OCR-Pose: Occlusion-aware Contrastive Representation for Unsupervised 3D Human Pose Estimation

Junjie Wang*
Zhenbo Yu*
dreamboy.gns@sjtu.edu.cn
yuzhenbo@sjtu.edu.cn
Shanghai Jiao Tong University

Zhengyan Tong
418004@sjtu.edu.cn
Shanghai Jiao Tong University

Hang Wang
wang--hang@sjtu.edu.cn
Huawei Hisilicon

Jinxian Liu
liujinxian@sjtu.edu.cn
Shanghai Jiao Tong University

Wenjun Zhang
zhangwenjun@sjtu.edu.cn
Shanghai Jiao Tong University

Xiaoyan Wu†
xiaoyanwu@sjtu.edu.cn
Shanghai Jiao Tong University

ABSTRACT

Occlusion is a significant problem in 3D human pose estimation from the 2D counterpart. On one hand, without explicit annotation, the 3D skeleton is hard to be accurately estimated from the occluded 2D pose. On the other hand, one occluded 2D pose might correspond to multiple 3D skeletons with low confidence parts. To address these issues, we decouple the 3D representation feature into view-invariant part termed occlusion-aware feature and view-dependent part termed rotation feature to facilitate subsequent optimization of the former. Then we propose an occlusion-aware contrastive representation based scheme (OCR-Pose) consisting of *Topology Invariant Contrastive Learning* module (**TiCLR**) and *View Equivariant Contrastive Learning* module (**VeCLR**). Specifically, TiCLR drives invariance to topology transformation, i.e., bridging the gap between an occluded 2D pose and the unoccluded one. While VeCLR encourages equivariance to view transformation, i.e., capturing the geometric similarity of the 3D skeleton in two views. Both modules optimize occlusion-aware constrastive representation with pose filling and lifting networks via an iterative training strategy in an end-to-end manner. OCR-Pose not only achieves superior performance against state-of-the-art unsupervised methods on unoccluded benchmarks, but also obtains significant improvements when occlusion is involved. Our project is available at https://sites.google.com/view/ocr-pose.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

3D vision, human pose estimation,representation learning

*Both authors contributed equally to this research.
†Corresponding Author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547780

ACM Reference Format:
Junjie Wang, Zhenbo Yu, Zhengyan Tong, Hang Wang, Jinxian Liu, Wenjun Zhang, and Xiaoyan Wu. 2022. OCR-Pose: Occlusion-aware Contrastive Representation for Unsupervised 3D Human Pose Estimation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547780

1 INTRODUCTION

Estimating 3D human poses from monocular images is an important step in various applications, such as robotics, Human-Computer Interaction, Virtual Reality, etc. Generally, methods under this task detect individuals in each image, estimate the 2D pose within each person's bounding box via off-the-shelf detectors, and finally convert the 2D pose to a 3D pose. Previous existing unsupervised methods are designed for the scenarios where human bodies are well captured without occlusion. However, occlusions (i.e., self-occlusion and object-occlusion) are very common in practical applications. And it is worth mentioning that occluded 3D human pose estimation is quite challenging, especially in an unsupervised setting. In this paper, we are interested in the occlusion problem for unsupervised monocular 3D human pose estimation.

The challenges of the occlusion problem in the task are two-fold: (a.) 3D skeleton is difficult to be accurately estimated from an occluded 2D pose lacking reliable 2D joints (see Fig. 1). We experimentally find that the performance of unsupervised 3D pose estimation is prone to drift seriously with occluded 2D poses. Previous methods [1, 17, 38] commonly exploit topology priors (e.g., kinematics priors, adversarial priors, geometric constraints, etc.) to alleviate the corresponding ambiguity. However, only relying on these topology priors is often ineffective under the occlusion setting. (b.) Multi-view data [5, 11, 15] can effectively address such kind of ambiguity. However, multi-view images require dedicated multi-camera equipment, which is often high-cost and unavailable. Yu et al. [38] utilize the generated pseudo view to construct multi-view consistency constraint, which hardly generalizes to the occlusion setting. Although Fig. 1 illustrates that either object-occluded or self-occluded part may be visible in another view, it remains challenging to exploit multi-view information directly.

To solve the above challenges, we decouple the representation feature generated by unsupervised pose lifting baseline [38] into the

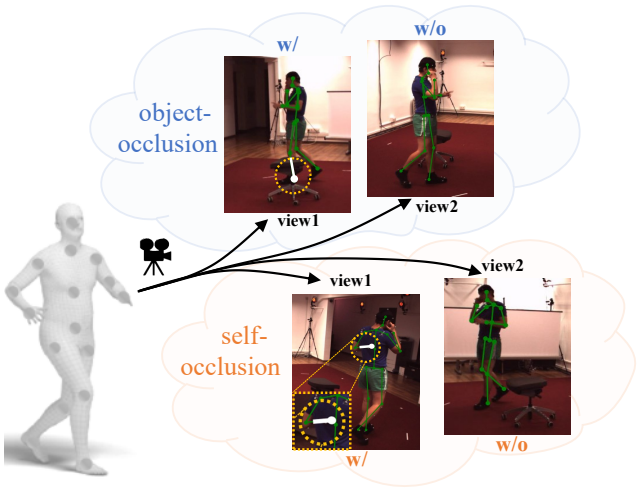


Figure 1: Illustration of occlusion problem (i.e., object-occlusion and self-occlusion in two views) for 3D pose estimation. Notably, either object-occluded part (e.g., left leg in the top yellow circle) or self-occluded part (e.g., left hand in the bottom yellow circle) is visible in another view.

view-invariant part termed occlusion-aware feature and the view-dependent part termed rotation feature to facilitate the subsequent optimization of the former feature. The decoupled semantic feature is experimentally proved to promote the following contrastive learning. Then, as is shown in Fig. 3, we propose an occlusion-aware contrastive learning-based scheme (OCR-Pose) consisting of: 1) *Topology Invariant Contrastive Learning* module (**TiCLR**) to drive invariance to topology transformation, i.e., bridging the gap between occluded 2D pose and unoccluded one. In TiCLR, the positive sample is the corresponding unoccluded 2D pose, whereas the negative samples refer to other occluded 2D ones. 2) *View Equivariant Contrastive learning* module (**VeCLR**), to encourage equivariance to view transformation, i.e., capturing geometric similarity of the 3D skeleton in two views. In VeCLR, the positive sample is the projected 2D pose with occlusion, which can be utilized to generate 3D skeleton in another pseudo view, while the negative samples correspond to other 3D skeletons with large differences filtered by attention mechanism. Furthermore, TiCLR and VeCLR encourage the latent space generated by positive pairs to lie close to each other, and push the one produced by negative pairs apart.

Extensive experiments demonstrate that our approach is effective in both unoccluded and occluded scenarios. Compared with the state-of-the-art unsupervised methods (e.g. [38]), our approach not only achieves superior performance on the unoccluded indoor benchmarks like Human3.6M [10], but also obtains significantly improvement (65.8% in terms of P-MPJPE) in simulated occlusion conditions. Furthermore, experiments on 3DPW [34] and 3DOH50K [40] illustrate that our approach provides reasonable predictions when real-world occlusion is involved.

2 RELATED WORK

Unsupervised 3D Pose Estimation. Most works [10, 11, 15, 21, 30, 39] perform fully / weakly supervised 3D pose estimation with densely annotated 3D joints [10] or 2D key-points [21]. The annotation process itself could be a laborious task, which motivates research towards unsupervised 3D pose estimation. Under such condition, the usage of any ground truth 3D pose information or relevant projection is not allowed, which is more practical and challenging compared to supervised setting. Rhodin *et al.* [25] propose to learn a geometry-aware body representation from multi-view images without annotations, where multi-view information is used as a guidance signal for learning geometry-aware representation. Chen *et al.* [1] exploit the geometric self-consistency through the lift-reproject-lift process. Several works [13, 17] aim to learn the 2D key-points via background/foreground disentangling. Recently, Kundu et.al. [16] propose to explicitly constrain the 3D pose by modeling it at its most fundamental form of rigid and non-rigid transformations, resulting in interpretable 3D pose predictions, even without any auxiliary 3D cues such as multi-view or depth information. Yu *et al.* [38] introduce a 2D pose scale estimation module and then map optimized 2D pose to 3D counterpart via a pose lifting module to alleviate the scale and pose ambiguity.

Pose Estimation with Occlusion. Pose estimation often suffers from performance degradation due to occlusion [4, 14, 26–28, 40, 41]. To tackle this problem, Zhou *et al.* [41] propose an occlusion-aware siamese network equipped with erasing and reconstruction submodule to obtain cleaner feature representation and reconstruct the information destroyed by occlusion. To improve the robustness under occlusion, CenterHMR [31] predicts the Center maps and the Parameter maps, which represent the location of each human body center and the corresponding parameter vector of 3D human mesh at each center. Cheng *et al.* [4] first filter out the unreliable estimations of occluded key-points and then feed incomplete 2D key-points to temporal convolutional networks, which further produces a complete 3D pose via constraining the temporal smoothness. To solve object occlusion, Zhang *et al.* [40] take a partial UV map representation for object-occluded 3D human body and convert the full 3D human pose estimation as an image inpainting problem. In this work, we propose two occlusion-aware contrastive learning module to explicitly address the occlusion problem.

Contrastive Representation Learning. Contrastive learning aims to learn representations from unlabeled data by instance discrimination. CPC [33] achieves this goal by maximizing the mutual information between correlated instances with an InfoNCE loss. SimCLR [3] presents a simple framework by maximizing agreement between differently augmented views of the same data example. Due to its effectiveness, contrastive representation learning has been applied to many vision tasks [3, 8, 18]. Recently, Mitra *et al.* [23] leverage multiview consistency to guide 3D human pose regression based on contrastive loss. Spurr *et al.* [29] use contrastive learning to learn the self-supervised representation with geometric consistency for 3D hand pose estimation. Li *et al.* [19] propose a cross-view contrastive learning framework for skeleton-based action representation, where cross-view consistent knowledge mining is developed to excavate useful samples across views. However, previous works are not carefully designed for occlusion problem.

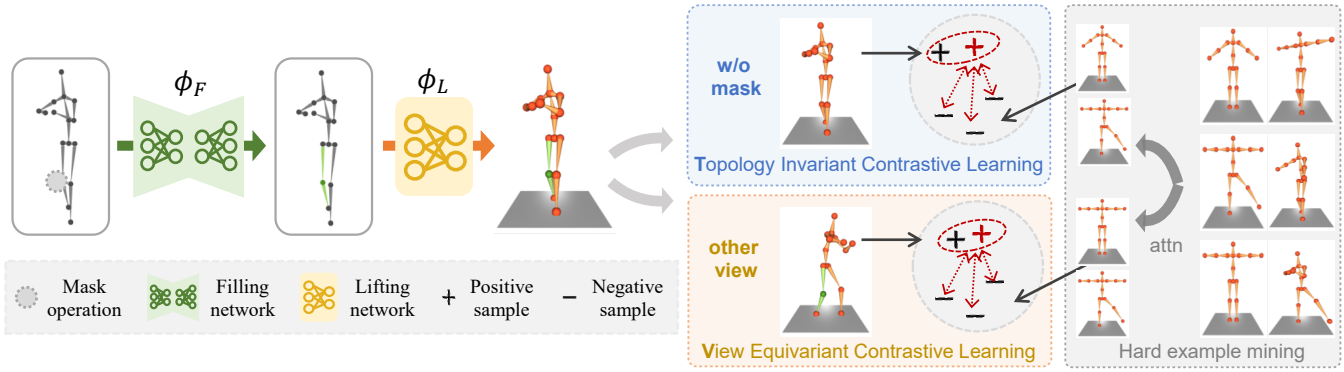


Figure 2: Overview of OCR-Pose. OCR-Pose exploits the backbone network (i.e., the pose filling and lifting networks) to obtain occlusion-aware contrastive learning representation, which is used to regress 3D skeleton directly. Notably, the occlusion-aware feature is optimized with TiCLR module and VeCLR module, simultaneously. Both modules encourage the latent space generated by positive pairs to lie close to each other, and push the one produced by negative pairs apart.

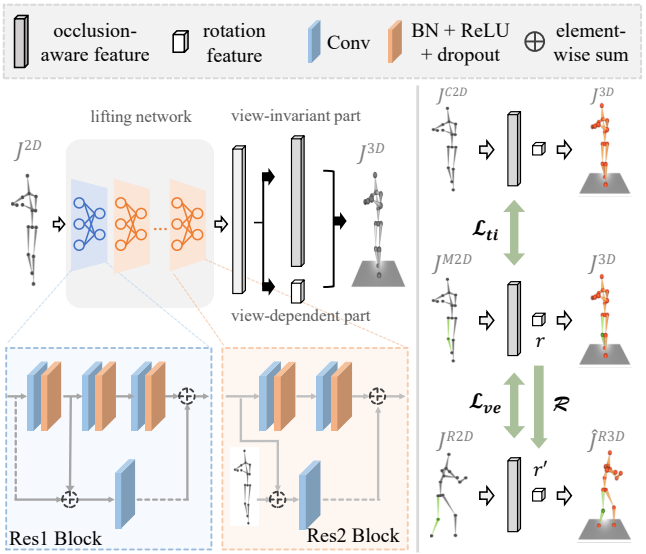


Figure 3: Illustration of feature decoupling process. The pose lifting network converts the 2D pose into a decoupled semantic feature containing a view-dependent part and a view-invariant part. Then an occlusion-aware constrastive learning-based scheme (OCR-Pose) containing TiCLR and VeCLR is proposed to optimize the view-invariant part termed occlusion-aware feature. Specifically, the view-dependent part is trained by relative rotation consistency during the view cycle pipeline. While the view-invariant part (i.e., occlusion-aware feature) is optimized with TiCLR and VeCLR constraints to enforce occlusion and view irrelevance. Finally, both two parts jointly regress the relative depth and then obtain the estimated 3D skeleton.

3 METHODS

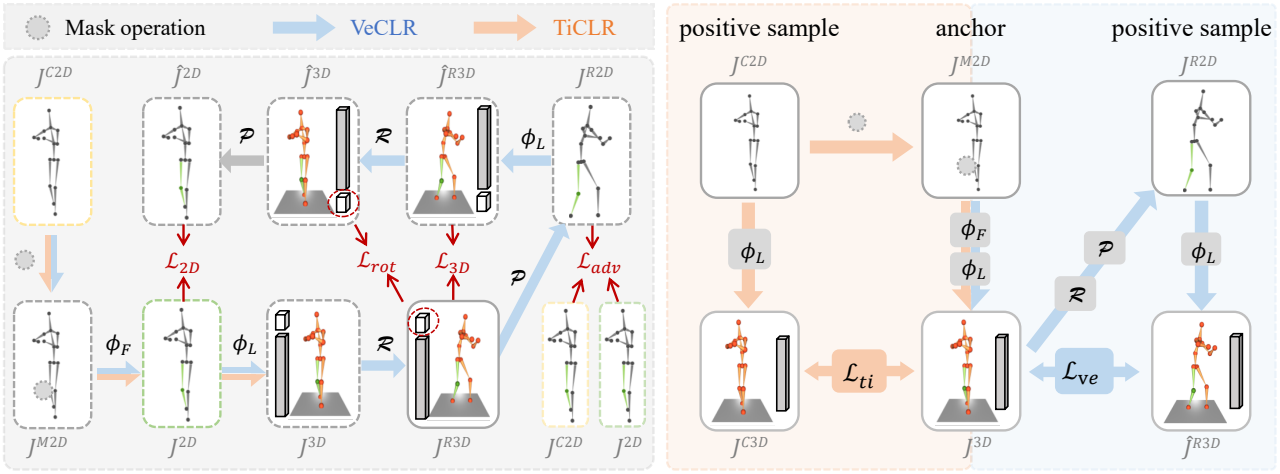
3.1 Overview

The overall framework of the proposed OCR-Pose is illustrated in Fig. 2, which consists of TiCLR and VeCLR modules. Additionally, the backbone of OCR-Pose contains two basic networks, i.e., a pose filling network Φ_F completing occluded 2D pose, and a pose lifting network Φ_L mapping 2D pose to 3D skeleton. We exploit off-the-shelf pose detectors [6, 35] to predict 2D joints with confidence score as input. Meanwhile, the confidence score produced by 2D detectors is an informative indicator for the reliability of joint locations. We experimentally find that masking 2D joints with a low confidence score exhibits better performance than using complete but incorrect ones (see Sec. 4.5). That is to say, our OCR-Pose can lift complete 2D poses, incomplete 2D poses, or complete 2D poses with confidence scores to 3D skeletons, which diminishes the dependency on the accuracy of 2D detectors. Notably, for complete 2D joints, we use masking operation to simulate occluded 2D joints. For incomplete 2D joints, we need to use an extra filling network to map incomplete 2D joints to complete 2D joints. Then the following training procedure is the same as that for complete 2D joints.

In Sec. 3.2 we make a review of the classic unsupervised 3D human pose estimation pipeline [1, 38]. In Sec. 3.3 we introduce our decoupled occlusion-aware feature representation based on the classic pipeline. In Sec. 3.4 and Sec. 3.5 we propose two novel contrastive modules to optimize the decoupled representation.

3.2 Pose Lifting Baseline

In this section, we make a review of classic unsupervised pose estimation baseline [1, 38]. The framework consists of a pose lifting module Φ_L and a discriminator Φ_D. The pose lifting network takes 2D pose $J^{2D} = (x_i, y_i)_{i=1}^{N_J}$ as input to regress relative depth d_i and the absolute depth can be calculated by $z_i = \max(1, d_i + D)$, where D is a pre-defined distance between the camera and the human skeleton. Then estimated 3D joints J^{3D} can be computed as $J_i^{3D} = (x_i z_i, y_i z_i, z_i) = \Phi_L(J^{2D})$. J^{3D} is then randomly transformed to $J^{R3D} = \mathcal{T}(J^{3D})$. The random transform $\mathcal{T}$ is composed of a random rotation $\mathcal{R}$ and a fixed translation to the plane $(0, 0, D)$. Further, transformed 3D joints J^{R3D} is projected to $J^{R2D} = \mathcal{P}(J^{R3D})$, re-lifted to $\hat{J}^{R3D} = \Phi_L(J^{R2D})$, inversely transformed to $\hat{J}^{3D} = \mathcal{T}^{-1}(\hat{J}^{R3D})$ and finally projected to $\hat{J}^{2D} = \mathcal{P}(\hat{J}^{3D})$. In this way, a loop closure is constructed and L2 loss is used to constrain the 2D /



(a) occlusion-aware feature generation (b) occlusion-aware feature optimization

Figure 4: Detailed structure of OCR-Pose. The basic cycle structure in the left contains one pose lifting network, one pose filling network and a discriminator, with one mask operation, two rotation operations, and two projection operations. Two contrastive representation learning modules in the right consist of TiCLR and VeCLR modules.

3D consistency as in Eq. 1.

$$\mathcal{L}_{2D} = \|J^{2D} - \hat{J}^{2D}\|_2 \quad \mathcal{L}_{3D} = \|J^{R3D} - \hat{J}^{R3D}\|_2 \tag{1}$$

Furthermore, a discriminator Φ_D is introduced to encourage the rationality of 2D pose and lifted 3D skeleton via adversarial training. Specifically, the discriminator is trained to discriminate input 2D pose as real, projected 2D poses as fake, while the pose lifting network is trained to produce realistic 3D estimations, the projections of which can fool the discriminator. The adversarial loss function is shown in Eq. 2.

$$\begin{aligned} \min_{\theta_L} \mathcal{L}_{adv}(\Phi_L) &= \mathbb{E}[\Phi_D(J^{R2D}) - 1]^2 \\ \min_{\theta_D} \mathcal{L}_{adv}(\Phi_D) &= \mathbb{E}[\Phi_D(J^{2D}) - 1]^2 + \mathbb{E}[\Phi_D(J^{R2D})]^2 \end{aligned} \tag{2}$$

By exploiting self-consistency and the prior learnt from adversarial training, the unsupervised baseline can produce reasonable 3D estimations. However, *a key problem* is the sensitivity to the distribution of input 2D joints. Missing joints(e.g. due to occlusion) or inaccurate estimations can significantly degrade the performance. To alleviate this problem, we introduce an extra pose filling network and two contrastive modules, which leverage the occlusion-aware representation for more robust and accurate 3D estimations.

3.3 Decoupled Semantic Feature

In this section, as illustrated in Fig. 3, we exploit the pose lifting network [38] to convert the 2D pose into a decoupled semantic feature comprised of a view-invariant part (occlusion-aware feature) and a view-dependent part (rotation feature), motivated by the unsupervised baseline architecture and its limitation discussed in Sec. 3.2. Then an occlusion-aware contrastive learning-based scheme containing TiCLR and VeCLR modules is utilized to optimize the occlusion-aware feature. We experimentally find that the decoupled semantic feature is able to facilitate the contrastive learning performance, yielding more stable convergence. Three major improvements on the pose lifting network are exhibited as follows: **1) A pose filling network Φ_F is designed to complete missing joints or refine inaccurate 2D joints.** Given incomplete 2D joints J^{M2D}(which can be caused by occlusion or obtained by masking complete one $J^{M2D} = \mathcal{M}(J^{C2D})$), Φ_F outputs the filled results J^{2D}. L2 loss in Eq. 3 is used to supervise Φ_F.

$$\mathcal{L}_{pf} = \|J^{2D} - J^{C2D}\|_2 \tag{3}$$

2) The proposed pose filling network is trained along with the pose lifting network in a unified manner. Not only Eq. 3 is optimized together with the pose lifting network, but also the pose filling model is incorporated into the adversarial training scheme, which leads to Eq. 4.

$$\begin{aligned} \min_{\theta_F,\theta_L} \mathcal{L}_{adv}(\Phi_{F,L}) &= \mathbb{E}[\Phi_D(J^{2D}) - 1]^2 + \mathbb{E}[\Phi_D(J^{R2D}) - 1]^2 \\ \min_{\theta_D} \mathcal{L}_{adv}(\Phi_D) &= \mathbb{E}[\Phi_D(J^{C2D}) - 1]^2 \\ &\quad + \mathbb{E}[\Phi_D(J^{2D})]^2 + \mathbb{E}[\Phi_D(J^{R2D})]^2 \end{aligned} \tag{4}$$

The advantages of unifying the two networks instead of using a standalone filling network are two-fold. Firstly, the diverse filled 2D joints produced by Φ_F make the pose lifting network more robust to input 2D distribution. Secondly, adversarial training encourages the pose filling model to produce more realistic 2D joints to fool the discriminator. **3) We decouple the 3D representation into a view-invariant part (occlusion-aware feature) and a view-dependent part (rotation feature).** The two features are learnt from separate linear layers and then concatenated to regress the relative depth z. The occlusion-aware feature is a latent vector $F_{occ} \in \mathbb{R}^{d_{occ}}$, while the rotation features $F_{rot} \in \mathbb{R}^6$ take the explicit form of rotation matrices(6D representation). The two matrices r, r' produced by two lifting passes are constrained by relative rotation consistency in Eq. 5.

$$\mathcal{L}_{rot} = \|r^{-1}r' - r^*\|_2 \tag{5}$$

r^* is the reference rotation matrix in random rotation $\mathcal{R}$. When decoupled from the view-dependent features, view-invariant features focus more on the topology itself, providing a representation better prepared for occlusion-aware learning. In Sec. 3.4 and Sec. 3.5 we propose two contrastive modules to optimize above representation.

3.4 TiCLR module

The proposed occlusion-aware representation should be robust to topology transformation. In other words, the representation should not change dramatically when part of the 2D pose is occluded. To this end, we design a topology invariant contrastive learning module(TiCLR), which optimizes over the occlusion-aware representation via an attention mechanism and a contrastive loss $\mathcal{L}_{ti}$.

A contrastive loss [7] is a metric indicating the similarity between the anchor sample and the corresponding positive one, and the dissimilarity between the anchor sample and the others. In Fig. 4 (b), the positive sample is the corresponding unoccluded 2D pose, whereas the negative samples refer to other occluded 2D ones. Furthermore, we design a hard example mining module Φ_A, consisting of an embedding layer, a dot product layer followed by softmax operation, which exploits the attention mechanism to widen the difference between the negative pairs.

Specifically, we denote $z_{ti}, z^+_{ti}, z^-_{ti} \in \mathbb{R}^{D_z}$ as the embeddings selected from the occlusion-aware features, corresponding to the anchor sample, the positive sample, and the negative samples. Then we ensemble the embedding features of one anchor, one positive sample and K negative samples into a sample matrix $M = (z^T_{ti}; (z^+_{ti})^T; (z^-_{ti})^T), M \in \mathbb{R}^{(K+2)\times D_z}$. The attention matrix is computed by $W = \Phi_A(M), W \in \mathbb{R}^{(K+2)\times(K+2)}$. We update the original sample matrix M by $M' = (z^T_{tia}; (z^+_{tia})^T; (z^-_{tia})^T) = WM$.

Finally, the contrastive loss is calculated as follows:

$$\mathcal{L}_{ti} = -\log \frac{\exp(z_{tia} \cdot z^+_{tia}/\tau)}{\exp(z_{tia} \cdot z^+_{tia}/\tau) + \sum_{j=1}^{K} \exp(z_{tia} \cdot z^-_{tia,j}/\tau)} \tag{6}$$

Where τ is the temperature hyper-parameter [9], and dot product $\cdot$ is to compute their similarity where each component is normalized. The sum is over one positive and K negative samples. Constrained by contrastive loss $\mathcal{L}_{ti}$, the TiCLR module is able to learn the information of missing joints in an unsupervised manner.

3.5 VeCLR module

With the feature decoupling strategy in Sec. 3.3, the occlusion-aware representation is expected to be view-equivariant(i.e. unchanged under any rotation transform). From Fig. 1, we can observe that a visible joint can be occluded in another viewpoint, which motivates us to design a view-equivariant contrastive learning module(VeCLR) to leverage the occlusion-aware representation. Specifically, we constrain the representation to be similar under a sequence of rotation, projection, masking and re-lifting operations(see Fig. 2). Similar to Eq. 6, we present the InfoNCE loss $\mathcal{L}_{ve}$ as follows:

$$\mathcal{L}_{ve} = -\log \frac{\exp(z_{vea} \cdot z^+_{vea}/\tau)}{\exp(z_{vea} \cdot z^+_{vea}/\tau) + \sum_{j=1}^{K} \exp(z_{vea} \cdot z^-_{vea,j}/\tau)} \tag{7}$$

where $z_{vea}, z^+_{vea}, z^-_{vea} \in \mathbb{R}^{D_z}$ indicate the embedding features with attention mechanism generated by the anchor sample, positive sample, and negative samples. Notably, the positive sample is the representation of the re-lifted 3D skeleton after the random rotation, while negative samples correspond to other 3D skeletons with a significant difference filtered by attention mechanism.

In the end, TiCLR and VeCLR both optimize occlusion-aware representation with pose filling and pose lifting network with $\mathcal{L}_{backbone}$ and $\mathcal{L}_{occlusion}$ via an iterative training strategy [38].

$$\begin{aligned}\mathcal{L}_{backbone} = w_{pf}\mathcal{L}_{pf} + w_{2D}\mathcal{L}_{2D} + w_{3D}\mathcal{L}_{3D} \\ + w_{adv}\mathcal{L}_{adv} + w_{rot}\mathcal{L}_{rot}\end{aligned} \tag{8}$$

$$\mathcal{L}_{occlusion} = w_{ti}\mathcal{L}_{ti} + w_{ve}\mathcal{L}_{ve} \tag{9}$$

Where $\mathcal{L}_{backbone}$ is used to train the backbone network (i.e., pose filling and lifting network) to obtain more reasonable 3D skeletons, and $\mathcal{L}_{occlusion}$ is utilized to optimize the pose filling network to achieve more accurate completed 2D poses. Both two losses are complementary to each other and reduce the learning difficulty by a large margin.

3.6 Implementation Details

Network Architecture. We use a simple auto-encoder structure with 4 layers and hidden dimension $h_1 = 128$ for the pose filling network. For the pose lifting network and the discriminator, we use residual building blocks like Yu et al. [38], with dimension $h_2 = 512$. View-invariant feature dimension d_{occ} is set to 512. The pose lifting network has 4 blocks and the discriminator has 2 blocks.

Training Details. For hyper-parameters regarding the pose lifting baseline, we keep the same with [38]. Constant D is set to 10, the azimuth is sampled from $[-7\pi/9, 7\pi/9]$ and elevation is sampled from $[-\pi/9, \pi/9]$ in random rotation.

In our design, the pose filling network and the pose lifting network are optimized in an iterative training strategy. Specifically, we first train the pose filling network with $\mathcal{L}_{pf}$ for 10 epochs as a warmup. Then we optimize the backbone network (i.e., pose filling and lifting network) along with the two contrastive modules via an iterative training strategy with $\mathcal{L}_{backbone}$ and $\mathcal{L}_{occlusion}$ for 100 epochs. Loss weights in Eq.9 are set to $w_{pf} = 10, w_{2D} = 5, w_{3D} = 0.5, w_{adv} = 1, w_{ti} = 0.3, w_{ve} = 0.1$. The temperature parameter in the contrastive modules is set to $\tau = 0.5$. We adopt Adam optimizer with initial learning rate equal to 0.0002 and batch size is set to 512. Since negative samples in both contrastive modules are sampled from the batch, we have $K = 511$.

4 EXPERIMENTS

4.1 Datasets And Metrics

Human3.6M [10] is a large scale in-door dataset, which is comprised of about 3.6 million frames with densely annotated 3D annotations. Human3.6M is the most widely used benchmark for 3D human pose estimation.

MPI-INF-3DHP [22] is another 3D human pose dataset, consisting of over 1.3 million frames from multiple viewpoints. Different from Human3.6M, some sequences in MPI-INF-3DHP are captured in the wild, making it more suitable for stronger algorithms.

3DPW [34] is a completely in-the-wild dataset with more complicated scenes. 3DPW consists of 60 video sequences. Diverse actions and occlusion make the dataset more challenging.

Supervision	Algorithm	GT	Pre
Fully Supervised	Martinez et al. [21]	37.1	47.7
	Pavllo et al. [24]	27.2	36.5
	Wang et al. [36]	-	32.7
Weakly Supervised	AIGN et al. [32]	79.0	97.4
	Kanazawa et al. [12]	-	67.5
	Drover et al. [5]	38.2	64.6
	Li et al. [20]	-	66.5
Unsupervised	Rhodin et al. [25]	-	98.2
	Chen et al. [1]	51.0	68.0
	Kundu et al. [16]	-	62.4
	Yu et al. [38]	46.0	54.9*
	Ours	**44.7**	**54.7**

Table 1: Experimental results on the test set of Human3.6M. The input 2D joints are unoccluded. GT / PRE denote results from ground truth 2D pose and estimated 2D pose by 2D detector respectively. *indicates single-frame results released in their github page[1].

Supervision	Algorithm	Trainset	PCK↑	AUC↑
Fully Supervised	Mehta et al. [22]	H36M	64.7	31.7
	Chen et al. [2]	3DHP	87.9	54.0
	Wang et al. [36]	3DHP	86.9	62.1
Weakly Supervised	Kanazawa et al. [12]	3DHP	77.1	40.7
	Zhou et al. [42]	H36M	69.2	32.5
Unsupervised	Chen et al. [1]	H36M	64.3	31.6
	Yu et al. [38]	H36M	82.2	46.6
	Ours	H36M	**83.4**	**47.3**

Table 2: Experimental results on the test set of MPI-INF-3DHP with unoccluded 2D joints.

Max #joints to mask(N)	Method	Visible	All
0	Yu et al. [38]	46.0	46.0
0	Ours	44.7	44.7
3	Yu et al. [38]	130.8	289.6
3	Ours(w/o decouple)	51.5	58.9
3	Ours(w/ decouple)	49.0	56.6
3	Ours(+VeCLR)	47.4	55.1
3	Ours(+VeCLR+TiCLR)	**47.0**	**54.8**
4	Ours(+VeCLR+TiCLR)	53.1	62.4
5	Ours(+VeCLR+TiCLR)	55.2	72.7

Table 3: Experimental results on the test set of Human3.6M under occluded conditions. A random number $n \sim \mathcal{U}(0, N)$ of joints are masked from the 2D inputs. "Visible" means evaluation is only performed on unmasked 2D key-points. "All" means evaluation is performed on all 2D key-points.

3DOH50K [40] is the first real 3D human dataset designed especially for occlusion scenarios in the problem of human reconstruction and pose estimation. It contains 51600 samples.

Evaluation Metrics We report the most widely used metric P-MPJPE, which is the euclidean distance between ground-truth 3D poses and predictions after rigid alignment. We also report the **P**ercentage of **C**orrect **K**eypoints (PCK3D) and **A**rea **U**nder **C**urve (AUC) for MPI-INF-3DHP.

4.2 Experimental Settings

To explore the performance of unsupervised 3D pose estimation under occlusion, our experimental settings are listed as follows: (a). **Unocclusion condition.** This setting is the same as the standard one. We use masking operation in the training procedure. We design this setting to show the superiority of our OCR-Pose compared with previous methods fairly. (b). **Occlusion condition.** This setting is only different from the standard one in the testing procedure. We use incomplete 2D joints as input instead of complete 2D joints. We present this setting to exhibit the effectiveness of our method under occlusion. (c). **Unocclusion condition with confidence score.** This setting has extra confidence scores compared the standard one a), which is very common when using off-the-shelf 2D detectors. (d). **Unocclusion condition with occluded training data.** This setting is only different from the standard one in the training procedure, where we use mixed 2D joints (i.e., complete and incomplete 2D joints). We propose this setting to show that our method can generalize to heavily occluded datasets including self-occlusion or object-occlusion.

4.3 Results in unoccluded Condition

To better prove the effectiveness of the occlusion-aware representation, we conduct experiments on simulated occlusion scenarios on Human3.6M [10]. During training, we select possibly occluded joints from 2D skeleton according to semantic segmentation labels [37] or key-point confidence, and then randomly mask out some joints. We evaluate the proposed OCR-Pose in unoccluded conditions and report the results in Tab. 1. On Human3.6M, our approach achieves 44.7 P-MPJPE when input is 2D ground-truth and 54.7 when input is 2D predictions. The performance is superior to most weakly-supervised and all unsupervised methods. In particular, our approach obtains slight improvement over a recent state-of-the-art unsupervised method Yu et. al. [38](we take the **single-frame** performance from their github page[1]). Moreover, when the model trained on Human3.6M is transferred to MPI-INF-3DHP, our approach surpasses [38] again, proving the generalization ability of our approach.

4.4 Results in occluded Condition

To verify the robustness of the proposed framework under occlusion, we evaluate on simulated occlusion scenarios on Human3.6M [10]. Besides, occlusion dataset 3DPW [34] and 3DOH50K [40] are also evaluated to demonstrate the generalization ability to real-world occlusion conditions.

[1] https://github.com/deepinsight/insightface/tree/master/body/human_pose/ambiguity_aware

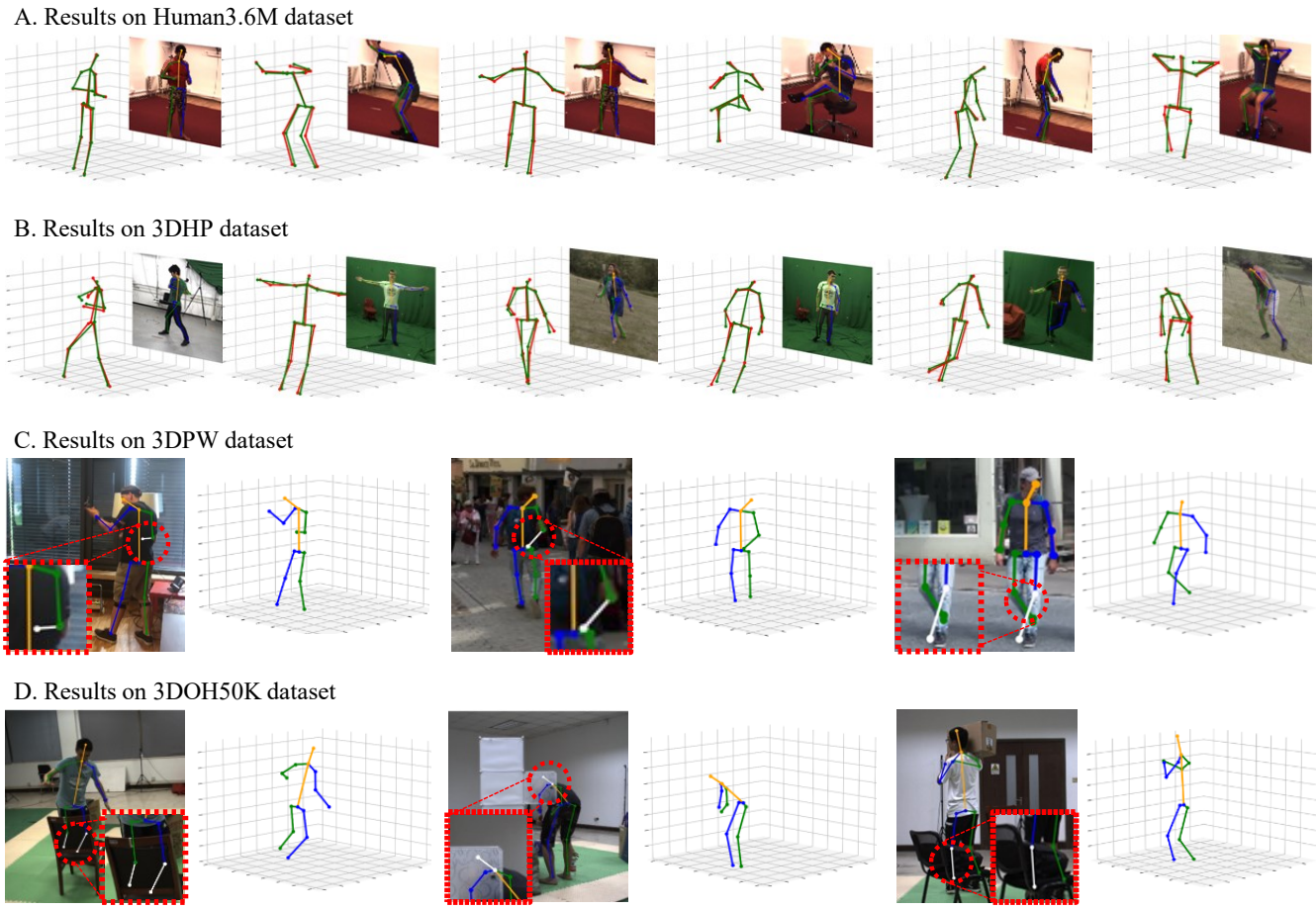


Figure 5: Qualitative results on 4 different datasets. Notably, predictions (red) along with ground truth (green) are illustrated in each part (A/B). 1st Row: Human3.6M. 2nd Row: MPI-INF-3DHP. 3rd Row: 3DPW. 4th Row: 3DOH50K.

As can be seen in Tab. 3, the performance of [38] drops dramatically with the introduction of occlusion. The reason is that classic 2D-3D pose lifting networks are sensitive to the change in input 2D distribution. In other words, outliers in the input-side (e.g. occlusion) will be a disaster for the network trained on unoccluded data. In contrast, in our approach, the pose filling network completes the missing parts conditioned on the whole skeleton, which also makes the input 2D distribution more stable and better prepared for the lifting network. We can observe that even when we mask out up to 5 parts in 2D inputs, the performance is still comparable to [32]. Notably, *visible only evaluation*(second last column in Tab. 3) in also shows significant improvements on *unmasked* key-points.

On 3DPW [34], unassigned 2D key-points are treated as occluded ones. On 3DOH-50K [40], we use AlphaPose [6] to obtain 2D predictions with confidence scores. Then key-point with confidence below 0.1 are treated as invisible (in Sec. 4.5 we exploit the usage of confidence). Qualitative results are shown in Fig. 5. Our approach provides reasonable pose completion and accurate 3D estimation for in-the-wild scenarios.

4.5 Results in Unocclusion with Confidence

During the past few years, mainstream 2D detectors [6] follow the design of key-point heatmaps, where confidence for each key-point is available. Our approach can make use of such scores to make more accurate predictions.

In scenarios with minor occlusion where 2D detector is competent, it's unnecessary to mask out 2D joints. Rather, it works better to refine 2D joints with pose filling network:

$$X_{out} = c_{det} X_{det} + (1 - c_{det}) X_{pf} \tag{10}$$

where c_{det} is the confidence score produced by 2D detectors to weight between original 2D estimations X_{det} and the output of pose filling model X_{pf}. Actually Eq. 10 can be unified with scenarios without key-point confidence by setting c_{det} for visible key-points to 1 and others to 0. To verify the effectiveness of such formulation, we use AlphaPose [6] to obtain 2D predictions on the test set of 3DOH50K and use our model trained on Human3.6M to obtain 3D estimations. When using the filling network to refine 2D predictions like Eq. 10, P-MPJPE reduces from 90.0 to 87.3. The reason is that the 2D detector exploits image cues to provide 2D estimations but rationality is ignored. Moreover, the pose filling network can adapt 2D inputs to the pose lifting network for better 3D estimation.

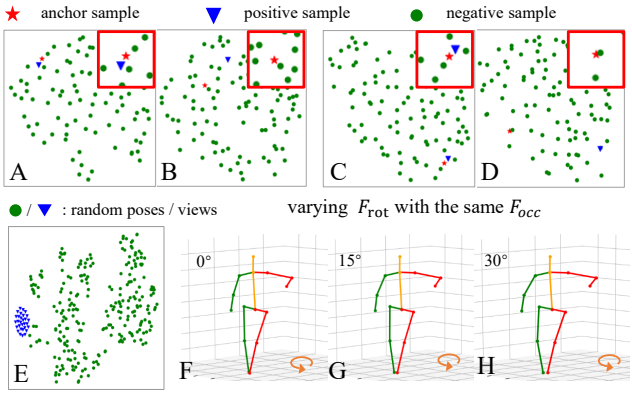
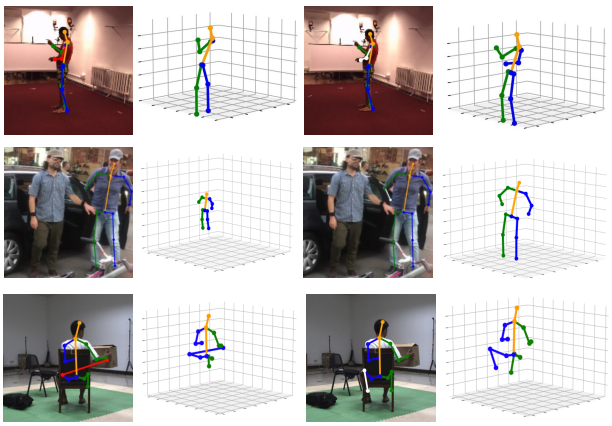


Figure 6: A-E: The t-SNE visualization of view-invariant features in the latent space. A/B correspond to w/ and w/o TiCLR. C/D indicate w/ and w/o VeCLR. E corresponds to 2D skeletons projected from the same 3D skeleton to different views (blue) and 2D skeletons projected from different 3D skeletons (green). F-H correspond to varying the view-dependent feature with the same view-invariant feature.



(a) Origin 2D (b) Yu (c) Refined 2D (d) Ours

Figure 7: Qualitative comparison of 2D completion and 3D estimation on Human3.6M, 3DPW and 3DOH50K. The 1st row and 2nd row show the simulated/real-world missing joint caused by occlusion. In the 3rd row, 2D detector provides an inaccurate estimation. Refined 2D (c) are produced by the proposed filling network. [38] (b) and ours (d) are the corresponding estimated 3D skeletons of two methods.

In particular, if there exist 2D joints with extremely low confidence (due to e.g. truncation, serious occlusion), we find it's better to mask out these key-points as missing ones and rely on the pose filling network to perform completion. We show some examples in Fig. 7. Quantitatively, we mask out the prediction from AlphaPose whose confidence is below 0.1 (i.e. set $c_{det} = 0$) and follow Eq. 10, P-MPJPE reduces from 87.3 to 86.8 on the test set of 3DOH50K.

4.6 Results in Unocclusion with Occluded Data

To verify that our approach can also generalize to scenarios where occlusion exists during training (i.e., there is no way to obtain complete counterparts), we simulate occluded training data on Human3.6M. Specifically, we randomly mask 10% 2D poses by 0-3 key-points at the beginning, and directly use the pose filling network to predict missing key-points. For TiCLR module, the positive samples come from a second-time random erasing since complete ones are unavailable. The performance on the same Human3.6M test set is 47.6, which is only a slight drop from 47.0 achieved with all training data unoccluded.

4.7 Module Analysis

Analysis on feature decoupling. a) To show the superiority of proposed decoupled representation, we experiment with a non-decoupled version in Tab. 3. The pure decoupled representation boosts the performance by ~ 1.5 points on MPJPE. Moreover, the coupled representation is unsuitable for view-equivariance optimization, while the decoupled version can further bring an improvement of ~ 2 points via leveraging two contrastive modules. **b)** To illustrate that we *do* decouple the view-invariant feature from the view-dependent part successfully, we make a visualization in the second row of Fig. 6. Firstly, we project one 3D skeleton to different 2D views and use the pose lifting network to extract the view-invariant features, which are visualized in blue color. Features extracted from different 3D skeletons' projections are visualized in green color. As expected, features from the same 3D skeletons but different views form a small cluster in the latent space and is well separated from those from different 3D poses. Secondly, we use the same view-invariant features and vary the view-dependent part(increase the rotation around y axis from 0 to 45 degrees), and then regress 3D skeletons. The results display little pose difference and gradual rotation around the y axis. We experimentally find that the network can't capture a significant change in the global feature(e.g. drift in the latent space). We attribute this to the intrinsic ambiguity of orientation regression from 2D poses only.

Analysis on contrastive learning modules. We quantitatively study the effectiveness of TiCLR and VeCLR modules. The results are reported in Tab. 3. With the introduction of VeCLR module, P-MPJPE reduces from 49.0 to 47.4. Moreover, the TiCLR module brings an improvement to 47.0. To verify that contrastive learning does help separate occluded samples from others from the anchor, we use t-SNE to visualize the features learnt by the pose lifting network in Fig. 6. Besides, the same skeleton, when projected to different views, produces closer features in latent space with the help of VeCLR module.

5 CONCLUSION

In this paper, we propose an occlusion-aware contrastive representation based scheme (OCR-Pose) consisting of TiCLR and VeCLR modules. Both modules optimize the corresponding representation via an iterative training strategy. Extensive experiments show that our model achieves the state-of-the-art performance on related human pose estimation datasets, and obtains the comparable performance under occlusion.

Acknowledgment This work was supported by National Science Foundation of China (U20B2072,61976137). The authors would like to give a personal thanks to the Student Innovation Center of SJTU for providing GPUs.

REFERENCES

[1] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. 2019. Unsupervised 3D Pose Estimation With Geometric Self-Supervision. In *CVPR*. IEEE, 5714–5724.

[2] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. 2020. Anatomy-aware 3D Human Pose Estimation in Videos. *CoRR* abs/2002.10322 (2020).

[3] Ting Chen, Simon Kornblith, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607.

[4] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby T. Tan. 2019. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. In *ICCV*. IEEE, 723–732.

[5] Dylan Drover, M. V. Rohith, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. 2018. Can 3D Pose Be Learned from 2D Projections Alone?. In *ECCV Workshops (Lecture Notes in Computer Science, Vol. 11132)*. Springer, 78–94.

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *ICCV*. IEEE, 2353–2362.

[7] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *CVPR*. IEEE, 1735–1742.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*. IEEE, 9726–9735.

[9] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015).

[10] Catalin Ionescu, Dragos Papava, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *TPAMI* 36, 7 (2014), 1325–1339.

[11] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. 2020. Weakly-Supervised 3D Human Pose Learning via Multi-View Images in the Wild. In *CVPR*. IEEE, 5242–5251.

[12] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-End Recovery of Human Shape and Pose. In *CVPR*. IEEE, 7122–7131.

[13] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. 2019. Unsupervised Keypoint Learning for Guiding Class-Conditional Video Prediction. In *NIPS*. 3809–3819.

[14] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. 2021. PARE: Part Attention Regressor for 3D Human Body Estimation. In *Proceedings International Conference on Computer Vision (ICCV)*. IEEE, 11127–11137.

[15] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. 2019. Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry. In *CVPR*. IEEE, 1077–1086.

[16] Jogendra Nath Kundu, Siddharth Seth, Mugalodi Rakesh, R. Venkatesh Babu, and Anirban Chakraborty. 2020. Self-Supervised 3D Human Pose Estimation via Part Guided Novel Image Synthesis. In *CVPR*. IEEE, 6151–6161.

[17] Jogendra Nath Kundu, Siddharth Seth, Rahul M. V., Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. 2020. Kinematic-Structure-Preserved Representation for Unsupervised 3D Human Pose Estimation. In *AAAI*. AAAI Press, 11312–11319.

[18] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021. Prototypical Contrastive Learning of Unsupervised Representations. In *ICLR*. OpenReview.net.

[19] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 2021. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. In *CVPR*. IEEE, 4741–4750.

[20] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. 2019. On Boosting Single-Frame 3D Human Pose Estimation via Monocular Videos. In *ICCV*. IEEE, 2192–2201.

[21] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *ICCV*. IEEE, 2659–2668.

[22] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In *3DV*. IEEE, 506–516.

[23] Rahul Mitra, Abhishek Sharma, and Arjun Jain. 2020. Multiview-Consistent Semi-Supervised Learning for 3D Human Pose Estimation. In *CVPR*. IEEE, 6906–6915.

[24] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised training. In *CVPR*. IEEE, 7753–7762.

[25] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. 2018. Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation. In *ECCV (Lecture Notes in Computer Science, Vol. 11214)*. Springer, 765–782.

[26] István Sárándi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. 2018. How Robust is 3D Human Pose Estimation to Occlusion? *CoRR* abs/1808.09316 (2018). arXiv:1808.09316 http://arxiv.org/abs/1808.09316

[27] István Sárándi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. 2018. Synthetic Occlusion Augmentation with Volumetric Heatmaps for the 2018 ECCV PoseTrack Challenge on 3D Human Pose Estimation. *CoRR* abs/1809.04987 (2018). arXiv:1809.04987 http://arxiv.org/abs/1809.04987

[28] István Sárándi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. 2021. MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation. *IEEE Trans. Biom. Behav. Identity Sci.* 3, 1 (2021), 16–30. https://doi.org/10.1109/TBIOM.2020.3037257

[29] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. 2021. Self-Supervised 3D Hand Pose Estimation From Monocular RGB via Contrastive Learning. In *ICCV*. IEEE, 11230–11239.

[30] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. 2017. Compositional Human Pose Regression. In *ICCV*. 2621–2630.

[31] Yu Sun, Qian Bao, Wu Liu, Yili Fu, and Tao Mei. 2020. CenterHMR: a Bottom-up Single-shot Method for Multi-person 3D Mesh Recovery from a Single Image. *CoRR* abs/2008.12272 (2020).

[32] Hsiao-Yu Fish Tung, Adam W. Harley, and Katerina Fragkiadaki. 2017. Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation from Unpaired Supervision. In *ICCV*. IEEE, 4364–4372.

[33] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018).

[34] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *ECCV (Lecture Notes in Computer Science, Vol. 11214)*. Springer, 614–631.

[35] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. 2021. Deep High-Resolution Representation Learning for Visual Recognition. *TPAMI* 43, 10 (2021), 3349–3364.

[36] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2020. Motion Guided 3D Pose Estimation from Videos. *CoRR* abs/2004.13985 (2020).

[37] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. 2020. Deep Kinematics Analysis for Monocular 3D Human Pose Estimation. In *CVPR*. IEEE, 896–905.

[38] Zhenbo Yu, Bingbing Ni, Jingwei Xu, Junjie Wang, Chenglong Zhao, and Wenjun Zhang. 2021. Towards Alleviating the Modeling Ambiguity of Unsupervised Monocular 3D Human Pose Estimation. In *ICCV*. IEEE, 8631.

[39] Zhenbo Yu, Junjie Wang, Jingwei Xu, Bingbing Ni, Chenglong Zhao, Minsi Wang, and Wenjun Zhang. 2021. Skeleton2Mesh: Kinematics Prior Injected Unsupervised Human Mesh Recovery. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 8599–8609. https://doi.org/10.1109/ICCV48922.2021.00850

[40] Tianshu Zhang, Buzhen Huang, and Yangang Wang. 2020. Object-Occluded Human Shape and Pose Estimation From a Single Color Image. In *CVPR*. IEEE, 7374–7383.

[41] Lu Zhou, Yingying Chen, Yunze Gao, Jinqiao Wang, and Hanqing Lu. 2020. Occlusion-Aware Siamese Network for Human Pose Estimation. In *ECCV (Lecture Notes in Computer Science, Vol. 12365)*. Springer, 396–412.

[42] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, and Kostas Daniilidis. 2019. MonoCap: Monocular Human Motion Capture using a CNN Coupled with a Geometric Prior. *TPAMI* 41, 4 (2019), 901–914.