Weakly Supervised 3D Multi-Person Pose Estimation for Large-Scale Scenes Based on Monocular Camera and Single LiDAR

Peishan Cong[1,†], **Yiteng Xu**[1,†], **Yiming Ren**[1], **Juze Zhang**[1],
Lan Xu[1,2], **Jingya Wang**[1,2], **Jingyi Yu**[1,2], **Yuexin Ma**[1,2,*]
[1]ShanghaiTech University
[2]Shanghai Engineering Research Center of Intelligent Vision and Imaging
{congpsh,xuyt1,mayuexin}@shanghaitech.edu.cn

Abstract

Depth estimation is usually ill-posed and ambiguous for monocular camera-based 3D multi-person pose estimation. Since LiDAR can capture accurate depth information in long-range scenes, it can benefit both the global localization of individuals and the 3D pose estimation by providing rich geometry features. Motivated by this, we propose a monocular camera and single LiDAR-based method for 3D multi-person pose estimation in large-scale scenes, which is easy to deploy and insensitive to light. Specifically, we design an effective fusion strategy to take advantage of multi-modal input data, including images and point cloud, and make full use of temporal information to guide the network to learn natural and coherent human motions. Without relying on any 3D pose annotations, our method exploits the inherent geometry constraints of point cloud for self-supervision and utilizes 2D keypoints on images for weak supervision. Extensive experiments on public datasets and our newly collected dataset demonstrate the superiority and generalization capability of our proposed method. https://github.com/4DVLab/FusionPose.git

Introduction

3D multi-person pose estimation (3D-MPE) in the wild, especially in large-scale outdoor scenes, has become an increasingly popular research field. It is an essential technique for human motion understanding, which can benefit many downstream real-world applications, including action recognition, sports analysis, surveillance, augmented/virtual reality(AR/VR), autonomous driving, assistive robots, etc. The goal is to localize semantic keypoints of human bodies in 3D space, namely the world coordinate system.

Most of previous works (Véges and Lőrincz 2019; Wang et al. 2020a) solve 3D-MPE based on the monocular camera, which is lightweight and convenient to be set up in general scenarios. However, the problem of depth estimation from monocular camera is ill-posed in essence (Mallot et al. 1991), causing many ambiguous predictions in global localization and local pose estimation, as Figure. 2 shows. Although researchers have proposed plenty of approaches to alleviate the problem, such as using geometric constraints by the prior knowledge of the height or bone length of human body (Zhang et al. 2022b), hybrid inverse kinematic constraints (Sun et al. 2021; Li et al. 2021), and motion consistency constraints existing in videos (Zhang et al. 2022a). These methods still perform limited due to the mathematically impossible mapping from the perspective view to 3D space. Although the settings of multi-view cameras (Dong et al. 2019; Zhang et al. 2021) and RGB-D cameras (Zimmermann et al. 2018) are proposed to escape from the trouble, they are not applicable for large-scale scenes due to the deployment difficulties or the physical limitations of sensors (RGB-D camera is available in about 5 meters and usually fail in outdoor scenes.).

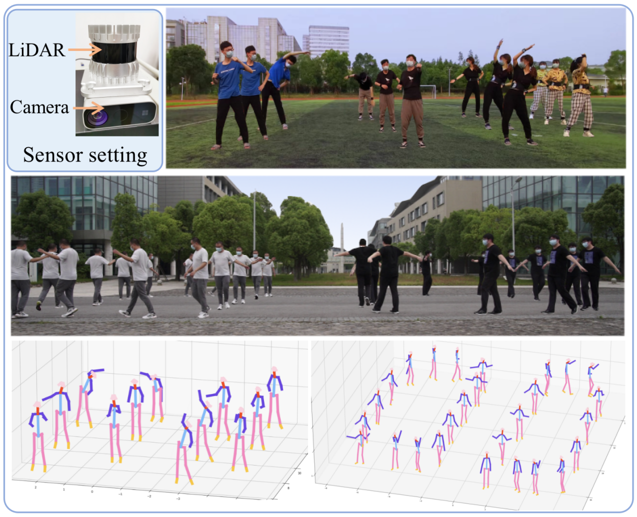


Figure 1: Our sensor setting and 3D multi-person pose estimation results for large-scale scenes. Based on synchronized images and LiDAR point clouds, we can capture continuous global locations and local poses accurately for each person.

LiDAR can provide accurate depth information and has been widely-used on autonomous vehicles and robots to perceive large-scale scenes. The effective range for capturing human with recognizable shape and scale could reach about 35m by common 128-beam mechanical LiDAR, making it feasible for 3D-MPE in long-range indoor or outdoor scenes. More importantly, unlike the sensitivity of camera to light, LiDAR could work day and night, which is appli-

[*]Corresponding author, † Equal contribution

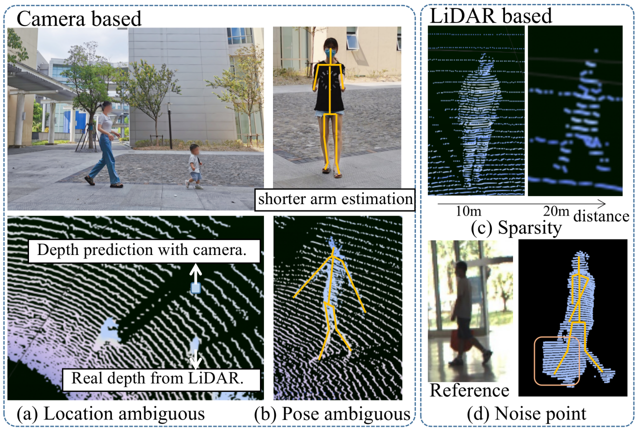


Figure 2: The limitation of 3D-MPE based on monocular camera or single LiDAR. (a) shows the ambiguity of global instance localization caused by the usage of statistics of human body in monocular camera-based methods, which has weak generalization capability for diverse persons. (b) shows the ambiguity of local poses due to the perspective view of image. (c) and (d) illustrate the limitations based on single LiDAR due to the sparse points in the distance and the noise point brought by carry-on objects. FusionPose takes advantage of both sensors to overcome above limitations.

cable for most scenarios. Recently, some researchers begin to exploit the LiDAR sensor in human motion capture (Li et al. 2022a; Zhao et al. 2022) and has achieved impressive performance. However, the point cloud captured by LiDAR is sparsity-varying along the distance with noise, leading to unstable pose predictions (Figure. 2). Considering the rich appearance features in images and geometry features in point clouds, (Fürst et al. 2021; Zheng et al. 2022) utilize both camera and LiDAR to estimate 3D poses of pedestrians to enhance scene understanding for autonomous vehicles. However, they fuse multi-modal features in a straight forward way and do not employ the temporal information and geometric supervision, leading to unsatisfactory performance on challenging poses, like doing sports.

We focus on solving 3D-MPE in large-scale scenes based on multi-modal sensors, including monocular camera and single LiDAR. There are two main challenges to overcome. First, the image captured by camera is dense and regular representation and contains the texture feature in the perspective view, while point cloud captured by LiDAR is sparse and unordered representation and provides the depth feature in 3D space. How to taking advantage of multi-modal data from totally different sensors and fuse them in an effective and interpretable way for accurate pose regression is one critical problem. Second, current deep learning-based 3D-MPE methods rely on huge annotated data, which is usually obtained by wearable IMUs devices or multi-view cameras. However, they are not applicable in large-scale scenes due to the drift problem in long distance of IMUs and difficult deployment of multi-view cameras. Manual annotation is expensive and time-consuming. How to conduct 3D-MPE in large-scale scenes without 3D annotations is the other core problem.

In this paper, we propose **FusionPose**, a novel 3D-MPE approach for large-scale scenes based on the single-LiDAR-camera setting, which has solved above problems. To fully utilize the global semantic feature in images and local geometric feature in point clouds, we present an effective Image-to-Point Attention Fusion (IPAFusion) method to fuse 2D and 3D information. Cross-attention is designed between two modalities to make the network learn the physical correspondence automatically, which can alleviate the dependence on accurate calibration of two sensors and make the fusion process effective and interpretable. To overcome the rely on 3D annotations, we take the best advantage of the self-supervision of the data, including the dynamic motion constraints and high-dimension feature consistency existing in consecutive frames of data, and the geometric constraints of human body points. We also use 2D keypoints generated by mature 2D pose estimation methods to further supervise the estimated 3D keypoints by back projection to image. To facilitate the 3D-MPE research on the multi-modal setting, we collected a new dataset, LiCamPose, in the wild. Extensive experiments show that our method achieves state-of-the-art performance on LiCamPose and other related open datasets. Main contributions of this paper are as follows:

1. Taking advantage of both LiDAR and camera sensors, we propose a novel method for multi-person 3D pose estimation in large-scale scenes with accurate localization. Specifically, our method is independent of 3D pose annotations.
2. We propose an IPAFusion method to fuse the information of 2D perspective-view images and 3D point cloud, which fully considers global semantic feature and local geometric feature of multimodal data and is free for calibration errors.
3. We exploit the motion cues and sequential consistency existing in temporal information to enhance the 3D pose estimation.
4. FusionPose achieves state-of-the-art performance on 3D pose datasets, including HybirdCap, 3DPW, STCrowd, and our new collected dataset, LiCamPose. We will release our novel data when the paper is published.

Related Work

Camera-based 3D Human Pose Estimation

Extensive methods have been proposed for 3D-MPE based on monocular camera. Early works focus on human-centric tasks without localizing individuals in the actual 3D space. (Pavlakos et al. 2017) directly regresses the joint positions from input images and (Tome, Russell, and Agapito 2017; Martinez et al. 2017; Rogez, Weinzaepfel, and Schmid 2019) feed the 2D keypoints into a 2D-to-3D lifting network to estimate 3D poses. To facilitate more real-world applications, researchers pay more attention to the camera-centric 3D-MPE recently. They (Moon, Chang, and Lee 2019; Véges and Lőrincz 2019; Wang et al. 2020a) usually decouple the problem into the root-relative 2.5D pose estimation and root depth estimation. However, the accurate depth estimation no matter for local keypoints or for objects is the core challenge for 3D-MPE. To address it, some works (Mehta, Sotnychenko, and etc. 2018; Mehta et al. 2020; Zhen et al. 2020; Zhang et al. 2022b) make use of geometry constraints by adding prior knowledge of the human body, such as the height or bone length, in the depth reasoning. Based on handcraft assumptions, such methods eliminate many poor results of 3D-MPE but become limited for the scenes with diverse people. Some other methods take advantage of hybrid inverse kinematics of motions (Sun et al. 2021; Li et al. 2021; Sun et al. 2022; Yu et al. 2021) by using SMPL (Loper et al. 2015) parametric human model or explore spatial and temporal relationships by enforcing temporal consistency across consecutive frames (Arnab, Doersch, and Zisserman 2019; Cheng et al. 2020; Zheng et al. 2021; Zhang et al. 2022a). However, the ambiguous depth estimation still exist for the monocular camera setting. Although the multi-camera (Dong et al. 2019; Zhang et al. 2021; Rhodin, Salzmann, and Fua 2018; Chen et al. 2019; Kocabas, Karagoz, and Akbas 2019; Wandt et al. 2021) and RGB-D (Mehta, Sridhar, and etc. 2017; Zimmermann et al. 2018; Ying and Zhao 2021) settings can, to some extend, alleviate the problem, they are not applicable for the large-scale outdoor scenes.

LiDAR-involved 3D Human Pose Estimation

LiDARs become more and more popular in 3d scene understanding (Cong et al. 2022; Zhu et al. 2020, 2021; Yin, Zhou, and Krähenbühl 2021; Han et al. 2022) due to its accurate measurement for the depth information in large-scale scenes, which has boosted the progress of autonomous driving and robotics. Recently, researchers begin to explore the potential usage of LiDAR in fine-grained human motion capture (Li et al. 2022a; Zhao et al. 2022) and has made impressive achievements especially for the long-range scenarios. However, LiDAR point cloud has sparse and unordered representation without much texture feature, which usually leads to unstable pose estimations with noise points caused by carry-on objects or clothes. To enhance the perception and understanding for pedestrians in traffic scenarios, (Fürst et al. 2021; Zheng et al. 2022) propose to use both camera and LiDAR to predict the 3D poses of pedestrians. However, they only rely on the 2D keypoint supervision or coarse 3D pseudo labels without considering temporal features, resulting in unsatisfactory results for more complicated actions. Our method leverages the comprehensive feature from images and point clouds, and motion cues in sequences to achieve more robust and accurate pose estimations in more general scenes.

Sensor-fusion Approaches for LiDAR and Camera

There are already many researches about LiDAR-camera-based sensor fusion methods for autonomous driving, which can be classified into three main categories. The first one is point-level fusion strategy (Vora et al. 2020; Wang et al. 2021; Zheng et al. 2022), which attaches the semantic feature extracted from the corresponding area of image to point, followed by a point cloud-based feature extractor. However, these hard-association methods reply heavily on the sensor calibration and will lose global context information of images. The second one is feature-level fusion strategy (Piergiovanni et al. 2021; Chen et al. 2017; Liang et al. 2018; Ku et al. 2018) by directing concatenating features from two modalities, which considers the fusion of global context but lacks local geometric corresponding. The third one (Bai et al. 2022; Li et al. 2022c; Prakash, Chitta, and Geiger 2021; Liu et al. 2022) utilizes transformer strategy by constructing queries in BEV space, which dynamically capture the correlations between image and LiDAR features. Such methods work well in detection and segmentation tasks by fusing features in BEV while ignoring fine-grained 3D postures, making them inapplicable for 3D pose estimation tasks. The only two related LiDAR-camera-based 3D-MPE methods (Fürst et al. 2021; Zheng et al. 2022) directly adopt above point-level and feature-level fusion strategies without specific design for 3D-MPE. In view of the fine-grained feature requirement of 3D-MPE, we propose a soft-association method based on the cross-attention mechanism, which fuse the local geometric features of point cloud with the global context feature of images in an effective manner.

Method

Problem Definition Given the synchronized image I and point cloud P captured by monocular RGB camera and single LiDAR, our task is to predict the 3D poses $\hat{J}_{3D} \in R^{K\times 3}$ for multiple people in the real world, where K denotes the number of keypoints of the 3D pose representation. Because all sensors are fixed during data capture, the LiDAR coordinate system equals to the world coordinate system, and $\hat{J}_{3D}$ can be projected to image through the intrinsic and extrinsic parameters of sensors.

Overview Our method is a top-down 3D-MPE method by first detecting persons and then estimating the 3D pose for each person according to the cropped image and point cloud. The whole pipeline of our method is illustrated in Figure 3, which contains two important components, including the Image-to-Point Attention Fusion (IPAFusion) module and Temporal Information Guided Pose Estimator. The former fuses the information of two distinct modalities of data to fully use the 3D geometry features of point cloud and the appearance features of images. The latter leverages temporal guidance existing in consecutive data to improve the pose accuracy by learning the dynamic rules of human motions and the pose consistency in high-dimension feature space. Furthermore, we utilize the raw point cloud to supervise the shape and scale of $\hat{J}_{3D}$ and 2D keypoint to weakly supervise the pose by back projection. In the following, we will introduce more details for above modules and losses.

Pre-processing

For 2D keypoints $J_{2D} \in R^{K\times 2}$ on images used for supervision, we generate from openpose(Cao et al. 2017). For the first-stage detection, we utilize the sate-of-the-art LiDAR-based 3D detector (Cong et al. 2022) and image-based 2D detector (YOLO v5) to process the input and obtain the paired persons in two modalities by projection and matching according to calibration matrix. Then we crop the images and point clouds by 2D and 3D bounding boxes for the following 3D pose estimation.

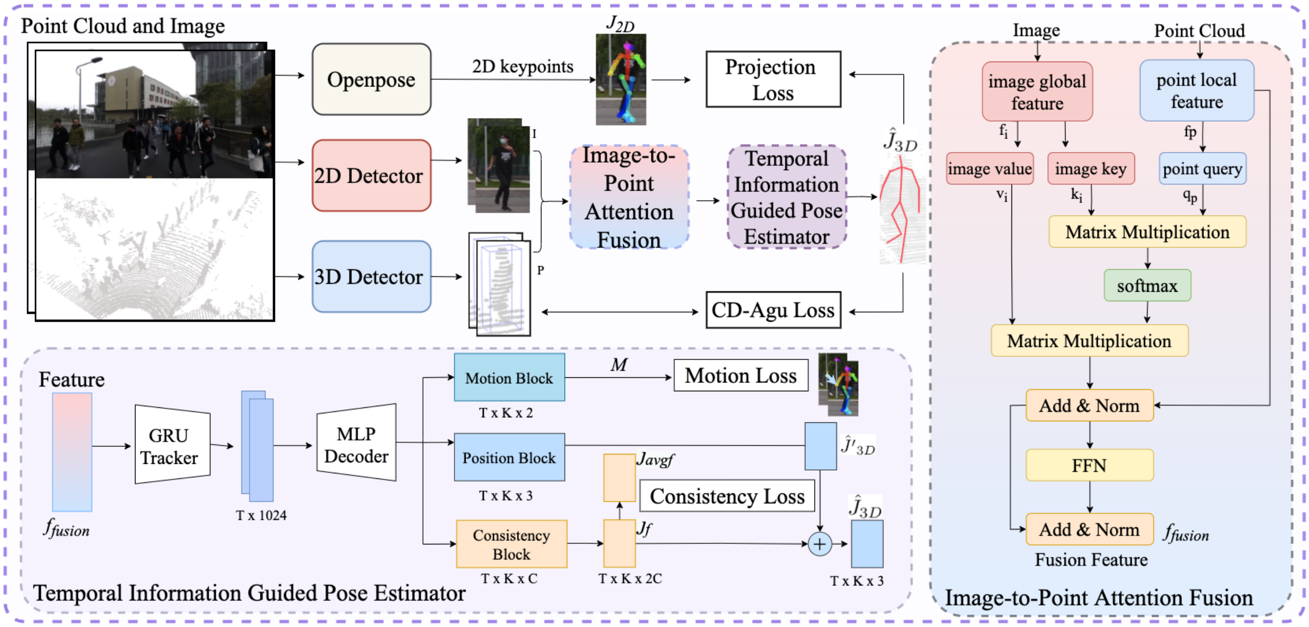


Figure 3: Pipeline of FusionPose. We first obtain cropped images and point clouds of each person by 2D and 3D detectors. Then the features extracted from multi-modal data are fed in the Image-to-Point Attention Fusion module to get the fused feature with rich texture and geometry information. Temporal Information Guided Pose Estimator is followed to estimate 3D poses by leveraging the temporal guidance. Finally, the raw point cloud and 2D keypoints are used for supervision by shape constraints.

Image-to-Point Attention Fusion

Previous fusion methods for LiDAR point clouds and images are designed for detection and segmentation tasks and are applied for autonomous driving. Different from them, 3D pose estimation requires us to pay attention to the fine-grained semantic and geometry features of human bodies. Thus, we propose an effective fusion method for 3D-MPE task, which can automatically learn the corresponding features between images and point cloud to eliminate the sensitivity to sensor calibrations and fully take advantage of global and local information of two modalities.

Point Cloud Feature Extraction The low dimensional point cloud input after downsampling $P \in R^{N\times3}$ ($N = 256$ denotes the number of the points) are fed into the PointNet (Qi et al. 2017) encoder to obtain the high-dimensional feature $p = PointNetEncoder(P), p \in R^{N\times256}$, then we use one layer of self-attention to integrate the global context feature to each point feature:

$$f_p = LN(p + SelfAttention(p)),$$

where LN is layer normalization.

Image Feature Extraction We use the pretrained model of HrNet (Wang et al. 2020b) to extract image features, which maintains multi-level resolutions of features and fine-grained local semantic features. The image input I is encoded into high dimensional feature of size $(256, H/8, W/8)$, where H and W represent the size of input. We flatten the spatial feature and integrate the channel information through Multi-Layer Perception (MLP) to get high-level semantic features: $i = MLP(Flatten(HrNet(I)))$. One layer of self-attention is also applied to involve the information from global context.

$$f_i = LN(i + SelfAttention(i)).$$

Cross-attention Fusion Fusing two modal features by direct projection relies heavily on the accurate calibrations of sensors and constrains the correspondence by totally physical mapping. Considering that different parts of the texture feature of images are not equally important to each point of the human body, we design IPAFusion to learn the correspondences between images and point cloud automatically by network, which can fuse features more reasonably and is calibration-free. The ***point query*** q_p are extracted from high-dimension point feature f_p, the ***image value*** v_i and the ***image key*** k_i are extracted from global semantic image feature f_i. For each query, it conducts a dot product with the image key to get the attention matrix and obtain the correlation from multi-model features. The higher value after the dot product indicates that the point cloud is highly correlated with the corresponding part of the image feature. After softmax normalization, the attention affinity matrix will be multiplied by image value to obtain new point cloud features weighted by the image information. The weighted point cloud features are then connected with the original point query and pass through two linear layers to obtain $f_{attention}$. The final fusion features f_{fusion} is acquired through FFN:

$$f_{attention} = LN(f_p + CrossAttention(q_p, k_i, v_i)),$$
$$f_{fusion} = LN(f_{attention} + FFN(f_{attention})).$$

By this way, IPAFusion can not only automatically learn the correspondences to fuse features of two modalities, but also fully use the global semantic information and local fine-grained geometric feature to boost accurate pose estimation.

Temporal Information Guided Pose Estimator

Human motions are changing continuously with each part of the body moving under specific dynamic constraints. Our method can learn the motion cues in sequential input data by the **Motion Block** to guide the estimation of more reasonable continuous poses, especially for the occlusion situations, where it is difficult to predict the pose only based on the current frame of data but can be inferred by adjacent poses. Meanwhile, the feature expression of the same keypoint in high dimensional semantic space should be similar, e.g. in high-level feature space, hands even in different frames should keep close and the body center should be far apart from the limbs. We consider the feature consistency in consecutive frames in the **Consistency Block**.

Figure. 3 shows the detailed operations of the temporal information guided pose estimator. First, the fusion features $f^t_{fusion}, t \in T$ of T consecutive frames are fed into bi-GRU tracker to extract temporal features. Then, the MLP decoders are followed to predict three different properties of K keypoints, including the motion map $\hat{M}^t \in R^{K\times 2}$ in the motion block, 3D positions in LiDAR coordinate system $\hat{J'}^t_{3D} \in R^{K\times 3}$ in the position block, and high-dimension features $\hat{J_f}^t \in R^{K\times C}$ in the consistency block, respectively.

The motion map M^t is calculated the by the difference between each keypoint position at the previous frame and current frame on the image pixel coordinates: $M^t = J^t_{2D} - J^{t-1}_{2D}$, the motion prediction is supervised by:

$$L_{motion} = \frac{1}{K}\sum_{j=1}^{K}\left\|\hat{M}^t_j - M^t_j\right\|,$$

so that the motion block can use the dynamic constraints to assist in more accurate pose estimation.

The consistency block expands the 3D keypoint positions into higher-level space by two extra layers MLP and calculates the temporal consistency loss $L_{consistency}$ to pull the feature $\hat{J}^t_f$ of each keypoint to its average feature $J^t_{avgf} = \frac{1}{T}\sum^T_{t=1} J^t_f$ cross multiple frames:

$$L_{consistency} = \frac{1}{K}\sum_{j=1}^{K}\left\|\hat{J}^t_{f_j} - J^t_{avgf_j}\right\|.$$

And then, the feature J^t_f is concatenated with the $\hat{J'}^t_{3D}$ and gets the final keypoints $\hat{J}^t_{3D}$ with function $\mathcal{F}$:

$$\hat{J}^t_{3D} = \mathcal{F}(J^t_f, \hat{J'}^t_{3D}).$$

Weakly Unsupervised Training

2D pose estimation from images has achieved great progress due to huge labeled training data. With the help of 2D poses J_{2D} automatically generated by algorithms, we can supervise $\hat{J}^t_{3D}$ by projecting to images according to the transform matrix $\mathcal{T}$. The projected result is represented as $\hat{J}^t_{2D} = \mathcal{T}(\hat{J}^t_{3D})$. The projection loss is defined as:

$$L_{proj} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{T}(\hat{J}^t_{3D}) - J^t_{2D}\right\|.$$

In addition, raw point cloud reflect the real shapes, scales, and postures of human body. An accurate estimated 3D pose ought to fit the captured point cloud well. We adopt the Chamfer Distance (CD) (Fan, Su, and Guibas 2017) to measure the similarity between 3D pose and the point cloud.

$$L_{CD}(P^t, \hat{J}^t_{3D}) = \frac{1}{\|P^t\|}\sum_{x\in P^t}\min_{y\in \hat{J}^t_{3D}}\|x-y\|^2_2$$
$$+\frac{1}{\|\hat{J}^t_{3D}\|}\sum_{y\in \hat{J}^t}\min_{x\in P^t}\|y-x\|^2_2,$$

where P^t denotes the point cloud, x and y represent the 3D coordinates of points. Only K keypoints is not comparable for N points numerically and geometrically, we further apply linear interpolation on $\hat{J}^t_{3D}$ and calculate the CD_agu Loss $L_{CD\_agu}(P^t, \hat{J}^t_{agu})$ where $\hat{J}^t_{agu}$ is the augmented keypoints. Then, our network can be trained by the loss L in self-supervised and weak-supervised manner as below:

$$L = \lambda_1 L_{motion} + \lambda_2 L_{consistency} + \lambda_3 L_{proj} + \lambda_4 L_{CD\_agu},$$

where λ are hyper-parameters.

Implementation Details

We implement our network using Pytorch 1.10.1 with CUDA 11.3. The point cloud branch is pretrained with simulated data as (Zhao et al. 2022) and the image branch utilizes the pretrained feature from HRNet (Wang et al. 2020b). The batch size is 8. K is set as 21 as (Cao et al. 2017) and T is set as 4 for continuous input frames.

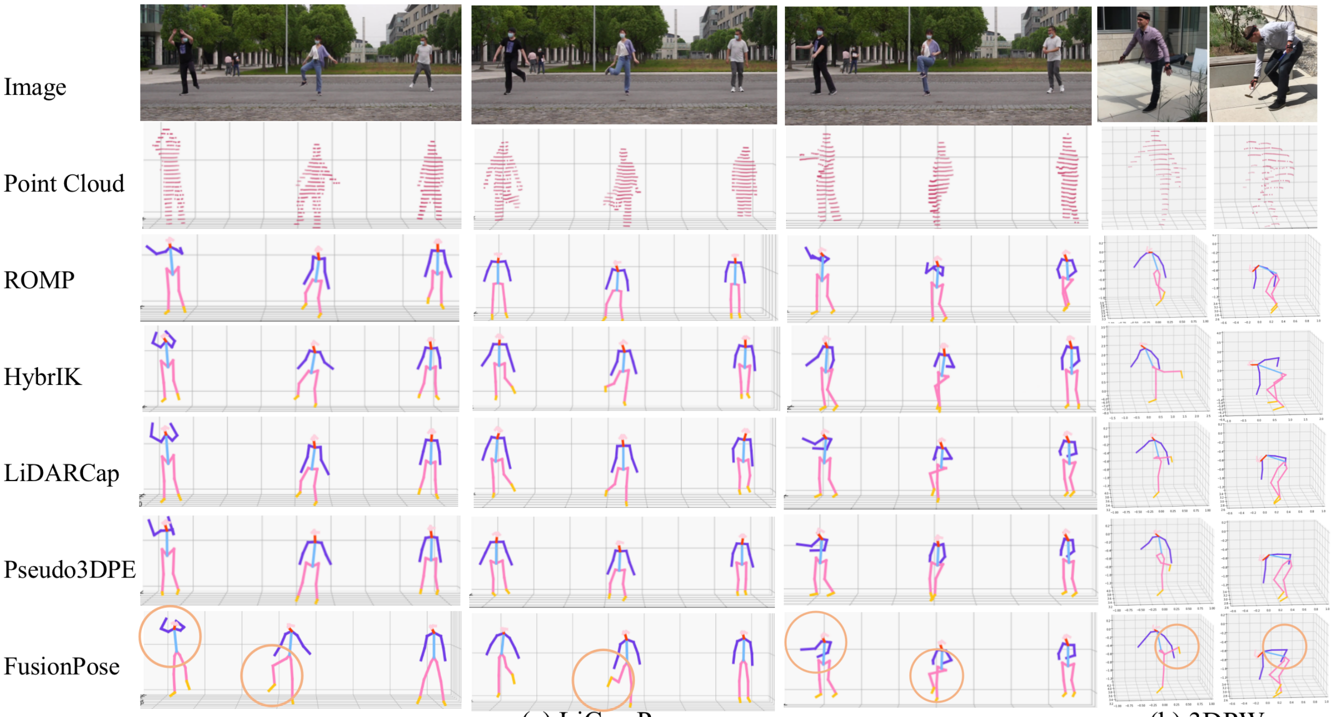
Experiment

We first introduce all datasets and evaluation metrics and then compare our method with current SOTA methods qualitatively and quantitatively. Extensive ablation studies are conducted for comprehensive assessment of FusionPose.

Dataset

LiCamPose is our new collected 3D-MPE dataset in long-range wild scenes with a 128-beam OuSTER-1 LiDAR and a camera, with totally 8,980 frames of synchronized multi-modal data. The ground truth is captured by Noitom Perception Neuron Studio(Noitom PN S). We divide the data half for the training set and half for the testing set. In addition, we collected extra 38,490 frames of data in the same setting but without pose annotations. These data is helpful for unsupervised methods to pretrain their models or validate the performance by visualization. LiCamPose contains various motions, including walking, running, doing KEEP, ball games, dancing, and Taekwondo in different scenes. We protect personal privacy by blurring faces in RGB images.

3DPW(Von Marcard et al. 2018) and **Hybridcap** (Liang et al. 2022) provide images and annotated SMPL models, where 3D poses can be acquired directly, but lack LiDAR point clouds. We follow (Cong, Zhu, and Ma 2021) to simulate the LiDAR point cloud data in a reasonable manner. Due to the simulation limitation, we follow the official protocol of dataset splitting and select valid data for experiments.



(a) LiCamPose (b) 3DPW

Figure 4: Visualization for local 3D pose results predicted by various methods on HybridCap, LiCamPose and 3DPW. We highlight some parts of estimated 3D poses by circles for detailed comparison.

		HybridCap			3DPW			LiCamPose		
	Sensor	PCK↑	MPJPE↓	CD↓	PCK↑	MPJPE↓	CD↓	PCK↑	MPJPE↓	CD↓
ROMP	Camera	45.3	187.9	-	53.9	160.6	-	53.2	159.9	-
HybrIK	Camera	75.8	113.4	-	80.2	107.1	-	73.0	109.2	-
LidarCap	LiDAR	86.5	88.1	21.2	70.7	119.2	28.6	75.8	119.9	26.8
Pseudo3DPE	LiDAR+Camera	70.7	130.8	28.1	73.5	116.4	17.6	68.9	124.3	**19.9**
FusionPose	LiDAR+Camera	**95.9**	**75.3**	**17.4**	**83.5**	**97.7**	**13.8**	**79.7**	**106.8**	21.7
FusionPose*	LiDAR+Camera	95.3	75.9	17.3	91.3	79.2	27.5	93.9	75.8	19.5

Table 1: Comparison results on HybridCap, 3DPW, and LiCamPose. * means fully supervised training mechanism.

STCrowd (Cong et al. 2022) is a pedestrian perception dataset with synchronized LiDAR point clouds and camera images, while it doesn't provide the 3D keypoints ground truth. Thus, we only provide the visualization results on it.

Evaluation Metric

Since LiDAR can provide accurate depth, we do not compare the depth estimations and only evaluate local poses. We use 1) **PCK↑**: percentage of correct keypoints that the normalized distance between the key point and its groundtruth is less than the set threshold (150mm) position error in millimeters; 2) **MPJPE↓**: mean per root-relative joint position error in millimeter; 3) **CD↓**: the chamfer distance between predict keypoints and raw point cloud in millimeter.

Performance Analysis

We compare with three kinds of SOTA methods for 3D-MPE, including monocular camera-based ROMP (Sun et al. 2021) and HybrIK (Li et al. 2021), LiDAR-based LidarCap (Li et al. 2022b), and LiDAR-camera multi-sensor-based Pseudo3DPE (Zhang et al. 2022b). We run their released code with provided parameters. The results on HybridCap, 3DPW, and LiCamPose are shown in Table **??**.

The camera-based methods are pretrained on MSCOCO, Human3.6M(Ionescu et al. 2013) and MPI-INF-3DHP (Mehta et al. 2017) in a full-supervision manner and then directly infer on the test data of these datasets. For LiDAR-based method, LiDARCap, we pretrain it on LiDARCap dataset with 3D annotations and show results by finetuning using our self-supervision losses. Pseudo3DPE and our method are weakly supervised methods with only 2D pose annotations. Compared with Pseudo3DPE, we get large improvement, illustrating the efficiency and generalization of our feature-fusion method and loss designs. Compared with camera-only and LiDAR-only methods, FusionPose has more accurate pose estimation even without any super-

	HybridCap			3DPW		
	PCK↑	MPJPE↓	CD↓	PCK↑	MPJPE↓	CD↓
Point-RGB	78.1	112.6	21.36	73.6	110.7	15.39
PixelFusion	80.3	107.8	20.05	75.6	106.9	14.91
LocalFusion	74.7	118.2	22.85	75.3	108.7	15.96
GlobalFusion	71.1	126.4	26.57	78.6	104.0	15.71
IPAFusion	**90.7**	**89.5**	**19.70**	**79.2**	**103.4**	**14.49**

Table 2: Ablation experiment for different fusion methods.

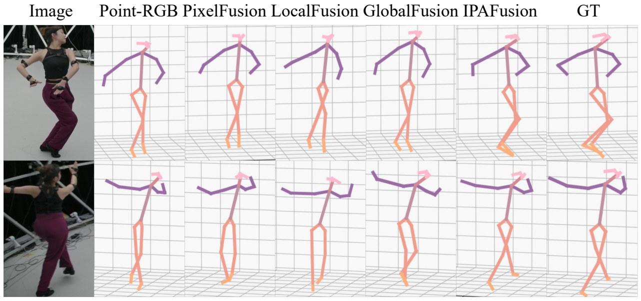


Figure 5: Comparison of different fusion methods on HybridCap Dataset. The last column is the ground truth.

vised pretrain stage. The visualization results are illustrated in Figure 4. Since camera-based methods are not good at global localization, we drag the 3D pose to the global position offered by LiDAR for local posture comparison. Due to the depth ambiguous, camera-based methods have poor performance on anterior and posterior amplitude of the limbs. While the LiDAR-based method is affected by the noise of the point cloud and generate rough local postures in some cases. Pseudo3DPE utilizes pseudo 3D labels projected by 2D points, which is not accurate enough and easy affected by calibration errors. Our method have superior performance on both local pose and depth estimation. In particular, we can see that our performance by weakly-supervised training is comparable to that by fully-supervised training, which further demonstrates the effectiveness of FusionPose.

Ablation Study

In this section, we validate the effectiveness of our sensor-fusion module and loss functions designed in our method.

Ablation Study for IPAFusion: We demonstrate the superiority of our IPAFusion model by replacing it with other fusion methods. We conduct the comparison on the basic network of FusionPose without extra temporal and CD supervision. Commonly used fusion strategies for images and point clouds in the perception area are as follows:

Point-RGB appends the raw representation of LiDAR point with corresponding RGB color according to calibration matrix. **PixelFusion** adds a k-dimensional image feature vector as a supplementary feature for each LiDAR point by projection. **LocalFusion** concatenates k-dimensional image feature vector to the corresponding high dimensional point feature for each point. Above three methods will get a feature-enhanced point cloud and then pass a point cloud-based backbone for further feature extraction. Their performances are sensitive to the sensor calibration, which is not stable in outdoor scenes. **GlobalFusion** directly concatenates the global image features with the global point cloud features. It is a totally high-level fusion strategy. Such method lacks actual mapping from two modal data, which downshifts the learning process and is not feasible for fine-grained pose tasks. **IPAFusion** integrates local high-dimensional geometric features of point cloud with global appearance features of images and automatically learn the projection by the cross-attention mechanism, which is calibration-free and maintain more detailed features. Table 2 and Figure 5 show that IPAFusion is significantly better than other fusion methods.

	PCK↑	MPJPE↓	CD↓
IPA	90.7	89.5	19.7
IPA+CB	93.6	82.3	18.8
IPA+CB+MB	95.4	77.0	17.6
IPA+CB+MB+CDA	**95.9**	**75.3**	**17.4**

Table 3: Ablation experiment for different components of FusionPose on HybridCap. IPA is the original IPAFusion baseline. CB and MB are consistency block and motion block, respectively, and CDA means CD_Agu optimization.

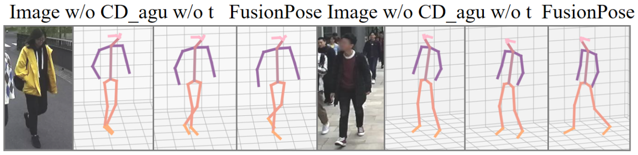


Figure 6: Ablation results on STCrowd. w/o CD_agu means eliminating the CD_agu loss and w/o t denotes eliminating temporal supervision (motion loss and consistency loss).

Ablation Study for Loss functions: We verify the loss functions of our method quantitatively and qualitatively, as Table. 3 and Figure. 6 shows. With the optimization of CD_agu loss with geometry constraints and temporal information guided model with dynamic constraints, the performance of FusionPose gets improved.

Conclusion

We propose a new 3D-MPE method for large-scale scenes based on single LiDAR and monocular camera. To fully use the appearance features of images and geometry features of LiDAR point clouds, we propose an effective sensor-fusion method to extract rich and fine-grained local pose features. In particular, our method does not require any 3D annotation by using motion cues and geometry constraints. Extensive experiments show our method achieves state-of-the-art performance on new collected dataset and open datasets.

Acknowledgements

This work was supported by NSFC (No.62206173, No.61976138), the National Key Research and Development Program (2018YFB2100500), STCSM (2015F0203-000-06), SHMEC (2019-01-07-00-01-E00003), Shanghai Sailing Program (No.22YF1428700, No.21YF1429400), and Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

Arnab, A.; Doersch, C.; and Zisserman, A. 2019. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 3395–3404.

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1090–1099.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 7291–7299.

Chen, C.-H.; Tyagi, A.; Agrawal, A.; Drover, D.; Mv, R.; Stojanov, S.; and Rehg, J. M. 2019. Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*, 5714–5724.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3D Object Detection Network for Autonomous Driving. 6526–6534.

Cheng, Y.; Yang, B.; Wang, B.; and Tan, R. T. 2020. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10631–10638.

Cong, P.; Zhu, X.; and Ma, Y. 2021. Input-output balanced framework for long-tailed lidar semantic segmentation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Cong, P.; Zhu, X.; Qiao, F.; Ren, Y.; Peng, X.; Hou, Y.; Xu, L.; Yang, R.; Manocha, D.; and Ma, Y. 2022. STCrowd: A Multimodal Dataset for Pedestrian Perception in Crowded Scenes. In *CVPR*, 19608–19617.

Dong, J.; Jiang, W.; Huang, Q.; Bao, H.; and Zhou, X. 2019. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7792–7801.

Fan, H.; Su, H.; and Guibas, L. J. 2017. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *CVPR*.

Fürst, M.; Gupta, S. T.; Schuster, R.; Wasenmüller, O.; and Stricker, D. 2021. HPERL: 3d human pose estimation from RGB and lidar. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 7321–7327. IEEE.

Han, X.; Cong, P.; Xu, L.; Wang, J.; Yu, J.; and Ma, Y. 2022. LiCamGait: Gait Recognition in the Wild by Using LiDAR and Camera Multi-modal Visual Sensors. *arXiv preprint arXiv:2211.12371.*

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Kocabas, M.; Karagoz, S.; and Akbas, E. 2019. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1077–1086.

Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; and Waslander, S. L. 2018. Joint 3d proposal generation and object detection from view aggregation. In *IROS*, 1–8. IEEE.

Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 3383–3393.

Li, J.; Zhang, J.; Wang, Z.; Shen, S.; Wen, C.; Ma, Y.; Xu, L.; Yu, J.; and Wang, C. 2022a. LiDARCap: Long-range Marker-less 3D Human Motion Capture with LiDAR Point Clouds. *ArXiv*, abs/2203.14698.

Li, J.; Zhang, J.; Wang, Z.; Shen, S.; Wen, C.; Ma, Y.; Xu, L.; Yu, J.; and Wang, C. 2022b. LiDARCap: Long-range Marker-less 3D Human Motion Capture with LiDAR Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20502–20512.

Li, Y.; Yu, A. W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Lu, Y.; Zhou, D.; Le, Q. V.; et al. 2022c. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17182–17191.

Liang, H.; He, Y.; Zhao, C.; Li, M.; Wang, J.; Yu, J.; and Xu, L. 2022. HybridCap: Inertia-aid Monocular Capture of Challenging Human Motions. *arXiv preprint arXiv:2203.09287.*

Liang, M.; Yang, B.; Wang, S.; and Urtasun, R. 2018. Deep continuous fusion for multi-sensor 3d object detection. In *ECCV*, 641–656.

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

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *TOG*, 34(6): 1–16.

Mallot, H. A.; Bülthoff, H. H.; Little, J.; and Bohrer, S. 1991. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological cybernetics*, 64(3): 177–185.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2640–2649.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, 506–516. IEEE.

Mehta, D.; Sotnychenko, O.; and etc. 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, 120–130. IEEE.

Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Elgharib, M.; Fua, P.; Seidel, H.-P.; Rhodin, H.; Pons-Moll, G.; and Theobalt, C. 2020. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *TOG*, 39(4): 82–1.

Mehta, D.; Sridhar, S.; and etc. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *TOG*, 36(4): 1–14.

Moon, G.; Chang, J. Y.; and Lee, K. M. 2019. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 10133–10142.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 7025–7034.

Piergiovanni, A.; Casser, V.; Ryoo, M. S.; and Angelova, A. 2021. 4d-net for learned multi-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15435–15445.

Prakash, A.; Chitta, K.; and Geiger, A. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7077–7087.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.

Rhodin, H.; Salzmann, M.; and Fua, P. 2018. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, 750–767.

Rogez, G.; Weinzaepfel, P.; and Schmid, C. 2019. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *TPAMI*, 42(5): 1146–1161.

Sun, Y.; Bao, Q.; Liu, W.; Fu, Y.; Black, M. J.; and Mei, T. 2021. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 11179–11188.

Sun, Y.; Liu, W.; Bao, Q.; Fu, Y.; Mei, T.; and Black, M. J. 2022. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 13243–13252.

Tome, D.; Russell, C.; and Agapito, L. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *CVPR*, 2500–2509.

Véges, M.; and Lőrincz, A. 2019. Absolute human pose estimation with depth prediction network. In *IJCNN*, 1–7. IEEE.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 601–617.

Vora, S.; Lang, A. H.; Helou, B.; and Beijbom, O. 2020. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4604–4612.

Wandt, B.; Rudolph, M.; Zell, P.; Rhodin, H.; and Rosenhahn, B. 2021. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *CVPR*, 13294–13304.

Wang, C.; Li, J.; Liu, W.; Qian, C.; and Lu, C. 2020a. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, 242–259. Springer.

Wang, C.; Ma, C.; Zhu, M.; and Yang, X. 2021. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11794–11803.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020b. Deep high-resolution representation learning for visual recognition. *TPAMI*, 43(10): 3349–3364.

Yin, T.; Zhou, X.; and Krähenbühl, P. 2021. Center-based 3D Object Detection and Tracking. *CVPR*.

Ying, J.; and Zhao, X. 2021. Rgb-D Fusion For Point-Cloud-Based 3d Human Pose Estimation. In *2021 IEEE International Conference on Image Processing (ICIP)*, 3108–3112. IEEE.

Yu, Z.; Wang, J.; Xu, J.; Ni, B.; Zhao, C.; Wang, M.; and Zhang, W. 2021. Skeleton2Mesh: Kinematics Prior Injected Unsupervised Human Mesh Recovery. In *ICCV*, 8619–8629.

Zhang, J.; Cai, Y.; Yan, S.; Feng, J.; et al. 2021. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34: 13153–13164.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022a. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13232–13242.

Zhang, J.; Wang, J.; Shi, Y.; Gao, F.; Xu, L.; and Yu, J. 2022b. Mutual Adaptive Reasoning for Monocular 3D Multi-Person Pose Estimation. *arXiv preprint arXiv:2207.07900*.

Zhao, C.; Ren, Y.; He, Y.; Cong, P.; Liang, H.; Yu, J.; Xu, L.; and Ma, Y. 2022. LiDAR-aid Inertial Poser: Large-scale Human Motion Capture by Sparse Inertial and LiDAR Sensors. *ArXiv*, abs/2205.15410.

Zhen, J.; Fang, Q.; Sun, J.; Liu, W.; Jiang, W.; Bao, H.; and Zhou, X. 2020. Smap: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, 550–566. Springer.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11656–11665.

Zheng, J.; Shi, X.; Gorban, A.; Mao, J.; Song, Y.; Qi, C. R.; Liu, T.; Chari, V.; Cornman, A.; Zhou, Y.; et al. 2022. Multi-modal 3D Human Pose Estimation with 2D Weak Supervision in Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4478–4487.

Zhu, X.; Ma, Y.; Wang, T.; Xu, Y.; Shi, J.; and Lin, D. 2020. SSN: Shape Signature Networks for Multi-class Object Detection from Point Clouds. *ECCV*.

Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Li, W.; Ma, Y.; Li, H.; Yang, R.; and Lin, D. 2021. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *TPAMI*.

Zimmermann, C.; Welschehold, T.; Dornhege, C.; Burgard, W.; and Brox, T. 2018. 3d human pose estimation in rgbd images for robotic task learning. In *ICRA*, 1986–1992. IEEE.