

JOTR: 3D Joint Contrastive Learning with Transformers for Occluded Human Mesh Recovery

Jiahao Li^{1,2†}, Zongxin Yang¹, Xiaohan Wang¹, Jianxin Ma², Chang Zhou², Yi Yang^{1‡}
¹ ReLER, CCAI, Zhejiang University ² DAMO Academy, Alibaba Group

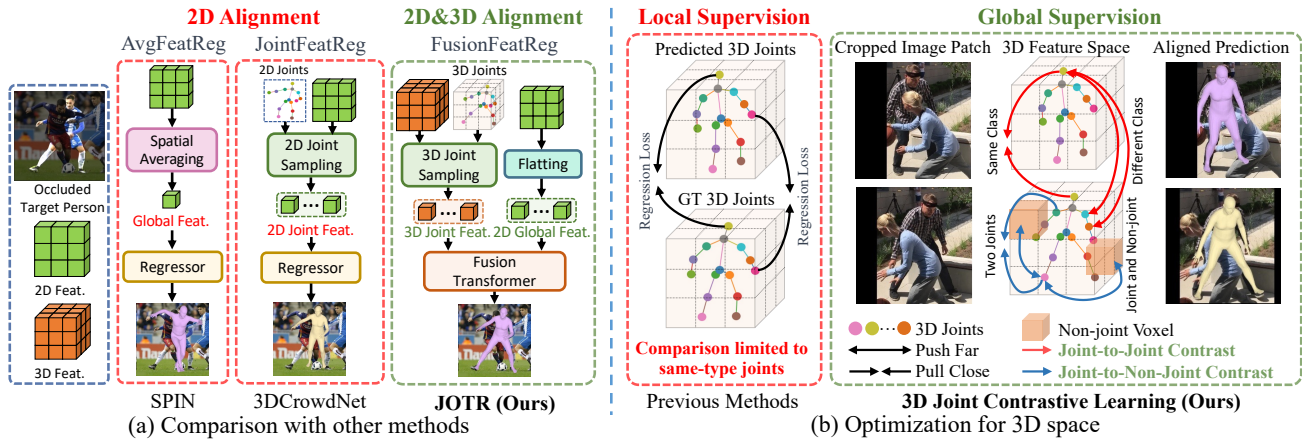


Figure 1: (a) Current methods for human mesh recovery could be classified into two categories: AvgFeatReg [25, 32] and JointFeatReg [10, 56, 72], which focus on improving 2D alignment by using average 2D features (feat.) for regression and employing sampled 2D joint feature for regression, respectively. In contrast, our proposed novel techniques (*i.e.*, FusionFeatReg), fuse 2D and 3D features for regression to enhance both 2D and 3D alignment. (b) Moreover, to provide global supervision for the entire 3D space, we introduce a 3D joint contrastive learning method, which stands in contrast to previous approaches that solely apply 3D joints as local supervision.

Abstract

In this study, we focus on the problem of 3D human mesh recovery from a single image under obscured conditions. Most state-of-the-art methods aim to improve 2D alignment technologies, such as spatial averaging and 2D joint sampling. However, they tend to neglect the crucial aspect of 3D alignment by improving 3D representations. Furthermore, recent methods struggle to separate the target human from occlusion or background in crowded scenes as they optimize the 3D space of target human with 3D joint coordinates as local supervision. To address these issues, a desirable method would involve a framework for fusing 2D and 3D features and a strategy for optimizing the 3D space globally. Therefore, this paper presents 3D JOint contrastive learning with TRansformers (JOTR) framework for handling occluded 3D human mesh recovery. Our method includes an encoder-decoder transformer architecture to fuse 2D and 3D representations for achieving 2D&3D aligned results in a coarse-to-fine manner and a novel 3D joint contrastive learning approach for adding explicitly global supervision for the 3D

feature space. The contrastive learning approach includes two contrastive losses: joint-to-joint contrast for enhancing the similarity of semantically similar voxels (*i.e.*, human joints), and joint-to-non-joint contrast for ensuring discrimination from others (*e.g.*, occlusions and background). Qualitative and quantitative analyses demonstrate that our method outperforms state-of-the-art competitors on both occlusion-specific and standard benchmarks, significantly improving the reconstruction of occluded humans. Code is available at <https://github.com/xljh0520/JOTR>.

1. Introduction

The estimation of 3D human meshes from single RGB images is an active area of research in computer vision with a broad range of applications in robotics, AR/VR, and human behavior analysis. In contrast to estimating the pose of general objects [69], human mesh recovery is more challenging due to the complex and deformable structure of the human body. Nevertheless, enhancing human-centric tasks can be achieved by combining visual features and prior knowledge about human anatomy through constructing multi-knowledge representations [66]. Generally, the

[†] Jiahao Li worked on this at his Alibaba internship.

[‡] Yi Yang is the corresponding author.

human mesh recovery task takes a single image as input and regresses human model parameters such as SMPL [46] as output.

Driven by deep neural networks, this task has achieved rapid progress [10, 21, 25, 28, 31–33, 41, 42, 56, 57, 72, 76]. Recent studies have focused on regressing accurate human meshes despite occlusions. To achieve this, most of them employ 2D prior knowledge (*e.g.*, UV maps [76], part segmentation masks [31] and 2D human key points [28]) to focus the model on visible human body parts for enhancing the 2D alignment of the predicted mesh. Additionally, some methods [10, 57] introduce 3D representations to locate 3D joints and extract 2D features from the corresponding regions of the 2D image.

Even though the above methods have achieved significant progress in occluded human mesh recovery, they still remain constrained to these two aspects: the pursuit of **2D alignment** and **local supervision** for 3D joints. (i) As shown in Fig. 1a, the above methods employing 2D prior knowledge mainly focus on **2D alignment** technologies, including spatial averaging and 2D joint sampling. However, in crowded or occluded scenarios, solely focusing on 2D alignment may acquire ambiguous features for the entire mesh due to the lack of estimation of hidden parts. Accordingly, the invisible human body parts would be aligned based on prior knowledge of the standard SMPL template, resulting in misalignment with visible parts and leading to inaccurate 3D reconstructions. (ii) Furthermore, creating a comprehensive and precise 3D representation from a single RGB image is an ill-posed problem as the inherently limited information. As illustrated in Fig. 1b, some methods that use 3D representations rely on localized 3D joints as **local supervision**, ignoring the rich semantic relations between voxels across different scenes. These “local” contents (*i.e.*, human joints) occupy only a small portion of the 3D space, while most voxels are often occupied by occlusions and background. Consequently, the lack of explicit supervision for the entire 3D space makes it difficult to differentiate target humans from other semantically similar voxels, resulting in ambiguous 3D representations.

Therefore, to improve occluded human mesh recovery, we consider investigating a fusion framework that integrates 2D and 3D features for **2D&3D alignment**, along with a **global supervision** strategy to obtain a semantically clear 3D feature space. By leveraging the complementary information from both 2D and 3D representations, the network could overcome the limitations of using only a single 2D representation, enabling obscured human parts to be detected in 3D representations and achieving 2D&3D alignment. Given a global supervision strategy, we could explicitly supervise the entire 3D space to highlight the representation of target humans and distinguish them from other semantically similar voxels, resulting in a semantically clear 3D feature

space.

Based on the above motivation, this paper proposes a novel framework, 3D JOint contrastive learning with TRansformers (JOTR), for recovering occluded human mesh using a fusion of multiple representations as shown in Fig. 1a. Unlike existing methods such as 3DCrowdNet [10] and BEV [57] that employ 3D-aware 2D sampling techniques, JOTR integrates 2D and 3D features through transformers [60] with attention mechanisms. Specifically, JOTR utilizes an encoder-decoder transformer architecture to combine 3D local features (*i.e.*, sampled 3D joint features) and 2D global features (*i.e.*, flatten 2D features), enhancing both 2D and 3D alignment. Besides, to obtain semantically clear 3D representations, the main objective is to strengthen and highlight the human representation while minimizing the impact of irrelevant features (*e.g.*, occlusions and background). Accordingly, we propose a new approach, 3D joint contrastive learning (in Fig. 1b), that provides global and explicit supervision for 3D space to improve the similarity of semantically similar voxels (*i.e.*, human joints), while maintaining discrimination from other voxels (*e.g.*, occlusions). By carefully designing 3D joint contrast for 3D representations, JOTR can mitigate the effects of occlusion and acquire semantically meaningful 3D representations, resulting in accurate localization of 3D human joints and acquisition of meaningful 3D joint features.

We conduct extensive experiments on both standard 3DPW benchmark [61] and occlusion benchmarks such as 3DPW-PC [56, 61], 3DPW-OC [61, 76], 3DPW-Crowd [10, 61], 3DOH [76] and CMU Panoptic [22], and JOTR achieves state-of-the-art performance on these datasets. Especially, JOTR outperforms the prior state-of-the-art method 3DCrowdNet [10] by **6.1** (PA-MPJPE), **4.9** (PA-MPJPE), and **5.3** (MPJPE) on 3DPW-PC, 3DPW-OC, and 3DPW respectively. Moreover, we carry out comprehensive ablation experiments to demonstrate the effectiveness of our framework and 3D joint contrastive learning strategy. Our contributions are summarized as follows:

- We propose JOTR, a novel method for recovering occluded human mesh using a fusion of 2D global and 3D local features, which overcomes limitations caused by person-person and person-object occlusions and achieves 2D&3D aligned results. JOTR achieves state-of-the-art results on both standard and occluded datasets, including 3DPW, 3DPW-PC, 3DPW-OC, 3DOH, CMU Panoptic, and 3DPW-Crowd.
- We develop a 3D joint contrastive learning strategy that supervises the 3D space explicitly and globally to obtain semantically clear 3D representations, minimizing the impact of occlusions and adapting to more challenging scenarios with the help of cross-image contrast.

2. Related Work

Based on the incorporation of a human body model [46, 53], Deep Neural Network-based 3D Human Mesh Recovery methods [9, 10, 12, 15, 21, 25, 28–33, 36, 41, 42, 50, 55–57, 59, 72, 76] can be divided into two categories. The first, SMPL-based approaches [21, 25, 30–32, 55–57, 72], maps input pixels to SMPL parameters [46] such as pose and shape and reconstructs meshes by SMPL models, while the second, SMPL-free methods [33, 41, 42], directly maps raw pixels to 3D mesh vertices without the assistance of SMPL models. In this paper, we mainly consider the first method as the implementation approach.

Human Mesh Recovery. Usually, human mesh recovery methods estimate 3D human mesh of a single person within a person bounding box, which is scaled to the same size. This allows us to assume that the distance between each individual and the camera is roughly equivalent in the cropped image patch. Early works [25, 32] employ spatial averaging on CNN features for obtaining global features and utilize Multi-Layer Perceptrons (MLPs) to regress SMPL parameters. However, global pooling is not suitable for achieving pixel-aligned results, leading to subpar performance in real-world scenarios. PARE [31] proposes using part segmentation masks to enhance pixel alignments. Zhang *et al.* [76] make use of occlusion segmentation masks to allow the model to attend to the visible human body parts, which also helps to reconstruct complete human mesh. OCHMR [28] employs load-global center maps to make the model regress the mesh of the referred person. While these methods make progress in occluded human mesh recovery by enhancing the ability to represent 2D information, they overlook the 3D structural information. 3DCrowdNet [10] and BEV [57] introduce 3D representations to locate human joints in 3D space. However, these approaches also have limitations since they extract 2D CNN features in the corresponding region of located 3D joints, thereby overlooking the full potential of 3D representations. Therefore, we design a fusion framework to integrate 2D and 3D features for mutual complementation.

Multi-Modality Transformers. Following the success of vision transformers in processing image [3, 6, 11, 45] or video [2, 67, 68], multi-modality transformers [7, 24, 34, 37–39, 73–75] are capable of processing input data from multiple modalities, such as text, image, audio, or video, in a single model. The attention mechanism is a key component of transformers, which enables them to selectively focus on relevant parts of the input sequence when generating the output. Returning to the present task, it is worth noting that although there is only one visual modality, two distinct representations (*i.e.*, 2D and 3D representations) are available. Thus, we propose using transformers with attention mechanisms to integrate multi-representation features, rather than relying on global average pooling or joint feature sampling

in CNN features.

Contrastive Learning. Contrastive learning [4, 5, 14, 17] is a type of unsupervised learning that aims to learn a similarity metric between data samples. The goal of contrastive learning is to bring similar examples closer together in feature space while pushing dissimilar examples farther apart. With regards to the current objective, in 3D space, human body joints occupy a relatively small proportion, with the majority of voxels in the space being occupied by other objects (*e.g.*, another person’s body, background elements, and empty elements). This poses a significant challenge in learning a similarity metric between data samples in 3D space. To address this issue, we propose a novel joint-based contrastive learning strategy inspired by the recent success of pixel contrastive learning in semantic segmentation [1, 19, 62, 77], which enables the network to learn a clear similarity metric in 3D space.

3. Method

Human Body Model. SMPL [46] represents a 3D human mesh by 3 low-dimensional vectors (*i.e.*, pose $\theta \in \mathbb{R}^{72}$, shape $\beta \in \mathbb{R}^{10}$ and camera parameters $\pi \in \mathbb{R}^3$). Following previous methods [10, 25, 31, 32, 56], we use the gender-neutral shape model. The SMPL model generates a 3D mesh $\mathcal{M}(\theta, \beta) \in \mathbb{R}^{6890 \times 3}$ through a differentiable function. By applying a pretrained linear regressor $W \in \mathbb{R}^{N \times 6890}$, we obtain the 3D joint coordinates $J_{3D} = W\mathcal{M} \in \mathbb{R}^{N \times 3}$, where $N = 17$, conveniently. Additionally, we obtain the 2D joints $J_{2D} = \Pi(J_{3D}, \pi) \in \mathbb{R}^{N \times 2}$ by projection.

Overview. We propose a method called JOTR, which utilizes transformers to fuse 2D and 3D features for 2D&3D alignment and a novel contrastive learning strategy to globally supervise the 3D space for target humans. Our pipeline is depicted in Fig. 2a and explained in Sec. 3.1, where JOTR regresses SMPL parameters by fusing 2D and 3D features obtained from a cropped image patch. The proposed 3D joint contrastive learning is illustrated in Fig. 3 and explained in Sec. 3.2, including two contrastive losses: joint-to-non-joint contrast and joint-to-joint contrast.

3.1. Fusing 2D and 3D Features with Transformers

As analyzed in Sec. 1, relying solely on 2D features for achieving 2D alignment to reconstruct the human mesh in occluded scenarios may result in suboptimal performance. To overcome this limitation, we propose integrating both 2D and 3D features with transformers in the reconstruction process. Drawing inspiration from the success of transformer models in multi-modality fusion [24, 34], we propose an encoder-decoder transformer architecture that enables the mutual complementation of 2D and 3D features for 2D&3D alignment.

Lifting Module. Unlike previous method [10] that lifts 2D features to 3D features via MLPs without integrating

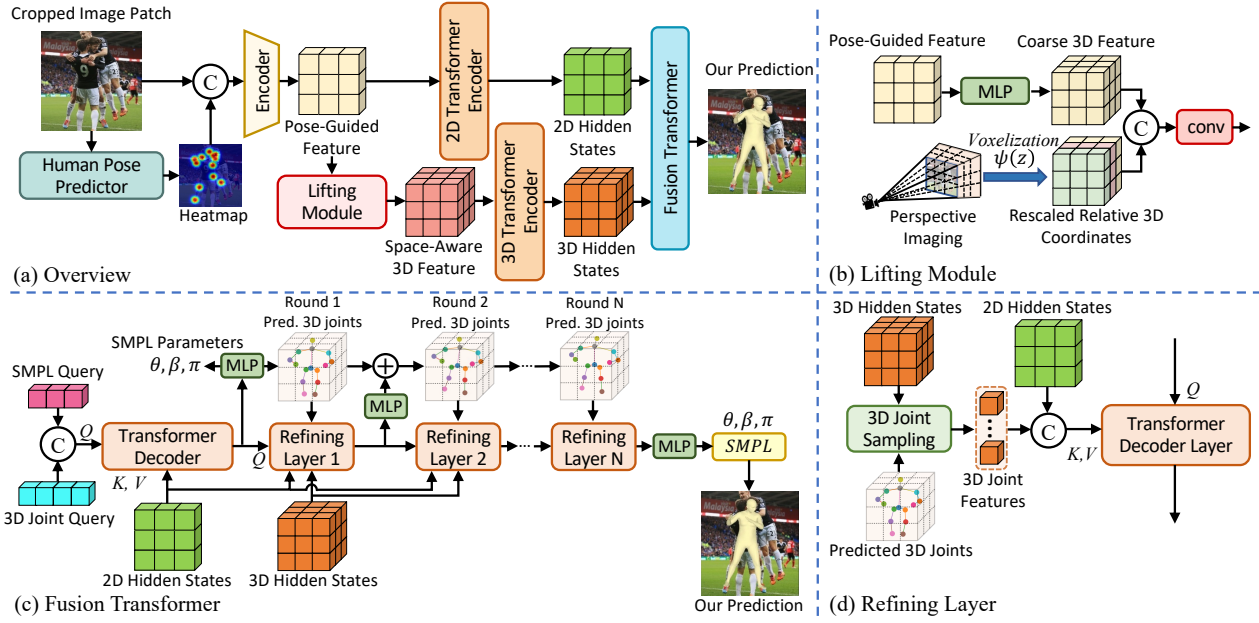


Figure 2: (a) The overview of our method. JOTR achieves 2D and 3D features from a cropped image patch and fuses them with a fusion transformer for 2D and 3D alignment. (b) The detail of the lifting module, which is responsible for lifting pose-guided 2D features to space-aware 3D features. (c) The fusion transformer is applied for fusing 2D and 3D features with attention mechanisms. (d) The refining layer combine sampled 3D joint features and 2D global features to refine the regression.

inductive bias or prior knowledge, we draw inspiration from bird’s eye view representations [8, 40, 54, 57, 63, 65] in 3D space. As analyzed in BEV [57], the farther a voxel is from the camera, the less information it carries. To put this hypothesis into practice, BEV employs pre-defined 3D camera anchor maps to impact the 3D feature. Similar to BEV, we design learnable Rescaled Relative 3D Coordinates (RRC) $C_{3D} \in \mathbb{R}^{D \times H \times W \times 3}$ in range (0, 1) to provide 3D spatial prior knowledge. In this representation, $C_{ijk} \in \mathbb{R}^3$ represents the relative location of voxel (x_k, y_j, z_i) and the x and y coordinates are uniformly distributed with equal intervals. For Z axis, we utilize a monotonically increasing convex function ψ to rescale z coordinates unevenly as $z' = \psi(z)$. In practice, we employ $\psi(z) = z^\lambda$, $\lambda > 1$ as rescaling function and λ is a learnable parameter with initial value of 3.0. The whole pipeline can be written as:

$$\begin{aligned} \hat{F}_{3D} &= MLP(F_{2D}), \\ \tilde{F}_{3D} &= CNN(Concat(\hat{F}_{3D}, C_{3D})), \\ H_{3D} &= TransformerEncoder(\tilde{F}_{3D}). \end{aligned}$$

JOTR first lifts pose-guided 2D feature $F_{2D} \in \mathbb{R}^{H \times W \times C}$ which is obtained from image and joint heatmap through CNN encoder to coarse 3D feature $\hat{F}_{3D} \in \mathbb{R}^{D \times H \times W \times C}$ via MLPs without any inductive bias or prior knowledge. Then, JOTR concatenates \hat{F}_{3D} and C_{3D} in channel dimension. Following CoordConv [44], we apply a convolutional block to refine the concatenated feature to achieve space-aware 3D feature $\tilde{F}_{3D} \in \mathbb{R}^{D \times H \times W \times C}$. Finally, we utilize a transformer encoder (*i.e.*, 3D transformer encoder in Fig. 2a) to enhance the global interaction of 3D space via self-attention mechanism,

$$Attention(Q, K, V) = softmax\left(\frac{QK}{\sqrt{C}}\right)V, \quad (1)$$

achieving the hidden state $H_{3D} \in \mathbb{R}^{D \times H \times W \times C}$. For the sake of simplicity, we omit the positional encoding and rearrangement of tensor in Eq. (1).

Fusion Transformer. In contrast to prior 2D alignment technologies such as spatial averaging and 2D joint feature sampling, we propose the use of attention mechanisms to selectively focus on semantically distinct areas (*i.e.*, visible human parts). Moreover, to estimate hidden information for achieving 3D alignment, we extend 2D features with 3D joint feature sampling. Drawing inspiration from the successful fusion of image and text representations in MDETR [24] and Moment-DETR [34], we design a transformer decoder-based fusion transformer to integrate 2D and 3D features and regress SMPL parameters in a coarse-to-fine manner leading to 2D & 3D alignment.

SMPL/Joint Query. Instead of concatenating or pooling on 2D and 3D features, JOTR decouples the SMPL parameters and 2D/3D joint features into separate query tokens, $Query \in \mathbb{R}^{N_q \times C}$, comprising two distinct parts. The N_s tokens belong to SMPL token, where $N_s = 3$, and are responsible for regressing pose, shape and camera parameters $\{\theta, \beta, \pi\}$ respectively. The remaining $N_j = N_q - N_s$ tokens are responsible for locating 3D joints of the human and extracting corresponding 3D joint features, which refine the SMPL parameters and provide auxiliary supervision for the 3D space.

2D-Based Initial Regression. As shown in Fig. 2c, we initially have no prior knowledge about the 3D joint loca-

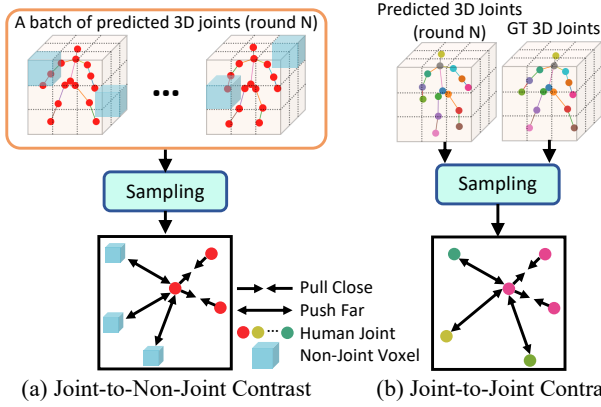


Figure 3: (a) The detail of joint-to-non-joint contrastive learning. (b) The detail of joint-to-joint contrastive learning.

tions. We regress the SMPL parameters and initial 3D joints with a transformer decoder reasoning in 2D hidden state $H_{2D} \in \mathbb{R}^{H \times W \times C}$ which is obtained by a transformer encoder (*i.e.*, 2D transformer encoder in Fig. 2a) working on F_{2D} . H_{2D} is set as K and V , and $Query$ tokens are treated as Q in Eq. (1). Subsequently, we obtain initial predictions for pose, shape, camera parameters, and 3D joint coordinates via MLPs working on the output of transformer decoder.

Refining with 3D Features. To conserve computing resources, we avoid directly concatenating the hidden states (*i.e.*, H_{2D} and H_{3D}). Instead, we use the initial prediction of 3D joints $J'_{3D} \in \mathbb{R}^{N_j \times 3}$ as reference points to sample “local” 3D joint features $H_{J_{3D}} = \mathcal{F}(H_{3D}, J'_{3D}) \in \mathbb{R}^{N_j \times C}$ like [78] in Fig. 2d, where $\mathcal{F}(\cdot)$ denotes feature sampling and trilinear interpolation. We then concatenate H_{2D} with $H_{J_{3D}}$ and feed them into another transformer decoder (*i.e.*, a stack of refining layers in Fig. 2c) as K and V in Eq. (1). Note that Z axis is not uniform in our 3D space. When sampling “local” 3D joint features, we also apply ψ to rescale z in J'_{3D} as mentioned earlier. Since the refining process consists of several identical transformer decoder layers, we naturally consider utilizing the outputs of each layer $H_d \in \mathbb{R}^{L \times N \times C}$ as a cascade refinement,

$$\beta^{l+1} = \beta^l + MLP(\beta^l, H_d^l), \quad (2)$$

where l denotes the l -th refining layer and $MLP(\beta^l, H_d^l)$ is responsible for learning the residual for correcting parameters via MLPs. Besides, we also regress and J'_{3D} and the input of Vposer [53] with cascaded refinement as shown above.

3.2. 3D Joint Contrastive Learning

As analyzed in Sec. 1, due to the lack of explicit “global” supervision for 3D representations, the “local” 3D joint coordinates may not provide accurate enough supervision for the 3D features. Especially when the target person is obstructed by other individuals, similarities in their semantic appearances could result in confusion. To address this challenge, we propose a 3D joint contrastive learning strategy inspired

by the success of pixel contrastive learning in semantic segmentation [62]. This approach enhances the representation of the target person while distinguishing them from other objects (*e.g.*, other people, occlusions, and background).

Vanilla Contrastive Learning. In computer vision, contrastive learning was originally applied for unsupervised representation learning, where the goal is to minimize the distance between similar images (*i.e.*, an image with its augmented version) while maximizing the distance between dissimilar images (*i.e.*, an image with another image in training set) in an embedding space. Usually, InfoNCE [16, 51] is used as the loss function for contrastive learning,

$$\mathcal{L}_I^{\text{NCE}} = -\log \frac{\exp(i \cdot i^+ / \tau)}{\exp(i \cdot i^+ / \tau) + \sum_{i^- \in \mathcal{N}_I} \exp(i \cdot i^- / \tau)}, \quad (3)$$

where I is the anchor image, $i \in \mathbb{R}^C$ is the representation embedding of I , i^+ is an embedding of a positive for I , \mathcal{N}_I contains embeddings of negatives, ‘ \cdot ’ denotes the inner (dot) product, and $\tau > 0$ is a temperature hyper-parameter. Note that all the embeddings in the loss function are ℓ_2 -normalized.

Joint-to-Non-Joint Contrast. As shown in Fig. 3a, to better distinguish occlusion cases, we consider *joint-to-non-joint contrast* between the n -th round predicted joints (in Fig. 2c) and the entire 3D space, as there are many voxels outside the joints. We augment Eq. (3) in our joint-to-non-joint contrast setting. Since we employ trilinear interpolation to acquire the joint embedding from H_{3D} , the joint embedding is a weighted sum of the 8 voxel embeddings in the 3D space. As a result, for an anchor joint j , the positive samples are other predicted joints (not restricted to belonging to the same class), and the negative samples are the voxels that have no contribution to any joint embeddings. The joint-to-non-joint contrastive loss is defined as:

$$\mathcal{L}_{j_{2n}}^{\text{NCE}} = \frac{1}{|\mathcal{P}_j|} \sum_{j^+ \in \mathcal{P}_j} -\log \frac{\exp(j \cdot j^+ / \tau)}{\exp(j \cdot j^+ / \tau) + \sum_{n^- \in \mathcal{N}_n} \exp(j \cdot n^- / \tau)}, \quad (4)$$

where \mathcal{P}_j is joint embedding collections of positive samples and \mathcal{N}_n denote non-joint voxel embedding collections of negative samples, for joint j .

Joint-to-Joint Contrast. As shown in Fig. 3b, to strengthen the internal connections among joints of the same category, we consider *joint-to-joint contrast* among human joints. We extend Eq. (3) for applying to our joint-to-joint contrast setting. Essentially, the data samples in our contrastive loss computation are the n -th round predicted joints (in Fig. 2c) and ground truth 3D joints. For an anchor joint j from predicted joints with its corresponding semantic label \bar{c} (*e.g.*, head, right hand, and neck), the positive samples are ground truth joints that also belong to the class \bar{c} , and the negative samples are the n -th round predicted joints belonging to the other classes $\mathcal{C} \setminus \{c_j\}$. As a result, the joint-to-joint

| Method | 3DPW-OC | | | 3DOH | | | 3DPW-PC | | | 3DPW-Crowd | | |
|------------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|--------------|-------------|-------------|--------------|
| | MPJPE↓ | PA-MPJPE↓ | PVE↓ | MPJPE↓ | PA-MPJPE↓ | PVE↓ | MPJPE↓ | PA-MPJPE↓ | PVE↓ | MPJPE↓ | PA-MPJPE↓ | PVE↓ |
| I2L-MeshNet [50] | 92.0 | 61.4 | 129.5 | - | - | - | 117.3 | 80.0 | 160.2 | 115.7 | 73.5 | 162.0 |
| SPIN [32] | 95.5 | 60.7 | 121.4 | 110.5 | 71.6 | 124.2 | 122.1 | 77.5 | 159.8 | 121.2 | 69.9 | 144.1 |
| PyMAF [72] | 89.6 | 59.1 | 113.7 | 101.6 | 67.7 | 116.6 | 117.5 | 74.5 | 154.6 | 115.7 | 66.4 | 147.5 |
| ROMP [56] | 91.0 | 62.0 | - | - | - | - | 98.7 | 69.0 | - | 104.8 | 63.9 | 127.8 |
| OCHMR [28] | 112.2 | 75.2 | 145.9 | - | - | - | - | - | - | - | - | - |
| PARE* [31] | 83.5 | 57.0 | 101.5 | 109.0 | 63.8 | 117.4 | 96.8 | 64.5 | 122.4 | 94.9 | 57.5 | 117.6 |
| 3DCrowdNet [10] | 83.5 | 57.1 | 101.5 | 102.8 | 61.6 | 111.8 | 90.9 | 64.4 | 114.8 | 85.8 | 55.8 | 108.5 |
| Ours | 75.7 | 52.2 | 92.6 | 98.7 | 59.3 | 104.8 | 86.5 | 58.3 | 109.7 | 82.4 | 52.0 | 103.4 |

Table 1: Comparisons to the state-of-the-art methods under severe occlusion. The units for mean joint and vertex errors are in mm. PARE* use a HRNet-32 backbone, others are with ResNet-50.

contrastive loss is defined as:

$$\mathcal{L}_{j2j}^{\text{NCE}} = \frac{1}{|\mathcal{P}_j|} \sum_{j^+ \in \mathcal{P}_j} -\log \frac{\exp(j \cdot j^+ / \tau)}{\exp(j \cdot j^+ / \tau) + \sum_{j^- \in \mathcal{N}_j} \exp(j \cdot j^- / \tau)}, \quad (5)$$

where \mathcal{P}_j and \mathcal{N}_j denote joint embedding collections of the positive and negative samples, respectively, for joint j .

Note that the positive and negative samples, as well as the anchor joint j in both *joint-to-non-joint* and *joint-to-joint* contrast are not necessarily limited to the same 3D space. The joint-to-non-joint contrastive loss in Eq. (4) and joint-to-joint contrastive loss in Eq. (5) are complementary to each other; the former enables the network to learn discriminative joint features that are distinctly different from those of other non-joint voxels (*e.g.*, occlusions), while the latter helps to regularize the joint embedding space by improving intra-class compactness and inter-class separability.

3.3. Loss Function.

Finally, we obtain refined SMPL parameters $\{\theta, \beta, \pi\}$. We can achieve mesh vertices $M = \mathcal{M}(\theta, \beta) \in \mathbb{R}^{6890 \times 3}$ and 3D joints from mesh $J_{3D} = W\mathcal{M} \in \mathbb{R}^{N \times 3}$ accordingly. We follow common practices [10, 25, 32] to project 3D joints on 2D space $J_{2D} = \Pi(J_{3D}, \pi) \in \mathbb{R}^{N \times 2}$ and add supervisions with 2D keypoints. Meanwhile, when 3D annotations are available, we also add 3D supervision on SMPL parameters and 3D joint coordinates. Overall, the loss function can be written as follows:

$$\mathcal{L} = \lambda_{3D} \mathcal{L}_{3D} + \lambda_{2D} \mathcal{L}_{2D} + \lambda_{SMPL} \mathcal{L}_{SMPL} + \lambda_{j2n} \sum_j \mathcal{L}_{j2n}^{\text{NCE}} + \lambda_{j2j} \sum_j \mathcal{L}_{j2j}^{\text{NCE}}, \quad (6)$$

where j is the sampled anchor joints and the first three is calculated as:

$$\begin{aligned} \mathcal{L}_{3D} &= \|J_{3D} - \hat{J}_{3D}\|, \\ \mathcal{L}_{2D} &= \|J_{2D} - \hat{J}_{2D}\|, \\ \mathcal{L}_{SMPL} &= \|\theta - \hat{\theta}\| + \|\beta - \hat{\beta}\|, \end{aligned}$$

where $\|\cdot\|$ denotes L1 norm. \hat{J}_{2D} , \hat{J}_{3D} , $\hat{\theta}$, and $\hat{\beta}$ denote the ground truth 2D keypoints, 3D joints, pose parameters and shape parameters, respectively.

4. Experiments

Implementation Detail. This proposed JOTR is validated on the ResNet-50 [18] backbone. Following 3DCrowdNet [10], we initialize ResNet from Xiao *et al.* [64] for fast convergence. We use AdamW optimizer [47] with a batch size of 256 and weight decay of 10^{-4} . The initial learning rate is 10^{-4} . The ResNet-50 backbone takes a 256×256 image as input and produces image features with size of $2048 \times 8 \times 8$. We build the 3D features with size of $256 \times 8 \times 8 \times 8$ and 2D features with size of $256 \times 8 \times 8$. As for weights for multiple different losses, we follow [27] to adjust them dynamically using learnable parameters. For joint-to-non-joint contrast, we sample 100 anchor joints per GPU in each mini-batch, which are paired with 1024 positive and 2048 and negative samples. For joint-to-joint contrast, we sample 100 anchor joints per GPU in each mini-batch, which are paired with 128 positive and 256 and negative samples. Both contrastive losses are set to a temperature of 0.07. More details can be found in the supplementary material.

Training. Following the settings of previous work [10, 25, 32], our approach is trained on a mixture of data from several datasets with 3D and 2D annotations, including Human3.6M [20], MuCo-3DHP [48], MSCOCO [43], and CrowdPose [35]. Only the training sets are used, following the standard split protocols. For the 2D datasets, we also utilize their pseudo ground-truth SMPL parameters [49] for training.

Evaluation. The 3DPW [61] test split, 3DOH [76] test split, 3DPW-PC [56, 61], 3DPW-OC [61, 76], 3DPW-Crowd [10, 61] and CMU-Panoptic [22] datasets are used for evaluation. 3DPW-PC and 3DPW-Crowd are the *person-object* occlusion subset of 3DPW, 3DPW-OC is the *person-object* occlusion subset of 3DPW and 3DOH is another *person-object* occlusion specific dataset. We adopt per-vertex error (PVE) in mm to evaluate the 3D mesh error. We employ Procrustes-aligned mean per joint position error (PA-MPJPE) in mm and mean per joint position error (MPJPE) in mm to evaluate the 3D pose accuracy. As for CMU-Panoptic, we only report mean per joint position error (MPJPE) in mm following previous work [10, 21, 56].

| Method | MPJPE↓ | PA-MPJPE↓ | PVE↓ |
|-----------------------------|-------------|-------------|-------------|
| HMR [25] | 130.0 | 76.7 | - |
| Kanazawa <i>et al.</i> [26] | 116.5 | 72.6 | 139.3 |
| GCMR [33] | - | 70.2 | - |
| DSD-SATN [58] | - | 69.5 | - |
| SPIN [32] | 96.9 | 59.2 | 116.4 |
| I2L-MeshNet [50] | 93.2 | 58.6 | 136.5 |
| PyMAF [72] | 92.8 | 58.9 | 110.1 |
| OCHMR [28] | 89.7 | 58.3 | 107.1 |
| EFT [23] | - | 54.2 | - |
| ROMP [56] | 89.3 | 53.5 | 105.6 |
| PARE [31] | 82.9 | 52.3 | 99.7 |
| 3DCrowdNet [10] | 81.7 | 51.2 | 98.3 |
| Ours | 76.4 | 48.7 | 92.6 |

Table 2: Comparisons to the state-of-the-art methods on standard 3DPW [61] test split.

| 2D Feature | 3D Feature | MPJPE ↓ | PA-MPJPE↓ | PVE↓ |
|------------|------------|-------------|-------------|-------------|
| sampling | none | 78.3 | 54.2 | 94.7 |
| flattening | none | 77.8 | 53.2 | 94.7 |
| sampling | sampling | 77.4 | 53.8 | 94.6 |
| flattening | sampling | 77.0 | 53.6 | 94.3 |

Table 4: Ablation study of utilization of 2D and 3D features. Flattening: flattening in height and weight for tokenization. Sampling: joint feature sampling for tokenization.

| w/o SMPL Token | MPJPE ↓ | PA-MPJPE↓ | PVE↓ |
|----------------|-------------|-------------|-------------|
| ✗ | 79.8 | 54.0 | 95.9 |
| ✓ | 77.0 | 53.6 | 94.3 |

Table 6: Ablation study of decoupling SMPL query. We apply average pooling on outputs of joint query tokens for regressing SMPL parameters when without SMPL query.

4.1. Comparison to the State-of-the-Art on Occlusion Benchmark

3DPW-OC [61, 76] is a person-object occlusion subset of 3DPW and contains 20243 persons. Tab. 1 shows our method achieve a new state-of-the-art performance on 3DPW-OC.

3DOH [76] is a person-object occlusion-specific dataset and contains 1290 persons in testing set, which incorporates a greater extent of occlusions than 3DPW-OC. For a fair comparison, we initialize PARE with weights that are not trained on the 3DOH training set, resulting in different performances from the results reported in [31]. Tab. 1 shows our method surpasses all the competitors with 59.3 (PA-MPJPE).

3DPW-PC [56, 61] is a multi-person subset of 3DPW and contains 2218 persons’ annotations under person-person occlusion. Tab. 1 shows our method surpasses all the competitors with 58.8 (PA-MPJPE).

3DPW-Crowd [10, 61] is a person crowded subset of 3DPW and contains 1923 persons. We slightly surpass previous state-of-the-art as shown in Tab. 1.

CMU-Panoptic [22] is a dataset with multi-person indoor scenes. We follow previous methods [10, 21] applying 4

| Method | Haggl. | Mafia | Ultim. | Pizza | Mean |
|---------------------------|-------------|--------------|--------------|--------------|--------------|
| Zanfir <i>et al.</i> [70] | 140.0 | 156.9 | 150.7 | 156.0 | 153.4 |
| Zanfir <i>et al.</i> [71] | 141.4 | 152.3 | 145.0 | 162.5 | 150.3 |
| Jiang <i>et al.</i> [21] | 129.6 | 133.5 | 153.0 | 156.7 | 143.2 |
| ROMP [56] | 111.8 | 129.0 | 148.5 | 149.1 | 134.6 |
| SPIN [32] | 124.3 | 132.4 | 150.4 | 153.5 | 133.1 |
| OCHMR [28] | 115.5 | 123.7 | 142.6 | 150.6 | 133.1 |
| REMIPS [13] | 121.6 | 137.1 | 146.4 | 148.0 | 138.3 |
| 3DCrowdNet [10] | 109.6 | 135.9 | 129.8 | 135.6 | 127.6 |
| BEV* [57] | 110.3 | 125.6 | 150.7 | 131.7 | 127.9 |
| Ours | 99.9 | 113.5 | 115.7 | 123.6 | 114.7 |

Table 3: Comparison on CMU-Panoptic [22]. The numbers denote MPJPE. For a fair comparison, we apply BEV* model that is not fine-tuned on AGORA [52] (*i.e.*, a synthetic 3D dataset.).

| Index of Refining Layer | MPJPE ↓ | PA-MPJPE↓ | PVE↓ |
|-------------------------|---------|-----------|-------|
| 0 | 276.5 | 124.7 | 308.8 |
| 1 | 145.5 | 103.2 | 185.9 |
| 2 | 109.8 | 68.7 | 131.6 |
| 3 | 77.0 | 53.6 | 94.3 |

Table 5: Validation of coarse-to-fine regression. We take intermediate outputs of refining layers for regressing SMPL parameters. Zero stands for 2D-based initial regression.

| J2N | J2J | MPJPE ↓ | PA-MPJPE↓ | PVE↓ |
|-----|-----|-------------|-------------|-------------|
| ✗ | ✗ | 77.0 | 53.6 | 94.3 |
| ✓ | ✗ | 75.9 | 52.5 | 93.0 |
| ✗ | ✓ | 76.6 | 52.2 | 93.2 |
| ✓ | ✓ | 75.7 | 52.2 | 92.6 |

Table 7: Ablation study of 3D joint contrastive learning. J2N: joint-to-non-joint contrast. J2J: joint-to-joint contrast.

scenes for evaluation without using any data from training set. Tab. 3, shows that our method outperforms previous 3D human pose estimation methods on CMU-Panoptic, which means our model also works well for indoor and daily life scenes.

4.2. Comparison to the State-of-the-Art on Standard Benchmark

3DPW [61] is the latest large-scale benchmark for 3D human mesh recovery. We do not use the training set and report performance on its test split which contains 60 videos and 3D annotations of 35515 persons. As shown in Tab. 2, Our method achieves state-of-the-art results among previous approaches. The results demonstrate the robustness of JOTR to a variety of in-the-wild scenarios.

4.3. Analysis.

In this section, we analyze the main components of JOTR and evaluate their impact on the mesh recovery performance. More details and ablation studies can be found in the supplementary material.

Utilization of 2D and 3D features: Tab. 4 demonstrates that the incorporation of 3D features is beneficial for mesh recovery

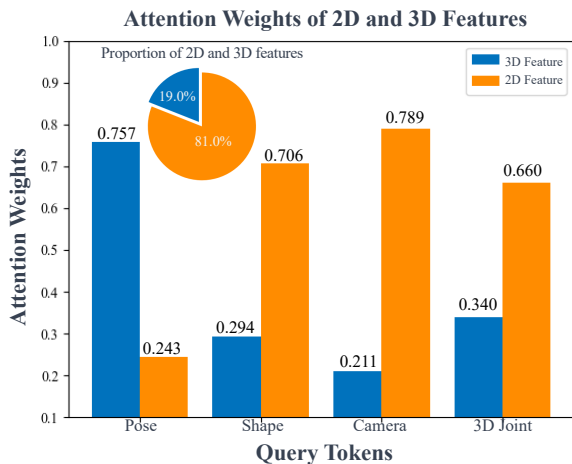


Figure 4: Visualization of cross-attention weights in the last refining layer. We randomly sample 100 persons from 3DPW test set and average the attention weights for visualization.

ery performance. For the utilization of 2D features, flattening shows better performance than sampling, which supports our hypothesis that sampling joint features in obscured regions could have a negative impact. For 3D features, we do not conduct experiments for flattening 3D features due to memory limitations. Moreover, we believe that 3D joint feature sampling is adequate for alleviating occlusion problems by attending to the accurate depth. Fig. 4 shows the attention weights in the last refining layer. The query tokens significantly pay more attention to 3D features, which validates the usefulness of our fusion framework.

Validation of coarse-to-fine regression: We validate the accuracy of intermediate predictions of fusion transformer in Tab. 5, which shows the coarse-to-fine regression process in JOTR.

Decoupling SMPL query: JOTR performance improvement is observed in Tab. 6 by decoupling SMPL query from joint query. In the experiment without decoupling, we employ mean pooling on the decoder’s output and regress SMPL parameters through MPLs. Decoupling SMPL query is presumed to enhance performance by reducing interference in executing other tasks (e.g., joint localization) during SMPL parameter regression.

3D Joint contrastive learning: The impact of 3D joint contrastive learning on the performance of JOTR is presented in Tab. 7. Both joint-to-non-joint and joint-to-joint contrastive losses result in improved performance, with the former being more effective as it incorporates global supervision for the entire 3D space. Our contrastive losses also lead to more compact and well-separated learned joint embeddings, as shown in Fig. 5. This indicates that our network can generate more discriminative 3D features, producing semantically clear 3D spaces and promising results.

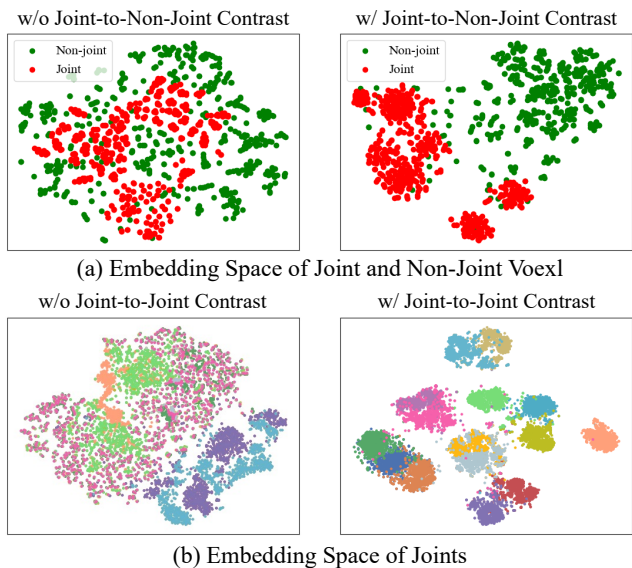


Figure 5: Visualization of features learned with (left) “local” joint supervision and (right) our “global” 3D joint contrast optimization objective (i.e., Eq. (4) and Eq. (5)) on 3DPW test set [61]. Each color stands for a kind of joint (e.g., head and right knee) in (b).

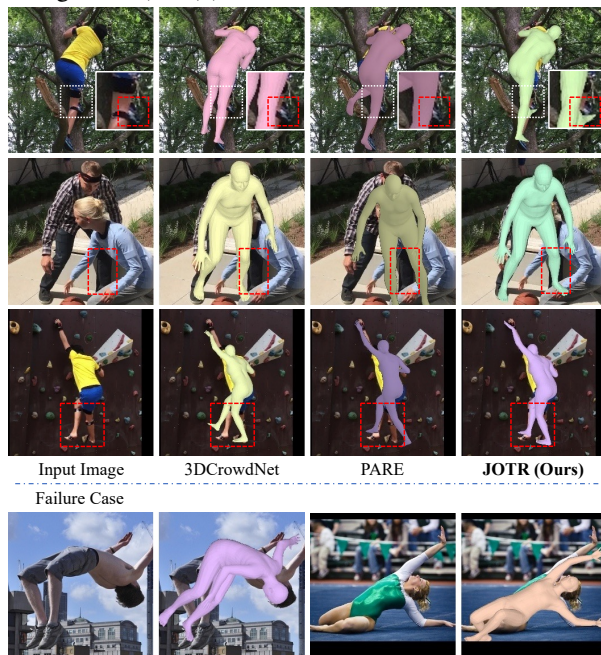


Figure 6: Qualitative results on 3DPW dataset [61]. Note that we use no data from 3DPW for training. The bottom row shows the failure cases of JOTR. JOTR performs badly on extreme poses due to the lack of training data. More qualitative results can be found in the supplementary material.

5. Conclusion

Many human mesh recovery methods focus on 2D alignment technologies, which would fail under occlusions or limited visibility. To address this limitation, we propose JOTR, a novel method that combines 2D and 3D features

using an encoder-decoder transformer architecture to achieve 2D&3D alignment. Furthermore, we introduce two novel 3D joint contrastive losses that enable global supervision of the 3D space of target persons, producing meaningful 3D representations. Extensive experiments on 3DPW benchmarks show that JOTR achieves the new state of the art.

Limitations and Broader Impact. 1) JOTR relies on the human pose predictor to detect 2D keypoints, leading to long inference times. 2) In the future, JOTR has the potential to be integrated with bottom-up 3D human mesh recovery methods for real-time applications.

Acknowledgements. This work was supported by the Natural Science Foundation of Zhejiang Province (DT23F020008) and the Fundamental Research Funds for the Central Universities (226-2023-00051).

References

- [1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021. 3
- [2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 3
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3
- [5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 3
- [6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3
- [7] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 3
- [8] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15793–15803, 2021. 4
- [9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787. Springer, 2020. 3
- [10] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 1, 2, 3, 6, 7
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [12] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J Black. Learning to regress bodies from images using differentiable semantic rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11250–11259, 2021. 3
- [13] Mihai Fieraru, Mihai Zanfir, Teodor Szenté, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. *Advances in Neural Information Processing Systems*, 34:19385–19397, 2021. 7
- [14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3
- [15] Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10472–10481, 2021. 3
- [16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 5
- [17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [19] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16291–16301, 2021. 3
- [20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. In *TPAMI*, 2014. 6
- [21] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 2, 3, 6, 7
- [22] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser

- Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. [2](#), [6](#), [7](#)
- [23] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. [7](#)
- [24] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. [3](#), [4](#)
- [25] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. [1](#), [2](#), [3](#), [6](#), [7](#)
- [26] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. [7](#)
- [27] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [6](#)
- [28] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1715–1725, 2022. [2](#), [3](#), [6](#), [7](#)
- [29] Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for monocular 3d human pose estimation. In *European Conference on Computer Vision*, pages 541–554. Springer, 2020.
- [30] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. [3](#)
- [31] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. [2](#), [3](#), [6](#), [7](#)
- [32] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. [1](#), [3](#), [6](#), [7](#)
- [33] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. [2](#), [3](#), [7](#)
- [34] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021. [3](#), [4](#)
- [35] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. [6](#)
- [36] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. [3](#)
- [37] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31th ACM International Conference on Multimedia*, 2023. [3](#)
- [38] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020.
- [39] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. [3](#)
- [40] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 1–18. Springer, 2022. [4](#)
- [41] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. [2](#), [3](#)
- [42] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12939–12948, 2021. [2](#), [3](#)
- [43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. [6](#)
- [44] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems*, 2018. [4](#)
- [45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. [3](#)
- [46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. [2](#), [3](#)
- [47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. [6](#)
- [48] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018. [6](#)

- [49] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. [6](#)
- [50] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. [3](#), [6](#), [7](#)
- [51] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. [5](#)
- [52] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. [7](#)
- [53] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. [3](#), [5](#)
- [54] Lennart Reiher, Bastian Lampe, and Lutz Eckstein. A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird’s eye view. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7. IEEE, 2020. [4](#)
- [55] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8887–8896, 2023. [3](#)
- [56] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021. [1](#), [2](#), [3](#), [6](#), [7](#)
- [57] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. [2](#), [3](#), [4](#), [7](#)
- [58] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019. [7](#)
- [59] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit Agrawal. Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. In *2020 International Conference on 3D Vision (3DV)*, pages 311–321. IEEE, 2020. [3](#)
- [60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. [2](#)
- [61] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. [2](#), [6](#), [7](#), [8](#)
- [62] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. [3](#), [5](#)
- [63] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. [4](#)
- [64] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. [6](#)
- [65] Weixiang Yang, Qi Li, Wenxi Liu, Yuanlong Yu, Yuexin Ma, Shengfeng He, and Jia Pan. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15536–15545, 2021. [4](#)
- [66] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021. [1](#)
- [67] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021. [3](#)
- [68] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 2022. [3](#)
- [69] Zongxin Yang, Xin Yu, and Yi Yang. Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2021. [1](#)
- [70] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. [7](#)
- [71] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. *Advances in Neural Information Processing Systems*, 31, 2018. [7](#)
- [72] Hongwen Zhang, Yating Tian, Xinchu Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021.

1, 2, 3, 6, 7

- [73] Jiangning Zhang, Xiangtai Li, Jian Li, Liang Liu, Zhucun Xue, Boshen Zhang, Zhengkai Jiang, Tianxin Huang, Yabiao Wang, and Chengjie Wang. Rethinking mobile block for efficient neural models. *ICCV*, 2023. 3
- [74] Jiangning Zhang, Xiangtai Li, Yabiao Wang, Chengjie Wang, Yibo Yang, Yong Liu, and Dacheng Tao. Eatformer: Improving vision transformer inspired by evolutionary algorithm. *arXiv preprint arXiv:2206.09325*, 2022.
- [75] Jiangning Zhang, Chao Xu, Jian Li, Wenzhou Chen, Yabiao Wang, Ying Tai, Shuo Chen, Chengjie Wang, Feiyue Huang, and Yong Liu. Analogous to evolutionary algorithm: Designing a unified sequence model. *NeurIPS*, 34:26674–26688, 2021. 3
- [76] Tianshu Zhang, Buzhen Huang, and Yangang 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*, pages 7376–7385, 2020. 2, 3, 6, 7
- [77] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021. 3
- [78] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 5