depth Keypoint Feature Aggregation The 2D image representation of depth data leads to the loss of 3D geometric structure, e.g., points that are far away in the real world may be adjacent in the depth images (Liu et al. 2022c). To address this problem, IPNet (Ren et al. 2023) proposes to fuse point cloud and 2D depth features and generate the point cloud features $F^{3d} \in \mathbb{R}^{(H\times W)\times C}$ through a 2D-3D projection module, which can effectively perceive the 3D geometric information of point cloud data. However, due to the inherent defects of depth sensors, depth data is prone to noise and depth holes at the edges of hands and objects, especially during motion. It is difficult to compensate for these edge noises by relying solely on the intra-modal fusion of different depth representations. On the other hand, RGB modality can provide more reliable edge information and texture information in most cases. Thus, we project the RGB feature into the 3D point cloud space through the real-world coordinates of each RGB feature pixel obtained in RGB KFAM. Then, we fuse the projected RGB image features $F_{proj}^{RGB} \in \mathbb{R}^{(H\times W)\times C}$ with the above point cloud features F^{3d}. Finally, we aggregate the depth keypoint feature $K_D \in \mathbb{R}^{J\times C}$ as:

$$K_D = FA(ReLU(BN(W_0 F_{proj}^{RGB} + W_1 F^{3d}))), \tag{4}$$

where $ReLU$, BN, and FA denote the ReLU activation function, batch normalization layer, and local feature aggregation module of IPNet (Ren et al. 2023); W_0 and W_1 are two learnable parameter matrices used for adjusting the contribution of RGB features and point cloud features.

As shown in Fig. 2, we construct the RGB and depth keypoint feature aggregation as iterative and serial modules, during which we refine the estimated hand pose, and the aggregation and cross-modal fusion in the subsequent stages can be performed based on more accurate positions of joints.

Cross-modal Keypoint Feature Interaction

The attention mechanisms of Transformer have been employed in previous 3D hand pose estimation methods for non-local interaction of image features to enhance the weakened local features caused by occlusion. However, due to the huge number of pixels in the image features, previous Transformer-based methods either perform intensive interaction with high computational complexity (e.g. HandOccNet (Park et al. 2022)), or use a multi-layer Transformer Encoder structure to reduce the dimension step by step, increasing the complexity of the network structure (e.g. METRO (Lin, Wang, and Liu 2021a)). To avoid the high computational costs caused by the dense cross-modal feature fusion and global-scale attention, we perform sparse cross-modal interaction at the joint level based on the obtained RGB and depth keypoint feature. Considering the serial arrangement of RGB-D KFAMs, we first perform intra-modal interaction between the aggregated keypoint feature of depth modality, and then perform inter-modal keypoint interaction to fuse the cross-modal features.

Intra-modal Keypoint Feature Interaction Since the aggregated keypoint feature contains rich local features around

Input Modality	Method	MPJPE↓					PA-MPJPE↓
		S0	S1	S2	S3	Average	
RGB	Spurr et al. (Spurr et al. 2020)	17.34	22.26	25.49	18.44	20.88	6.83
RGB	Liu et al. (Liu et al. 2021a)	15.28	-	-	-	-	6.58
RGB	Tse et al. (Tse et al. 2022)	16.05	21.22	27.01	17.93	20.55	-
RGB	METRO (Lin, Wang, and Liu 2021a)	15.24	-	-	-	-	6.99
RGB	HandOccNet (Park et al. 2022)	14.04	-	-	-	-	5.80
Depth	A2J (Xiong et al. 2019)	23.93	25.57	27.65	24.92	25.52	-
Depth	AWR (Huang et al. 2020c)	11.23	-	-	-	-	-
RGB-D	SA-Fusion (Liu et al. 2023)	9.51	-	-	-	-	-
RGB-D	DiffHand (Li et al. 2023)	12.10	-	-	-	-	4.98
D&Point Cloud	IPNet (Ren et al. 2023)	8.03	9.01	8.60	7.80	8.36	-
RGB-D	**Keypoint-Fusion**	**6.94**	**8.64**	**7.56**	**7.02**	**7.54**	**4.79**

Table 1: Comparison with SOTA methods of the MPJPE and PA-MPJPE (mm) on the DexYCB dataset.

Input	Method	MPJPE↓	MPJPE↓ (align.)
RGB	METRO (Lin, Wang, and Liu 2021a)	-	2.89
RGB	Liu et al. (Liu et al. 2021a)	3.00	3.17
RGB	I2L-MeshNet (Moon and Lee 2020)	2.68	2.60
RGB	Keypoint TR (Hampali et al. 2022)	-	2.57
RGB	Zheng et al. (Zheng et al. 2021a)	-	2.51
RGB	ArtiBoost (Li et al. 2021)	-	2.53
RGB	HandOccNet (Park et al. 2022)	2.49	2.40
RGB	Hampali et al. (Hampali et al. 2020)	-	3.04
RGB	Hasson et al. (Hasson et al. 2019)	5.52	3.18
Voxel	HandVoxNet++ (Malik et al. 2021)	2.46	-
RGB-D	DiffHand (Li et al. 2023)	2.37	-
D&PCL	IPNet (Ren et al. 2023)	1.81	2.01
RGB-D	**Keypoint-Fusion**	**1.79**	**1.87**

Table 2: Comparison with SOTA methods of the MPJPE (cm) before and after scale-translation alignment on HO-3D.

joints while lacking long-range interactions, we first perform intra-modal keypoint-level interaction utilizing the self-attention mechanism to model the long-range correlation of keypoint feature. As shown in Fig. 3, for the input depth keypoint feature $K_D \in \mathbb{R}^{C\times J}$, we employ a single Transformer encoder to perform self-attention between keypoint feature, and then use an FC layer to estimate the refined hand pose J_{refine}.

Inter-modal Keypoint Feature Interaction We model cross-modal keypoint correlation through a Transformer decoder and a Transformer encoder. Specifically, we take the RGB keypoint feature $K_{RGB} \in \mathbb{R}^{J\times C}$ as query Q, and take the depth keypoint feature $K_D \in \mathbb{R}^{J\times C}$ which models intra-modal long-range keypoint correlation as Key K and Value V. Then, the transformer decoder performs cross-attention to fuse the RGB keypoint feature with the depth keypoint feature that contains rich geometry information and long-range correlation. Next, the Transformer encoder performs self-attention to model the non-local dependency of fused keypoints, Finally, we apply an FC layer on the fused keypoint feature to estimate the final hand pose J_{fuse}.

Input	Method	MPJPE↓
Depth	2D CNN baseline (Huang et al. 2020c)	11.23
RGB-D	1-stage Keypoint-Fusion	7.23
RGB-D	2-stages Keypoint-Fusion	6.94
RGB-D	3-stages Keypoint-Fusion	6.89

Table 3: Comparison of MPJPE (mm) among different numbers of fusion stages of Keypoint-Fusion on DexYCB.

Experiments

Experiments Setup

DexYCB dataset DexYCB is a hand-object dataset captured by multiple RGB-D cameras, containing 582K RGB-D frames over 1,000 sequences of 10 subjects grasping 20 different objects from 8 views. DexYCB has four official dataset splits of train/val/test, namely S0, S1, S2, and S3, split by the sequences, subjects, views, and objects, respectively. We conduct performance comparisons on all four splits and use the default S0 split in ablation studies.

HO-3D dataset HO-3D is an RGB-D hand-object interaction dataset, containing 66,034 training images and 11,524 test images from a total of 68 sequences. The sequences are captured in multi-camera and single-camera setups and contain 10 different subjects manipulating 10 different objects from YCB dataset. Evaluation of the HO-3D test set is conducted at an online submission server.

Implementation Details Our experiments are conducted with an NVIDIA RTX 4090 GPU. The network is implemented based on PyTorch (Paszke et al. 2019). We use an AdamW optimizer (Kulon et al. 2019) with an initial learning rate of 8e-4. The whole training process takes 15 and 25 epochs on DexYCB and HO-3D, respectively. For data augmentation, we crop the input RGB-D images to the size of 128×128, and perform random rotation ∈ [-180, 180], random scaling ∈ [0.9, 1.1], and random translating ∈ [-10, 10].

Input	Method	IF	MPJPE↓
RGB-D	SA-Fusion (Liu et al. 2023)	7.54	9.51
D&PCL	IPNet (Ren et al. 2023)	19.91	8.03
RGB-D	Dense Fusion baseline	16.19	9.76
RGB-D	**1-stage Keypoint-Fusion**	**12.47**	**7.23**

Table 4: Comparison of the inference time (ms) and MPJPE (mm) with the fusion-based 3D hand pose estimation methods on DexYCB. *IF* represents the inference time.

ID	RGB KFAM		Depth KFAM			MPJPE↓
	HWM	GAM	PCL	RGB	REF	
0						11.23
1	✓	✓				7.54
2	✓	✓	✓		✓	7.17
3	✓	✓	✓	✓		7.11
4			✓	✓	✓	7.09
5	✓		✓	✓	✓	6.97
6		✓	✓	✓	✓	6.96
7			✓		✓	8.76
8	✓		✓		✓	7.19
9		✓	✓		✓	7.27
10	✓	✓	✓	✓	✓	6.94

Table 5: Ablation study for the KFAM on DexYCB. *HWM* and *GAM* represent heatmap-based weight map and geometry adjacency map. *PCL* and *RGB* represent the aggregation of point cloud feature and projected RGB feature. *REF* represents pose refinement with the depth keypoint feature.

We evaluate our method using the metric of Mean Per Joint Position Error (MPJPE).

Comparisons with State-of-the-arts

We compare the performance with SOTA 3D hand pose estimation methods on DexYCB. As shown in Table 1, Keypoint-Fusion achieves the lowest MPJPE on all four official splits of DexYCB. Compared with the SOTA method IPNet (Ren et al. 2023), our method achieves an average MPJPE reduction of 13.6% (from 8.03mm to 6.94mm) on the default S0 split. Additionally, several RGB-based methods are also evaluated in terms of the MPJPE after Procrustes Alignment (PA-MPJPE) on DexYCB S0 split, and our method achieves the best performance on all metrics.

The performance comparison with SOTA methods on the HO-3D dataset is shown in Table 2. We adopt two ConvNeXt-T (Liu et al. 2022b) as the 2D backbones. We report the hand pose results based on MPJPE and MPJPE after scale-translation alignment of the root joint, which are two significant metrics of 3D hand pose estimation. As can be seen, both for the MPJPE before and after alignment, our method achieves the best performance on HO-3D.

Ablation Study

Number of Iterative Fusion Stages We design the keypoint feature aggregation as iterative and serial modules, so the cross-modal interaction in the later stages can be performed based on more accurate joint positions. To verify the effectiveness of the above iterative fusion design, we conduct an ablation study on Keypoint-Fusion with various numbers of fusion stages on DexYCB. For the 2D CNN baseline, we adopt ResNet-18 as the backbone. As shown in Table 3, our model reduces the Mean Per Joint Position Error (MPJPE) by 4.0mm through a single fusion stage compared with the 2D CNN baseline. In addition, by increasing the number of fusion stages, the performance can be further significantly improved.

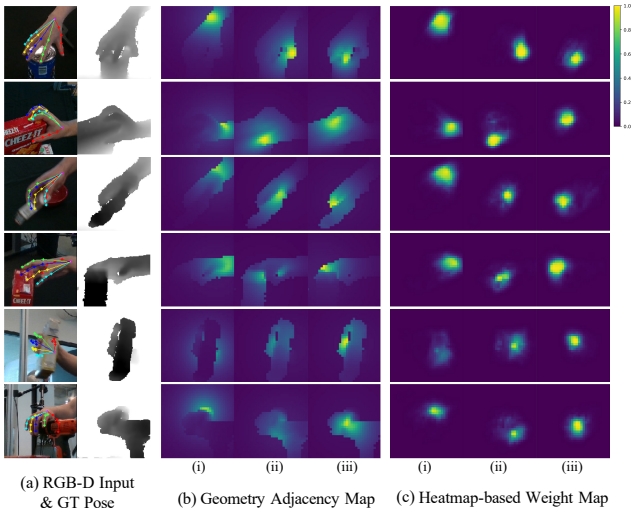
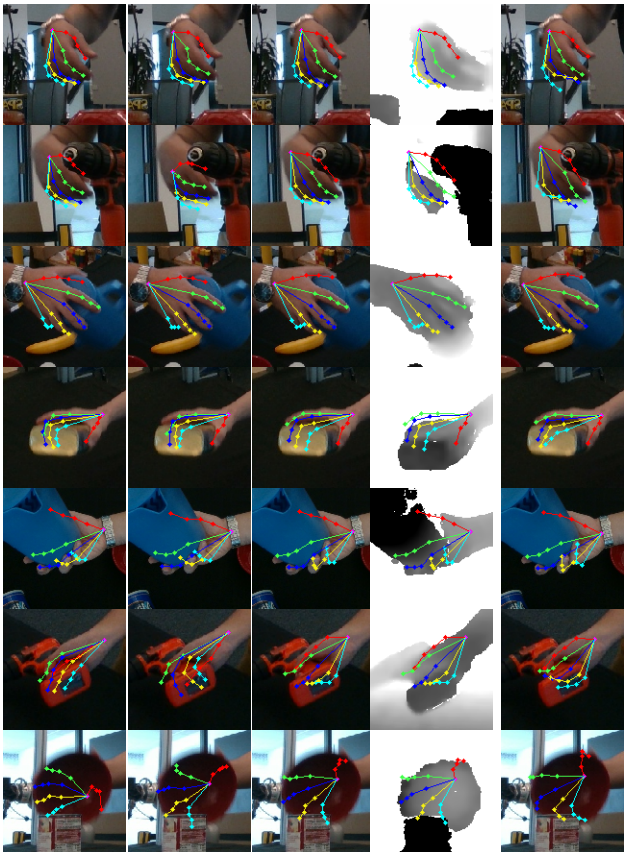


Figure 4: Visualization of the proposed geometry adjacency map and heatmap-based weight map. We show (i) wrist, (ii) thumb tip, and (iii) index tip among 21 joints of DexYCB.

Efficiency Analysis To verify the efficiency of the proposed sparse fusion strategy, we conduct an ablation study on the inference time with SA-Fusion (Liu et al. 2023), an existing RGB-D image-based fusion method, and with IPNet (Ren et al. 2023), which constructs point cloud features efficiently. Additionally, we consider a dense fusion baseline by ablating the proposed KFAMs and performing cross-modal features interaction in a dense manner. Table 4 demonstrates that: (i) Compared with previous RGB-D image-based fusion method, our method achieves significant performance improvements with an acceptable additional inference time; (ii) Compared with IPNet, our method enables cross-modal fusion in 3D space and shows higher efficiency and effectiveness. (iii) Compared with the dense fusion baseline, our method can perform cross-modal feature interaction in a more efficient way.

Design of Keypoint Feature Aggregation Module We ablate various designs of the KFAM as illustrated in Table 5. First, by individually adopting RGB KFAM (ID 1) or depth KFAM (ID 4), the performance of the network is significantly improved compared with the depth-only based baseline. Second, discarding RGB feature fusion (ID 2) or hand pose refinement (ID 3) in the depth feature aggregation causes performance degradation. Next, the performance of



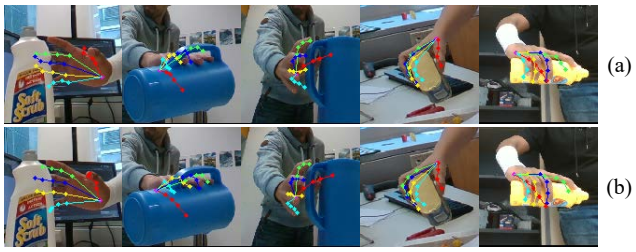
(a) IPNet (b) Initial (c) Keypoint-Fusion (d) GT

Figure 5: Qualitative results on the DexYCB dataset.

the network can be reduced without adopting the geometry adjacency map (ID 5) or heatmap-based weight map (ID 6) in RGB feature aggregation. Furthermore, without the complement of local RGB features in the depth KFAM, the performance gap among different designs of the RGB KFAM can be further widened (ID 7-9).

Visualization of Keypoint-Fusion

To explore the perception ability of the proposed RGB KFAM to local feature information, we visualize the geometry adjacency map W_{geo} and heatmap-based weight map W_{hm} in Fig. 4. In ordinary cases (row 1 to row 4), the heatmap-based weight map focuses on the relevant spatial regions around the joints in terms of color features, and the geometry adjacency map provides local geometric structure information to help distinguish between joints and invalid pixels, such as backgrounds, objects in contact, and other hand regions. For occluded joints (row 5 and thumb tip in row 6), visual features with similar colors and adjacent distances are comprehensively considered, and the geometry adjacency map suppresses features in occluded regions that are visually adjacent but far away from the joint. In addition, when it is hard to determine the relative distance between joints from single RGB images due to hand color similarity, such as the wrist and index tip in row 6, the geometry adjacency map pinpoints relevant feature regions.



(a)

(b)

Figure 6: Qualitative results of (a) the initial hand pose and (b) the output hand pose of Keypoint-Fusion on HO-3D.

Qualitative Results

We present the qualitative results on DexYCB in Fig. 5. The results of our method are additionally shown on the depth map to intuitively demonstrate its robustness to depth holes and noise. Compared with the SOTA IPNet, our method demonstrates superior performance in handling challenging samples with noise and depth holes (row 1, row 3, row 4), motion blur (row 2), and severe occlusion (row 5, row 6, row 7). Our method accurately predicts the position of visually blurred joints and mitigates the impact of insufficient depth information. Additionally, ours achieves a more accurate estimation of unseen joints. Meanwhile, through cross-modal disambiguation and interaction, our method can effectively refine the initial hand pose estimated by the 2D CNN backbone in the presence of edge noise and object occlusion.

The qualitative results on HO-3D are shown in Fig. 6. Since the ground truth joint positions of the HO-3D test set are not provided, we only present the initial hand pose and output hand pose estimated by our method. Compared with the initial hand pose, our method more accurately estimates the joints that are occluded during interaction with objects.

Conclusion

In this work, we propose Keypoint-Fusion for RGB-D based 3D hand pose estimation, which effectively eliminates intra-modal ambiguous information and efficiently performs cross-modal feature interaction. We first propose a Keypoint Feature Aggregation Module to aggregate local features around the hand joints, and leverage the inherent advantages of complementary modalities to provide disambiguation clues mutually. In particular, during RGB aggregation, we construct a per-joint local geometric structure representation using depth data, to clarify the ambiguous weights of RGB feature pixels. In turn, during depth aggregation, we project RGB features to the 3D point cloud space to complement the edge information lost due to depth holes and noise. Then, our method efficiently performs cross-modal feature interaction based on sparse keypoints. Experiments on DexYCB and HO-3D datasets demonstrate that our method significantly outperforms other SOTA methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants (62171057, 62101064, 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, in part by the Project funded by China Postdoctoral Science Foundation (2023TQ0039).

References

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, 1090–1099.

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

Chen, X.; Lin, K.-Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; and Zeng, G. 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *ECCV*, 561–577. Springer.

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

Chen, X.; Zhang, T.; Wang, Y.; Wang, Y.; and Zhao, H. 2022a. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; and Zhao, F. 2022b. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. *arXiv preprint arXiv:2207.10316*.

Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; Zhao, F.; Zhou, B.; and Zhao, H. 2022c. Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection. *arXiv preprint arXiv:2201.06493*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, K.; Lin, X.; Sun, Y.; and Ma, X. 2019. Crossinfonet: Multi-task information sharing based hand pose estimation. In *CVPR*, 9896–9905.

Fang, L.; Liu, X.; Liu, L.; Xu, H.; and Kang, W. 2020. Jgrp2o: Joint graph reasoning based pixel-to-offset prediction network for 3d hand pose estimation from a single depth image. In *ECCV*, 120–137. Springer.

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

Ge, L.; Ren, Z.; and Yuan, J. 2018. Point-to-point regression pointnet for 3d hand pose estimation. In *ECCV*, 475–491.

Hampali, S.; Rad, M.; Oberweger, M.; and Lepetit, V. 2020. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 3196–3206.

Hampali, S.; Sarkar, S. D.; Rad, M.; and Lepetit, V. 2022. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, 11090–11100.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, X.; Yang, K.; Fei, L.; and Wang, K. 2019. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing*, 1440–1444. IEEE.

Huang, L.; Tan, J.; Liu, J.; and Yuan, J. 2020a. Hand-transformer: non-autoregressive structured modeling for 3d hand pose estimation. In *ECCV*, 17–33. Springer.

Huang, L.; Tan, J.; Meng, J.; Liu, J.; and Yuan, J. 2020b. Hot-net: Non-autoregressive transformer for 3d hand-object pose estimation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3136–3145.

Huang, W.; Ren, P.; Wang, J.; Qi, Q.; and Sun, H. 2020c. Awr: Adaptive weighting regression for 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11061–11068.

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

Kim, Y.; Park, K.; Kim, M.; Kum, D.; and Choi, J. W. 2022. 3D Dual-Fusion: Dual-Domain Dual-Query Camera-LiDAR Fusion for 3D Object Detection. *arXiv preprint arXiv:2211.13529*.

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

Kulon, D.; Wang, H.; Güler, R. A.; Bronstein, M.; and Zafeiriou, S. 2019. Single image 3d hand reconstruction with mesh convolutions. *arXiv preprint arXiv:1905.01326*.

Li, K.; Yang, L.; Zhan, X.; Lv, J.; Xu, W.; Li, J.; and Lu, C. 2021. ArtiBoost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. *arXiv preprint arXiv:2109.05488*.

Li, L.; Zhuo, L.; Zhang, B.; Bo, L.; and Chen, C. 2023. Diff-Hand: End-to-End Hand Mesh Reconstruction via Diffusion Models. *arXiv preprint arXiv:2305.13705*.

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

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

Liu, H.; Zhang, J.; Yang, K.; Hu, X.; and Stiefelhagen, R. 2022a. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*.

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

Liu, X.; Ren, P.; Chen, Y.; Liu, C.; Wang, J.; Sun, H.; Qi, Q.; and Wang, J. 2023. SA-Fusion: Multimodal Fusion Approach for Web-Based Human-Computer Interaction in the Wild. WWW '23, 3883–3891. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *ICCV*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A convnet for the 2020s. In *CVPR*, 11976–11986.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.; and Han, S. 2022c. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. *arXiv preprint arXiv:2205.13542*.

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

Malik, J.; Shimada, S.; Elhayek, A.; Ali, S. A.; Theobalt, C.; Golyanik, V.; and Stricker, D. 2021. Handvoxnet++: 3d hand shape and pose estimation using voxel-based neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8962–8974.

Moon, G.; and Lee, K. M. 2020. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 752–768. Springer.

Park, J.; Oh, Y.; Moon, G.; Choi, H.; and Lee, K. M. 2022. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *CVPR*, 1496–1505.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Ren, P.; Chen, Y.; Hao, J.; Sun, H.; Qi, Q.; Wang, J.; and Liao, J. 2023. 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*.

Ren, P.; Sun, H.; Qi, Q.; Wang, J.; and Huang, W. 2019. SRN: Stacked Regression Network for Real-time 3D Hand Pose Estimation. In *BMVC*, 112.

Seichter, D.; Köhler, M.; Lewandowski, B.; Wengefeld, T.; and Gross, H.-M. 2021. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation*, 13525–13531. IEEE.

Spurr, A.; Iqbal, U.; Molchanov, P.; Hilliges, O.; and Kautz, J. 2020. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *ECCV*, 211–228. Springer.

Sun, L.; Yang, K.; Hu, X.; Hu, W.; and Wang, K. 2020. Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. *IEEE robotics and automation letters*, 5(4): 5558–5565.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *ECCV*, 3–19.

Xiong, F.; Zhang, B.; Xiao, Y.; Cao, Z.; Yu, T.; Zhou, J. T.; and Yuan, J. 2019. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *ICCV*, 793–802.

Yang, H.; Shi, C.; Chen, Y.; and Wang, L. 2022. Boosting 3D Object Detection via Object-Focused Image Fusion. *arXiv preprint arXiv:2207.10589*.

Zheng, X.; Ren, P.; Sun, H.; Wang, J.; Qi, Q.; and Liao, J. 2021a. Joint-aware regression: Rethinking regression-based method for 3d hand pose estimation. In *British Machine Vision Conference*, volume 10.

Zheng, X.; Ren, P.; Sun, H.; Wang, J.; Qi, Q.; and Liao, J. 2021b. 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*, 99–108. IEEE.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.