2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Pose Relation Transformer Refine Occlusions for Human Pose Estimation

Hyung-gun Chi*[1], Seunggeun Chi*[1], Stanley Chan[1], Karthik Ramani[1]

***Abstract*— Accurately estimating the human pose is an essential task for many applications in robotics. However, existing pose estimation methods suffer from poor performance when occlusion occurs. Recent advances in NLP have been very successful in predicting the missing words conditioned on visible words. We draw upon the sentence completion analogy in NLP to guide our model to address occlusions in the pose estimation problem. We propose a novel approach that can mitigate the effect of occlusions motivated by the sentence completion task of NLP. In an analogous manner, we designed our model to reconstruct occluded joints given the visible joints utilizing joint correlations by capturing the implicit joint connectivity through the attention mechanism. In this work, we propose a *POse Relation Transformer* (PORT) that captures the global context of the pose using self-attention and a local context by aggregating adjacent joint features. To supervise PORT in learning joint correlations, we guide PORT to reconstruct randomly masked joints, which we call Masked Joint Modeling (MJM). PORT trained with MJM adds to existing keypoint detection methods and successfully refines occlusions. Notably, PORT is a model-agnostic plug-and-play module for pose refinement under occlusion that can be plugged into any keypoint detector with substantially low computational costs. We conducted extensive experiments to demonstrate the advantage of PORT mitigating the occlusion on the hand and body pose PORT improves the pose estimation accuracy of existing human pose estimation methods by up to 16% with only 5% of additional parameters. The code is publicly available at https://github.com/stnoah1/PORT.**

I. INTRODUCTION

Human pose estimation (HPE) has attracted significant interest due to its importance to various tasks in robotics, such as human-robot interaction [1], [2], hand-object interaction in AR/VR [3], imitation learning for dexterous manipulation [4], and learning from demonstration [5]. However, in a single-view camera setup, various occlusions such as self-occlusion, occlusion by the object, and out-of-frame occur. As a consequence, the occlusion confuses the existing keypoint detectors, an essential intermediate step of pose estimation, and produces incorrect poses (see examples in Fig. 3) that result in errors in applications such as lost tracking and gestural miscommunication in human-robot interaction. In this work, we aim to mitigate the effects of occlusions to provide a more reliable solution for the HPE task. Specifically, we improve the keypoint detection accuracy under occlusion, an important intermediate step for human pose estimation methods.

The recent success of Masked Language Modeling (MLM) [6], a widely used pretraining task in Natural Language Processing (NLP), inspires our work. Inferring occluded joints of the human skeleton is analogous to inferring missing words in a sentence, as illustrated in Fig.1. The objective of MLM is to train the model to predict blank words in a sentence. Through this process, the model learns to capture the contextual relations between the words. Intuitively, the model learns 'where and how much' to attend in order to predict the masked words.

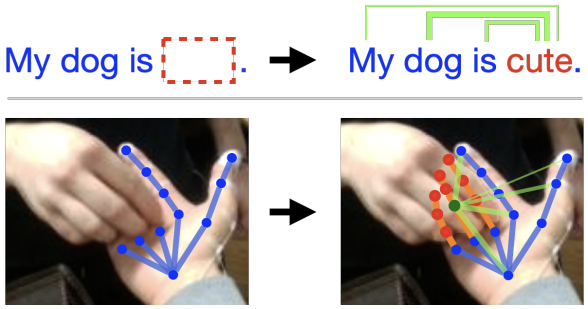

Fig. 1: Analogous to the fill-in-the-blank task that captures word correlations to predict blanked words **(Top)**, PORT predicts occluded joints by capturing the semantic connectivity between hand joints **(Bottom)**.

We propose an occlusion refinement framework using a pose refinement module, which we named POse Relation Transformer (PORT). PORT refines locations of occluded joints from existing keypoint detectors. Like MLM, we train PORT by randomly masking joints and reconstructing them, which we call Masked Joint Modeling (MJM). Through this process, PORT learns to capture joint correlation and utilizes it to reconstruct occluded joints. Then, the trained PORT is used to refine occluded joints when plugged into existing keypoint detectors. We found occluded joints in keypoint detectors tend to have lower confidence and higher errors. Therefore, we improve the detection accuracy by replacing these joints with the reconstructed joints from PORT.

We design the architecture of PORT to capture the global and local context of the human pose since these contexts provide an important clue to reconstructing occluded joints. The architecture of PORT is based on the transformer following the BERT [6] that first introduced MLM. The transformer's self-attention mechanism captures the pose's global semantic context. To further utilize the semantic knowledge embedded in the skeleton, we use graph convolution [7] for the embedding and projection process of the transformer. Graph representation has been widely adopted to model the human skeleton [8]–[11] because of its versatility in capturing physical constraints, relations, and semantics of

[1]School of Electrical and Computer Engineering, Purdue University. {hgchi, sgchi, stanchan, ramani}@purdue.edu
*These authors contributed equally to this work.

the skeleton. Graph convolution enables the PORT to be able to extract the local context along with the global semantics from self-attention.

Notably, PORT has several advantages that make it adaptable to existing keypoint detectors. First, PORT is model-agnostic and therefore be plugged into any keypoint detector. Second, PORT is light-weighted since the input format of PORT is a joint location instead of an image. With only 5% of the parameters, PORT reduces up to 16% of the error of the existing keypoint detector. Lastly, PORT does not require additional finetuning. PORT trained with MJM refines occlusions without further end-to-end training after plugging into the keypoint detector.

To demonstrate the effectiveness of PORT in refining occluded joints, we evaluate PORT on four datasets that cover various occlusion scenarios. We prove that PORT improves the performance of existing keypoint detectors and demonstrates the occlusion refinement ability through analysis. In short, our contributions are as follows,

- We introduce a novel architecture named POse Relation Transformer (PORT) that captures global and local semantic joint relations.
- We propose a model-agnostic occlusion refinement framework. PORT trained with Masked Joint Modeling (MJM) works as a plug-in and refines occluded joints of any keypoint detector.
- Through extensive experiments on various datasets, we prove that PORT mitigates the effect of occlusions and improves the pose estimation accuracies of existing keypoint detectors.

II. Related works

Human Pose Estimation 3D Human pose estimation is categorized into regression-based [12]–[14] and detection-based methods [15]–[18]. Regression-based methods directly predict the 3D coordinates of human joints, whereas detection-based methods estimate 2D keypoints for its intermediate step. Due to their accurate pose estimation performance, detection-based methods are more popular than regression-based methods. However, detection-based methods are more sensitive to occlusion than regression-based methods [19] since the keypoint detectors tend to fail to detect keypoints under occlusion. To alleviate the errors from the occlusion, recent approaches exploit large-scale datasets [15], [20]–[22], or synthetic images [16], [23], [24] that include various occlusion cases. However, the data-driven approaches do not explicitly tackle the occlusion prediction problem. Cheng *et al.* [25] filter out the unreliable estimations of occluded joints using optical flow, Ye *et al.* [26] model different distributions for visible and occluded joints using hierarchical mixture density networks, and *et al.* [27] utilizes the concept of geometric constraints to refine the estimated joints. The mixture density model [26], [28] exploits multi-modal gaussian distribution to mitigate the effect of occlusion and shows successful improvement. However, since they do not consider the context of the pose, output poses are often awkward. Unlike previous approaches, we focus on mitigating the occlusion effect for the existing keypoint detectors by utilizing the semantic connectivity of human skeletons.

Transformer in Computer Vision The Transformer [29] was first introduced to solve the Neural Machine Translation problem in NLP. It has been a de-facto standard for NLP tasks due to its outstanding performance and scalability. The success of the Transformer in NLP has encouraged researchers to exploit the Transformer in other domains, including but not limited to Computer Vision. For example, Transformer variants have been used in image recognition [30], [31], object detection [32], [33], semantic segmentation [34], and action recognition [35], [36], while achieving state-of-the-art performances in each field. In the HPE field, transformer-based approaches [37]–[41] achieve successful performance due to their ability to capture the semantic context. In this work, we utilize a transformer architecture to refine occluded joints by capturing the context of the pose.

III. Method

A. Overview

We propose a PORT to refine occluded joints by capturing the both global and local context of the pose through graph convolution and self-attention. We introduce Masked Joint Modeling (MJM) to supervise PORT in reconstructing randomly masked joints. PORT, trained with MJM, produce refined joints $\hat{\mathbf{J}}$ by replacing occluded joints from a keypoint detector with reconstructed joints from PORT $\mathbf{J}^{\text{recon}}$, as illustrated in Fig. 2.

B. Preliminaries

Keypoint Detection. The goal of this task is to detect the locations of N keypoints $\mathcal{J} = \{\mathbf{j}_n\}_{n=1}^{N}$ from an image $\mathbf{I}$. Detection-based approaches [42]–[46] utilize heatmaps $\mathcal{H} = \{\mathbf{H}_n\}_{n=1}^{N}$ from an image to estimate the pose. A joint location of n-th joint $\mathbf{j}_n$ can be derived from a heatmap, i.e., argmax function $\arg\max_{(i,j)}[\mathbf{H}_n]_{i,j}$ or weighted sum after applying soft-argmax operation [47] to the heatmaps.

$$\mathbf{j}_n = (x_n, y_n) = (\sum_{i}^{W}\sum_{j}^{H} i[\mathbf{H}_n]_{i,j}, \sum_{i}^{W}\sum_{j}^{H} j[\mathbf{H}_n]_{i,j}). \quad (1)$$

A confidence value c_n of the inferred joint is defined as

$$c_n = [\mathbf{H}_n]_{\lfloor x_n \rceil, \lfloor y_n \rceil}, \quad (2)$$

where $0 \leq c_n \leq 1$ and $\lfloor \cdot \rceil$ denotes round operation.

Masked Language Modeling. Masked Language Modeling (MLM) [6] is a widely used pretraining task in NLP to train the model to learn the context of the language. During the training, the words in a sentence are randomly masked, and the model reconstructs the masked words by learning the correlations between the words. Let $\mathcal{W} = \{\mathbf{w}_n\}_{n=1}^{T}$ denotes the sequence of words, and $\mathcal{M}$ denotes a set of masked word indices, then the objective of MLM is to maximize the log-likelihood of masked word $\mathbf{w}_i$ conditioned on visible words $\mathcal{W}_{\text{vis}}$ which are not masked.

$$\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \log p(\mathbf{w}_i|\mathcal{W}_{\text{vis}}) \,. \quad (3)$$

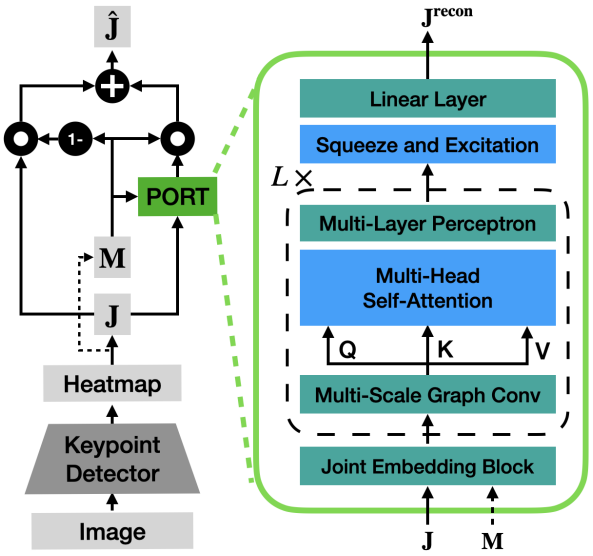

Fig. 2: **(Left)** Overview of proposed occlusion refinement framework and **(Right)** detailed architecture of PORT. $\mathbf{J}$ denotes estimated joints from a keypoint detector, and $\mathbf{M}$ is a mask indicating the occluded joints.

Multi-Scale Graph Convolution. Graph convolution [48] is an effective method to extract skeleton features since the human skeleton can be represented as a graph with joints as nodes and bones as edges. Let the C-dimensional node feature matrix be $\mathbf{X} \in \mathbb{R}^{N \times C}$ and the adjacency matrix be binary matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where $\mathbf{A}_{i,j}$ is 1 if i-th and j-th joins are connected with a bone otherwise 0. Then, graph convolution is formulated as $\tilde{\mathbf{A}}^k\mathbf{X}\mathbf{W}$, where $\tilde{\mathbf{A}}$ is a symmetrically normalized form of $\mathbf{A}+\mathbf{I}$, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{W} \in \mathbb{R}^{C \times C'}$ are learnable weights. Note that we omit non-linear activation since we use graph convolution for feature projection and embedding. Especially, we utilize *Multi Scale Graph Convolution* (MSGC) for PORT that aggregates skeleton features with different kernel sizes. MSGC is formulated as

$$\text{MSGC}(\mathbf{A}, \mathbf{X}) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \tilde{\mathbf{A}}^k \mathbf{X} \mathbf{W}, \tag{4}$$

where $\mathcal{K}$ is a set of exponents for the adjacency matrix.

C. POse Relation Transformer (PORT)

PORT consists of a joint embedding block, an encoder, and a regression head (see Fig. 2 green box). In the embedding block, we transform the skeleton features to the embedding dimension using MSGC and use it as input for the encoder. The encoder is built based on the Transformer [29] encoder, and it captures the global and local context of the pose using self-attention and graph convolution. Lastly, the regression head projects the output of the encoder to the joint location.

Joint Embedding Block. We first transform the skeleton joint locations $\mathbf{J} \in \mathbb{R}^{N \times 2}$ to D-dimensional joint embeddings using Multi-Scale Graph Convolution (MSGC).

$$\mathbf{Z}_{(0)} = \text{MSGC}(\mathbf{A}, \mathbf{J}), \tag{5}$$

where $\mathbf{J}_i = \mathbf{j}_i$ and $\mathbf{Z}_{(l)} \in \mathbb{R}^{N \times D}$ indicates a feature embedding of l-th encoding layer. Unlike the transformer, positional encoding for positional information is not added since the graph convolution employs an adjacency matrix, which implicitly includes positional information.

Encoder. The encoder consists of L encoding layers. It captures the context of the pose utilizing self-attention and graph convolution. To embed the local context, inputs are first transformed to Key, Query, and Value (denoted as $\mathbf{K}, \mathbf{Q}, \mathbf{V} \in \mathbb{R}^{N \times D}$, respectively) using MSGC in each encoding layer.

$$\mathbf{Q}_{(l)}, \mathbf{K}_{(l)}, \mathbf{V}_{(l)} = \text{MSGC}(\mathbf{A}, \mathbf{Z}_{(l-1)}). \tag{6}$$

Then, the attention is calculated as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}})\mathbf{V}. \tag{7}$$

We especially use the Multi-head Self-Attention (MSA) that allows the model to explore different feature representation subspaces. The overall encoding process of the encoding layer is formulated as

$$\mathbf{Z}'_{(l+1)} = \mathbf{Z}_{(l)} + \text{MSA}(\text{LN}(\mathbf{Z}_{(l)})), \tag{8}$$

$$\mathbf{Z}_{(l+1)} = \mathbf{Z}'_{(l+1)} + \text{MLP}(\text{LN}(\mathbf{Z}'_{(l+1)})), \tag{9}$$

where $\text{LN}(\cdot)$ denotes layer normalization [49]. Two linear layers with ReLU [50] for activation is used for MLP.

Regression Head. The regression head transforms the output of the last encoding layer $\mathbf{Z}_{(L)}$ to the joint location. To explicitly model channel inter-dependencies, we use Sequence-and-Excitation (SE) [51] module,

$$\text{SE}(\mathbf{Z}) = \text{Sigmoid}(\text{MLP}(\frac{1}{N}\sum_{i}^{N} \mathbf{Z}_i)), \tag{10}$$

where the output $\text{SE}(\mathbf{Z}) \in \mathbb{R}^{1 \times D}$ is weight for channel. Finally, the entire decoding process is defined as

$$\mathbf{J}^{\text{recon}} = (\text{SE}(\mathbf{Z}_{(L)}) \odot \mathbf{Z}_{(L)})\mathbf{W}', \tag{11}$$

where $\odot$ denotes broadcasted element-wise product and $\mathbf{W}' \in \mathbb{R}^{D \times 2}$ is a linear projection that is learnable.

D. Masked Joint Modeling (MJM)

We propose MJM, a training strategy for PORT. The objective of MJM is to reconstruct masked joints given visible joints. We randomly select joint indices for the masking ($\mathcal{M}$) and train the PORT to reconstruct masked joints. Similar to [31], rather than masking the input joints, we replace corresponding rows of joint embedding $\mathbf{Z}_{(0)}$ with a learnable mask embedding $\mathbf{E}^{\text{mask}} \in \mathbb{R}^{1 \times D}$. To train PORT, we set the target distribution of i-th joint to follow two dimensional gaussian $\mathcal{N}_i(\mu_i, \sigma_i\mathbf{I})$ with a ground truth joint location as a center $\mu_i = \mathbf{J}_i^{GT}$ and a fixed variance $\sigma_i = 1$. Then, PORT is trained to minimize reconstruction loss $\mathcal{L}$, defined as negative gaussian log-likelihood.

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \frac{||\mathbf{J}_i^{\text{recon}} - \mu_i||_2^2}{\sigma_i^2} = \sum_{i \in \mathcal{M}} \frac{||\mathbf{J}_i^{\text{recon}} - \mathbf{J}_i^{GT}||_2^2}{|\mathcal{M}|}. \tag{12}$$

Methods	FPHB [20]		CMU panoptic [52]		RHD [53]	
	EPE ↓	P-EPE ↓	EPE ↓	P-EPE ↓	EPE ↓	P-EPE ↓
HRNet_w48 [43]	8.49	5.25	13.88	6.91	5.89	2.15
+PORT	**8.19** (-0.30)	**4.82** (-0.46)	**13.85** (-0.03)	**6.53** (-0.38)	**5.86** (-0.03)	**2.08** (-0.07)
HRNetv2_w18 [46]	8.31	5.01	15.52	6.27	6.30	2.25
+PORT	**7.61** (-0.08)	**4.54** (-0.47)	**15.42** (-0.10)	**6.16** (-0.11)	**6.28** (-0.02)	**2.20** (-0.05)
MobileNetv2 [54]	9.57	6.29	15.27	7.76	6.96	2.75
+PORT	**8.94** (-0.63)	**5.27** (-1.02)	**15.15** (-0.12)	**7.46** (-0.30)	**6.96** (-0.00)	**2.69** (-0.06)
ResNet50 [55]	10.59	6.32	13.63	7.16	6.45	2.34
+PORT	**10.39** (-0.20)	**5.96** (-0.36)	**13.62** (-0.01)	**6.86** (-0.30)	**6.43** (-0.02)	**2.31** (-0.03)

Methods	H36M [56]		H36M_masked	
	EPE ↓	P-EPE ↓	EPE ↓	P-EPE ↓
HRNet_w32 [43]	10.10	8.56	20.05	16.99
+PORT	**9.86** (-0.24)	**8.16** (-0.40)	**19.24** (-0.81)	**16.09** (-0.90)
HRNet_w48 [43]	7.60	6.29	15.07	12.65
+PORT	**7.52** (-0.08)	**6.15** (-0.14)	**14.48** (-0.59)	**11.97** (-0.68)

TABLE I: Keypoint detection performance comparison for various keypoint detectors with and without PORT on **(Top)**: hand and **(Bottom)**: human body test sets. Bold figures indicate the results with PORT, and blue figures denote the improvement.

E. Occlusion Refinement

Here we propose a model-agnostic pose refinement framework for occlusion using PORT trained with MJM (see Fig. 2). We observe that estimated joints from the keypoint detector tend to have low confidence under occlusion, leading to high pose estimation error (see Fig. 4). Therefore, by refining the joints with low confidence, overall performance can be improved. To do so, we mask the estimated joints with low confidence and then reconstruct them using PORT. We add PORT at the end of the keypoint detector and let the PORT refine the estimated joint from the keypoint detector based on their confidence values. We define occluded joints as the set of joints whose confidence c_n from the keypoint detector is less than the predefined threshold δ based on the statistic of the training set.

$$m_n = \begin{cases} 1 & \text{if } c_n < \delta, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Finally, the refined joint $\hat{\mathbf{J}}$ is derived by replacing joint with low confidence $\mathbf{J}$ with reconstructed joint $\mathbf{J}^{\text{recon}}$ from PORT.

$$\hat{\mathbf{J}} = (1 - \mathbf{M}) \odot \mathbf{J} + \mathbf{M} \odot \mathbf{J}^{\text{recon}}, \tag{14}$$

where $\mathbf{M} \in \mathbb{R}^{N \times 1}$ is a masking matrix with $\mathbf{M}_n = m_n$.

IV. EXPERIMENT

To demonstrate the effectiveness of PORT under occlusion, we carried out the keypoint detection task by adding PORT to existing keypoint detectors. To cover various occlusion scenarios, we test PORT on hand (FPHB [20], CMU panoptic [52], RHD [53]) and body (Human 3.6M [56] and Human 3.6M with synthetic mask) datasets.

We evaluate our results using two metrics, End Point Error (EPE) and Procrustes analysis End Point Error (P-EPE). EPE quantifies the pixel differences between the ground truth and the predicted results. P-EPE quantifies the pixel differences after aligning the prediction with the ground truth via a rigid transform. We use P-EPE for all our analysis and ablation studies since it properly reflects occlusion refinement by measuring the pose similarity.

A. Datasets

FPHB [20] First-Person Hand action Benchmark (FPHB) is a collection of egocentric videos of hand-object interactions. We select the dataset to explore the scenario of self-occlusion and occlusion by the object. We use *action*-split of FPHB in our experiments.

CMU Panoptic [52] CMU Panoptic dataset contains third-person view hand images. We select this dataset to test PORT to various scenarios in third-person view images.

RHD [53] Rendered Hand pose Dataset (RHD) contains rendered human hands and their keypoints, which comprised 41,258 training and 2,728 testing samples.

H36M [56] Human 3.6M dataset (H36M) contains 3.6 million human poses. Following the previous works [57]–[59], we train our model with five subjects (1, 5, 6, 7, 8) and test with two subjects (9, 11). However, images on H36M are not much occluded since they are recorded on single-person action in the indoor environment. Therefore, to simulate the occlusion scenario, we introduce an additional test set, which we call *H36_masked*, by synthesizing occlusion with a random mask patch following [60], [61]. In this test set, synthetic masks are randomly colored 30×30 pixel-sized square centered on the joint. We generate the patches for each joint following binomial distribution $B(n = 17, p = 0.02)$.

B. Implementation Details

All experiments are performed using PyTorch [62] on NVIDIA TITAN RTX. Our model is trained with ADAM [63] optimizer with batch size 128 and an initial learning rate (LR) 5e-4. Cosine LR decay with warm-up step 1,000 is used. In PORT, we employ four encoder layers (L) with four heads and set hidden dimensions (D) for embeddings as 64 unless otherwise stated. The max epoch is set to 100. For experiments, we train PORT with MJM using joint locations as input and test its refinement ability after adding the PORT to different keypoint detectors. We use

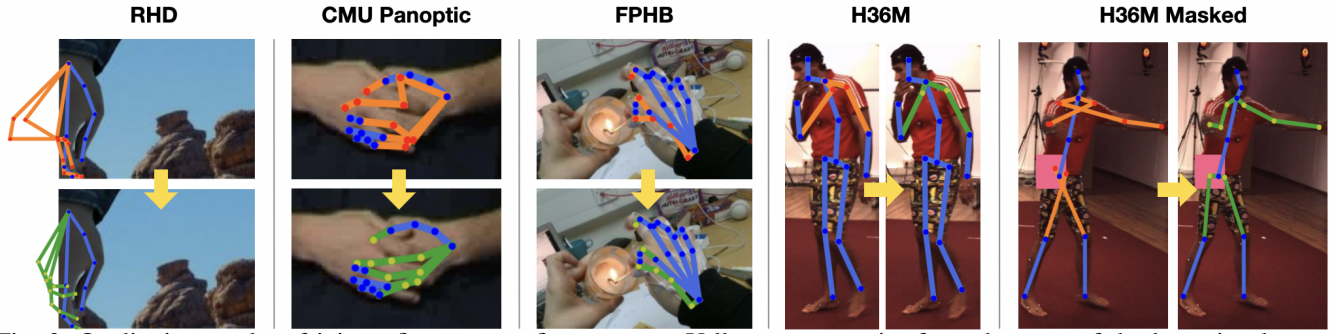

Fig. 3: Qualitative results of joint refinement on five test sets. Yellow arrows point from the pose of the keypoint detector (HRNet_w48) to the refined pose of PORT. **Blue** points indicate visible joints, which have a high confidence value, whereas **Orange** indicate occluded joints, which have a low confidence value. **Green** points represent refined joints from PORT.

pretrained keypoint detectors provided by MMPose [64][1]. We set the masking ratio for MJM to be 40% and $\mathcal{K} = \{1, 2\}$ for all our experiments unless otherwise stated. The confidence threshold value δ is empirically selected based on the keypoint detector outputs on the training sets.

C. Experimental Results

We investigate the effect of PORT on the various keypoint detectors (HRNet [43], HRNetv2 [46], MobileNetv2 [54], ResNet [55]) on five test sets. In Table I, we compare the error of estimated joints $\mathbf{J}$ from the pretrained keypoint detectors and refined joints $\hat{\mathbf{J}}$ from PORT. Bold figures indicate the results with PORT, and blue figures denote the improvement. We observe that PORT reduces the errors of all keypoint detectors under different test sets in terms of both EPE and P-EPE. We also find that P-EPE improvements are more significant than EPE over all results. This result implies that PORT tends to refine the results into plausible poses than fix each joint into the exact location.

Qualitative results in Fig. 3 on five test sets further prove the occlusion refinement ability of the proposed method. PORT successfully refines the occluded joints by replacing them with the structurally plausible joint when the hand does not fully appear in the frame (RHD), occluded by the other hand or patches (CMU Panoptic, H36M_masked), and self-occluded (FPHB). We also observe that PORT can refine incorrectly detected visible joints (H36M, H36M_masked), which also have low confidence.

D. Analysis

Occlusion Refinement We analyze PORT's effectiveness on occlusion using experimental results from the HRNet_w48 keypoint detector. In Fig. 4, we plot the error distribution with and without PORT on five test sets to observe the effect of PORT at different confidence values. We use box plots to group joints based on their confidence values, connecting the mean values of each box with blue and orange lines. Blue lines (without PORT) are duplicated on the right plot for easy comparison. We observe that the error distribution with confidence less than δ (vertical red lines), which we assume

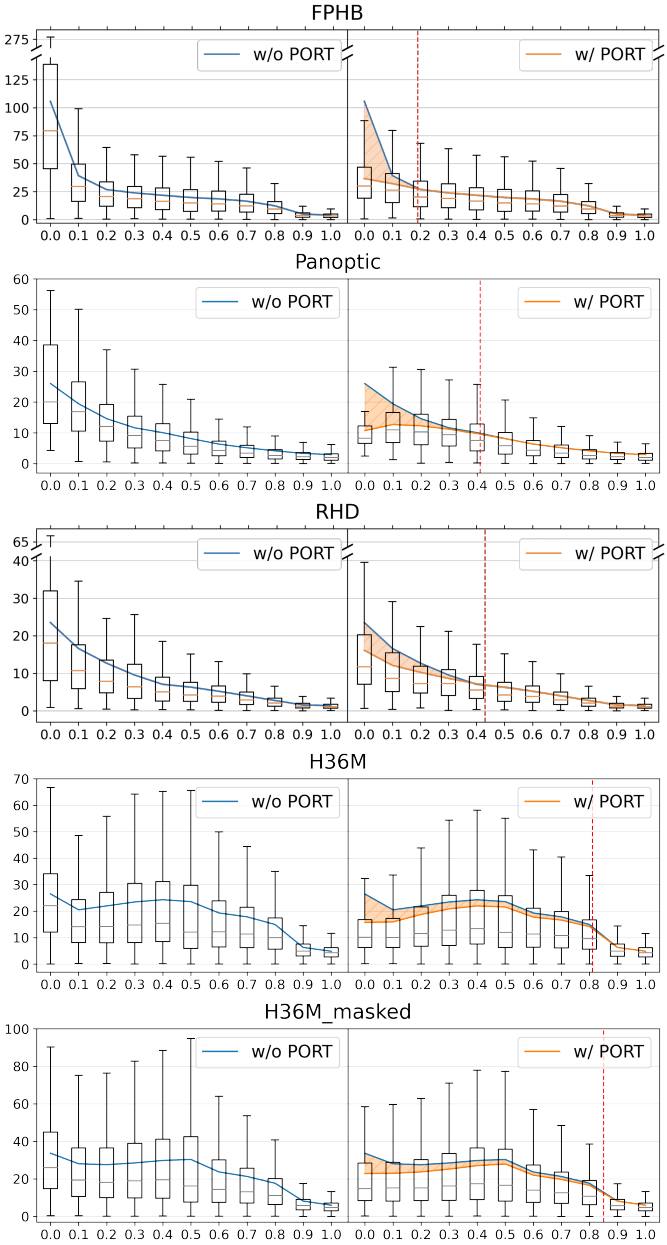

Fig. 4: Error distribution over different confidence values **(Left)** without and **(Right)** with PORT on five test sets. The vertical red lines indicate each test set's confidence threshold δ. The orange area highlights the error reduction by PORT.

[1]Since no pretrained models exist for the FPHB dataset, we train keypoint detectors by ourselves with hyper-parameters provided by original works.

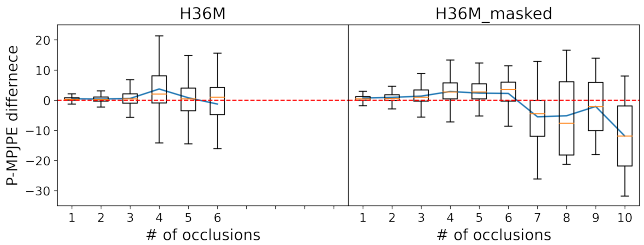

Fig. 5: Distribution of error difference between with and without PORT over the number of occlusion on H36M **(Left)** and H36M_masked **(Right)**. Positive values indicate the reduction of errors.

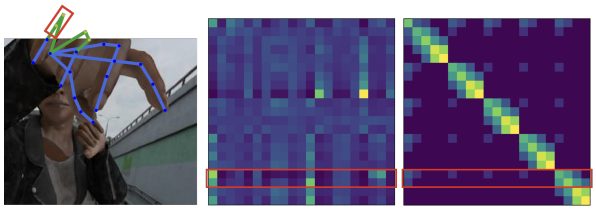

Fig. 6: Example of the occluded skeleton **(Left)** and their corresponding self-attention map from PORT **(Middle)**, and 2-hop adjacency matrix **(Right)**. The occluded joints of the pinky finger and their corresponding weights are marked as red boxes.

as occlusion, is reduced across all test sets. Additionally, we note that the effect of PORT is greater on lower confidence joints. These results demonstrate that PORT successfully reduces error by refining low-confidence joints.

We investigate the relation between PORT's refinement performance and the number of occlusions. In Fig. 5, we visualize the error difference distribution between with and without PORT on the H36M and H36M_masked test sets over the number of occlusions. PORT reduces the error when the number of occlusions is less than 5 for H36M, but exacerbates the error after that. We conjecture that the context for refining occlusion is insufficient under severe occlusion, explaining why PORT shows similar P-EPE improvement on H36M and H36M_masked despite the latter having 30% more occlusions (4.6% vs. 5.3% in Table I).

Skeleton Feature Extraction. Fig. 6 gives an example of an attention map and an adjacency matrix of the occluded skeleton that represent global and local features of the skeleton, respectively. Here we focus on the occlusion case of the pinky finger, marked with red boxes. We first see that the model attends more to its adjacent joints to reconstruct occluded joints from the observation that the attention and adjacency matrix has a similar pattern on occluded joints. Still, the attention map attends to all the other joints. It shows that PORT refines the result by combining local features from adjacent joints and global features.

Parameter Comparison. We compare the number of parameters of existing keypoint detectors with that of PORT in Table II. We note that the number of parameters of PORT is significantly smaller than keypoint detectors. The number of parameters of the PORT is only 0.8% $\sim$ 5.2% of keypoint detectors, which proves PORT can be a light-weighted plug-and-play to keypoint detectors.

Methods		# Parameters (M)
Keypoint Detectors	HRNet_w32 [43]	28.5
	HRNet_w48 [43]	63.6
	HRNetv2_w18 [46]	9.6
	MobileNetv2 [54]	9.6
	ResNet50 [55]	34.0
PORT		**0.5**

TABLE II: Comparison of the number of parameters of keypoint detectors and PORT.

Methods	P-EPE ↓
Linear	7.65
MSGC	
w/ $\mathcal{K} = \{1\}$	7.54
w/ $\mathcal{K} = \{2\}$	7.40
w/ $\mathcal{K} = \{3\}$	7.64
w/ $\mathcal{K} = \{1, 2\}$	**7.36**
w/ $\mathcal{K} = \{1, 3\}$	7.43
w/ $\mathcal{K} = \{2, 3\}$	7.38
w/ $\mathcal{K} = \{1, 2, 3\}$	7.39

(a)

Masking Ratio	P-EPE ↓
10%	7.63
20%	7.57
30%	7.45
40%	**7.36**
50%	7.39
60%	7.37

(b)

TABLE III: P-EPE comparisons on different **(a)** feature projection methods and **(b)** masking ratios for MJM.

E. Ablation Studies

In this section, we examine different configurations of PORT. All ablation studies are conducted on CMU panoptic [52] dataset with MobileNetv2 [54] as a keypoint detector.

Feature Projection. We compare the performance of different projection methods for feature projection in Table III-(a). When compared to the Linear layer, MSGC shows better P-EPE. We also observed that increasing the cardinality of the kernel set does not always lead the performance improvement. Among different kernel sets, $\mathcal{K} = \{1, 2\}$ shows the best performance.

Masking Strategies. We report the performances of different masking strategies in Table III-(b). Among the different masking ratios, the model trained with 40% shows the best performance. Until 40%, increasing the masking ratio reduces P-EPE, but the ratios higher than 40% show slightly worse results.

V. CONCLUSION

This work introduces a POse Relation Transformer (PORT) to refine occluded joints in the intermediate step of human pose estimation, trained with Masked Joint Modeling (MJM) to capture the local and global pose context and reconstruct occluded joints. Our experiments on four human pose datasets demonstrate that PORT is a model-agnostic plug-and-play module that mitigates occlusion for keypoint detectors with minimal additional computational cost.

Acknowledgement This work was partially supported by US National Science Foundation (FW-HTF 1839971). We also acknowledge the Feddersen Chair Funds for Professor Karthik Ramani.

REFERENCES

[1] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1986–1992. 1

[2] J. Liang, A. Handa, K. Van Wyk, V. Makoviychuk, O. Kroemer, and D. Fox, "In-hand object pose tracking via contact feedback and gpu-accelerated robotic simulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6203–6209. 1

[3] H. Liu, Z. Zhang, X. Xie, Y. Zhu, Y. Liu, Y. Wang, and S.-C. Zhu, "High-fidelity grasping in virtual reality using a glove-based system," in *2019 international conference on robotics and automation (icra)*. IEEE, 2019, pp. 5180–5186. 1

[4] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," *arXiv preprint arXiv:2108.05877*, 2021. 1

[5] J. Li, J. Wang, S. Wang, and C. Yang, "Human–robot skill transmission for mobile robot via learning by demonstration," *Neural Computing and Applications*, pp. 1–11, 2021. 1

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016. 1

[8] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall, "Hope-net: A graph-based model for hand-object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6608–6617. 1

[9] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8561–8568. 1

[10] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 143–152. 1

[11] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603. 1

[12] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660. 2

[13] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4733–4742. 2

[14] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2602–2611. 2

[15] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3196–3206. 2

[16] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 807–11 816. 2

[17] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand manipulating an object from rgb-d input," in *European Conference on Computer Vision*. Springer, 2016, pp. 294–310. 2

[18] M. Schroder and H. Ritter, "Hand-object interaction detection with fully convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 18–25. 2

[19] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 2020. 2

[20] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 409–419. 2, 4

[21] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 813–822. 2

[22] S. Kim and H.-g. Chi, "First-person view hand segmentation of multi-modal hand activity video dataset." *BMVC 2020*, 2020. 2

[23] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 49–59. 2

[24] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1154–1163. 2

[25] Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3d human pose estimation in video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 723–732. 2

[26] Q. Ye and T.-K. Kim, "Occlusion-aware hand pose estimation using hierarchical mixture density network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–817. 2

[27] X. Guo and Y. Dai, "Occluded joints recovery in 3d human pose estimation based on distance matrix," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1325–1330. 2

[28] C. Li and G. H. Lee, "Generating multiple hypotheses for 3d human pose estimation with mixture density network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9887–9895. 2

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. 2, 3

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 2

[31] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31. 2, 3

[32] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020. 2

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229. 2

[34] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890. 2

[35] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 244–253. 2

[36] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "Infogcn: Representation learning for human skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 186–20 196. 2

[37] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 939–12 948. 2

[38] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 11 313–11 322. 2

[39] L. Huang, J. Tan, J. Liu, and J. Yuan, "Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 17–33. 2

[40] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 802–11 812. 2

[41] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in

Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11 656–11 665. 2

[42] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 483–499. 2

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703. 2, 4, 5, 6

[44] C. Jiang, K. Huang, S. Zhang, X. Wang, and J. Xiao, "Pay attention selectively and comprehensively: Pyramid gating network for human pose estimation without pre-training," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2364–2371. 2

[45] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 440–10 450. 2

[46] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020. 2, 4, 5, 6

[47] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 529–545. 2

[48] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017. 3

[49] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016. 3

[50] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814. 3

[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. 3

[52] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3334–3342. 4, 6

[53] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4903–4911. 4

[54] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. 4, 5, 6

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 4, 5, 6

[56] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013. 4

[57] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[58] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5255–5264. 4

[59] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018. 4

[60] T. Zhang, B. Huang, and Y. Wang, "Object-occluded human shape and pose estimation from a single color image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7376–7385. 4

[61] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3d human pose estimation to occlusion?" *arXiv preprint arXiv:1808.09316*, 2018. 4

[62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019. 4

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, cite arxiv:1412.6980 Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980 4

[64] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020. 5