FreeCap: Hybrid Calibration-Free Motion Capture in Open Environments

Aoru Xue[1,*], **Yiming Ren**[1,*], **Zining Song**[1], **Mao Ye**[2], **Xinge Zhu**[3], **Yuexin Ma**[1,†]

[1]ShanghaiTech University
[2]Inceptio Technology
[3]Shanghai Jiao Tong University
{xuear2024, renym2022, mayuexin}@shanghaitech.edu.cn

Abstract

We propose a novel hybrid calibration-free method **FreeCap** to accurately capture global multi-person motions in open environments. Our system combines a single LiDAR with expandable moving cameras, allowing for flexible and precise motion estimation in a unified world coordinate. In particular, We introduce a local-to-global pose-aware cross-sensor human-matching module that predicts the alignment among each sensor, even in the absence of calibration. Additionally, our coarse-to-fine sensor-expandable pose optimizer further optimizes the 3D human key points and the alignments, it is also capable of incorporating additional cameras to enhance accuracy. Extensive experiments on Human-M3 and FreeMotion datasets demonstrate that our method significantly outperforms state-of-the-art single-modal methods, offering an expandable and efficient solution for multi-person motion capture across various applications.

Introduction

Vision-based human motion capture refers to the process of accurately predicting and reconstructing human meshes in open environments by utilizing visual data from LiDAR or camera. The approach eliminates the need for wearable sensors, enabling the capture of natural and unrestricted human motion. It is particularly advantageous in scenarios such as daily activities and sports, where the use of wearable devices may be impractical or restrictive.

Camera-based methods (Belagiannis and Zisserman 2016; Miezal, Taetz, and Bleser 2017; Rajasegaran et al. 2022; Shin et al. 2024) can effectively capture the accurate human local pose by image texture information but lack depth information, and depth camera-based methods (Sridhar et al. 2015; Wei, Zhang, and Chai 2012; Yu et al. 2017, 2019; Zheng et al. 2018) suffer from lighting conditions and the limited sensor range. LiDAR is widely used in robotics and autonomous driving (Zhu et al. 2021, 2020; Cong et al. 2022; Xu et al. 2023; Lu et al. 2023) for perceiving depth information in large-scale scenes. Previous LiDAR-based methods (Li et al. 2022; Fan et al. 2023b; Ren et al. 2024) can estimate the global human pose in large-scale scenes by accurate depth information. However, the sparsity of long-distance point clouds and the lack of texture information can lead to inaccuracies in predictions.

To leverage both image texture and LiDAR depth information for more accurate global human motion predictions in open environments. The previous method(Cong et al. 2023; Fan et al. 2023a) utilizes calibrated multiple cameras and LiDARs for involving the fusion of raw data, which can lead to difficulties in unifying data distribution and reduced generalization ability. In contrast, as shown in Fig. 1, our system consists of the single LiDAR and any number of moving cameras. The LiDAR captures extensive global depth information, while the moving cameras provide detailed texture information in local areas. This more flexible and expandable setup enables accurate multi-person motion estimation within a unified world coordinate system, making it suitable for diverse applications.

In this paper, we propose a novel calibration-free multi-model method **FreeCap**. We can not directly get the cross-sensor human matching in large-scale multi-person scenes without the calibration and the data-driven method can not intentionally learn the calibration when the camera is moving. To address these challenges, we introduce an efficient and high-quality local-to-global matching module, Pose-aware Cross-sensor Matching. Specifically, we use RTM-Pose (Jiang et al. 2023) and WHAM (Shin et al. 2024) to estimate the 2D key points and body pose in image and use LiveHPS (Ren et al. 2024) to estimate the 3D key points and body pose in point cloud. The body poses are used for local matching and the global key points are used for matching refinement. Next, we design a coarse-to-fine sensor-expandable pose optimizer to optimize the coarse align matrix calculated based on the matched 2D key points and 3D key points, which can unify the calibration-unknown multi-modal data and guide the network to further optimize the 3D human key points. Furthermore, considering the narrow of the camera range, our network can accept expandable camera data, the more cameras can assist LiDAR to predict more accurate results. Extensive experiments conducted on the multi-person large-scale dataset Human-M3(Fan et al. 2023a) and the multi-view sensor dataset FreeMotion(Ren et al. 2024), demonstrate that our method achieves significant improvements in human pose compared to other single-modal SOTA methods.

[*]These authors contributed equally.
[†]Corresponding author.

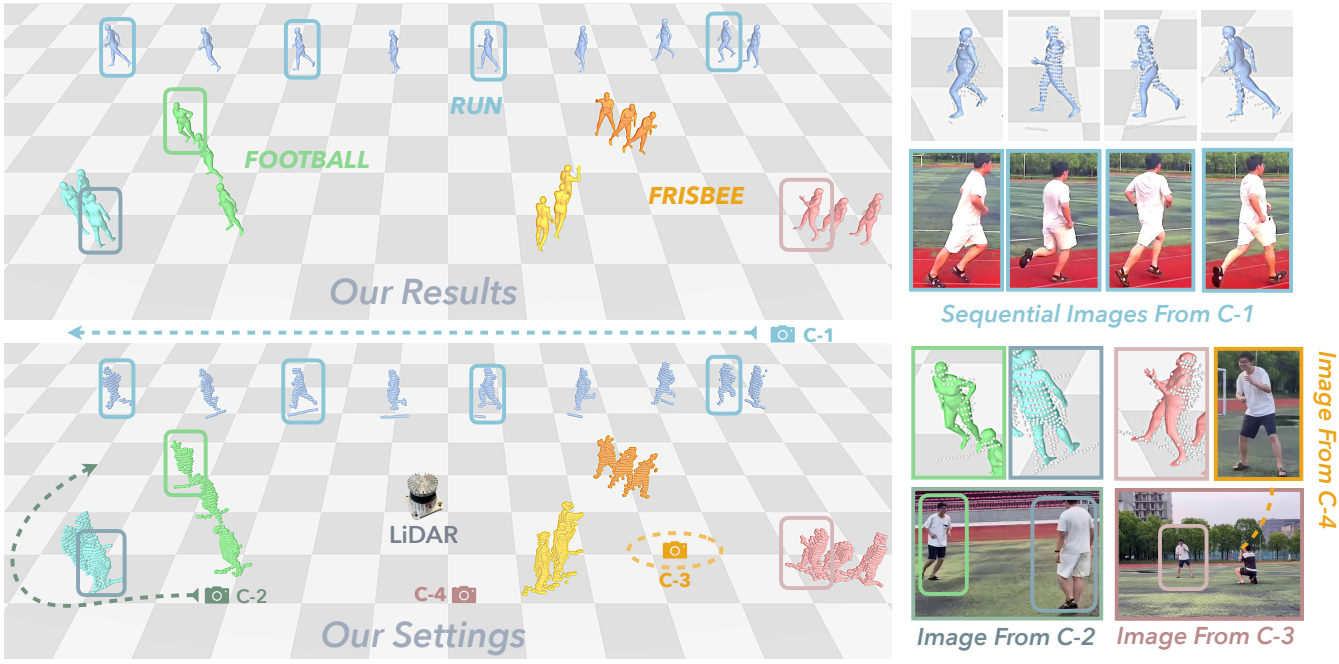

Figure 1: Visualization of our FreeCap in a real-time captured scenario. Our settings include a single LiDAR and four cameras. Camera-1 follows the running person, camera-2 surrounds two people playing soccer, camera-3 focuses on the main person playing frisbee and camera-4 captures three persons. We zoom in some cases to the right.

Our main contributions can be summarized as follows:

- We present the first calibration-free and sensor-expandable system designed for capturing multi-person motions in open environments.
- We propose a local-to-global cross-sensor human-matching approach for predicting alignment matrix.
- We design an effective multi-modal coarse-to-fine fusion method by optimizing human key points.
- Our method achieves SOTA on the large-scale dataset Human-M3 and multi-sensor dataset FreeMotion.

Related Work

Wearable Sensor-based Methods

Early motion capture system reconstructs human mesh relays on dense markers (Loper, Mahmood, and Black 2014; Park and Hodgins 2008; Song and Godøy 2016; Raskar et al. 2007; Xu et al. 2019) attach in human body. It enables the capture of high-quality human motions and is wildly applied in industry. However, system is costly and limited in indoor scenes with good lighting conditions. The freer system uses inertial measurement units (Yi, Zhou, and Xu 2021; Yi et al. 2022; Ren et al. 2023) for capturing orientations and accelerations of human key bones which reflect the human motion representation based on SMPL model. It can work in relatively open scenarios, but it suffers from drift and can be affected by magnetic field. Moreover, above methods require actors to wear sensors, causing constrained human motions.

Camera-based Methods

To make mocap method applied in daily usage, the marker-less methods (Liu et al. 2013; Rhodin et al. 2016; Kanazawa et al. 2018) use cameras only has made great progress. The multi-view camera-based methods (Malleson, Collomosse, and Hilton 2019; Zhang et al. 2020; Malleson, Collomosse, and Hilton 2020) predict accurate human motions based on optimization from full perspective motion information. Though performers do not need to wear markers, the multi-camera studio needs complex deployment in closed areas. The monocular-based method (Mehta et al. 2017; Sapp and Taskar 2013) only requires a single camera, it is light-weight while the performer is free to move, but a single camera can not perceive depth information and the sensor range is narrow. Recently, some methods (Shin et al. 2024; Rajasegaran et al. 2022), begin to pay attention to estimating the global human trajectory from a monocular dynamic camera. However, the trajectory is relative to a beginning frame and accumulated errors occur, which means that methods cannot infer long-term data and cannot predict multi-person global motions in a unified world coordinate system.

LiDAR-based Methods

In order to estimate the global human motions in large-scale open environments, various LiDAR-based mocap methods have been proposed. LiDARCap (Li et al. 2022) first proposes a single LiDAR-based human motion capture method, but it only estimates the local pose based on the graph-based framework. LiDAR-HMR (Fan et al. 2023b) utilizes the point cloud geometric information to reconstruct the hu-

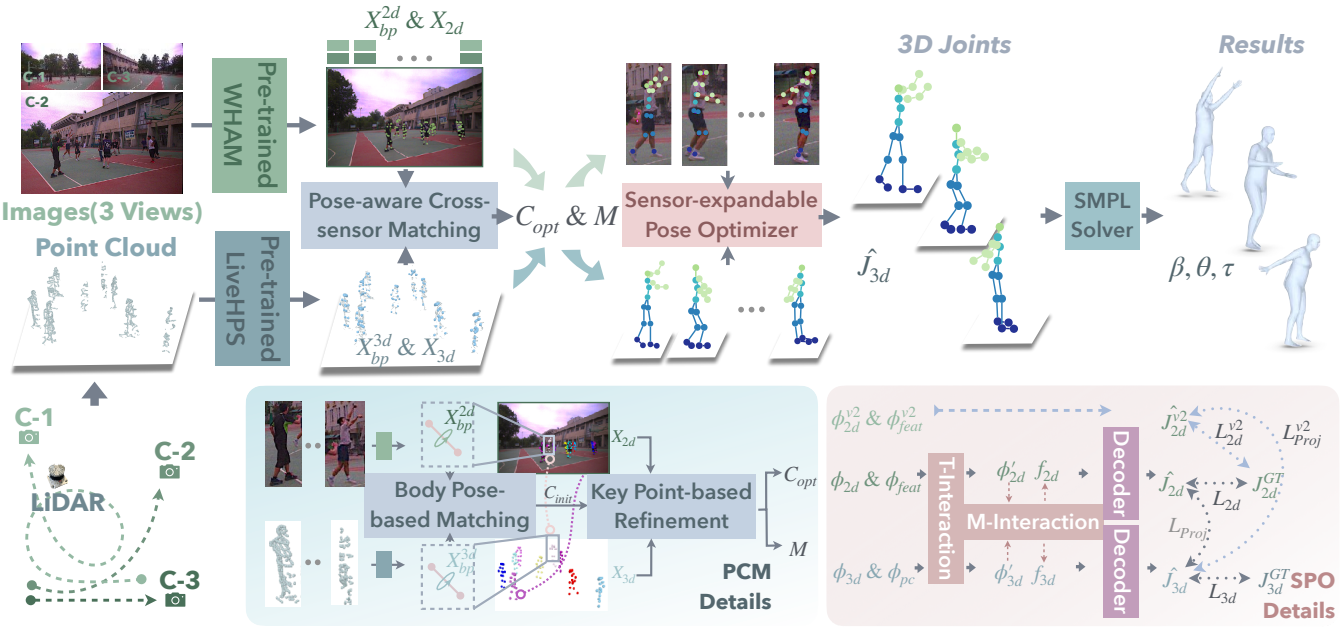

Figure 2: The pipeline of FreeCap. It consists of three main modules, including the pose-aware cross-sensor matching estimate of the optimal pairs and aligns matrix, the sensor-expandable pose optimizer predicts the 3D human joints, and the SMPL solver regresses the SMPL parameters. We also show the details of PCM and SPO.

man mesh, but it can not provide the template-based motion representations, the scope of application is limited. PointHPS (Cai et al. 2023) provides a cascaded network to predict the human pose and shape, but the network framework requires dense points in close range. LiveHPS and LiveHPS++ (Ren et al. 2024, 2025) estimate the full SMPL parameters, including the human local pose, body shape, and global translation, fully utilizing the depth information of point cloud. However, the sparsity of data and the lack of texture information also lead to inaccurate predictions.

Hybrid Sensor-based Methods

As all single modality sensor-based methods have corresponding disadvantages, recent works begin to explore the hybrid sensor-based method. Some methods utilize cameras and IMUs (Liang et al. 2022) to estimate more accurate human local poses, but still suffer from loss of depth information. LIP (Ren et al. 2023) uses the single LiDAR to provide depth information and four IMUs attached in for limbs to optimize the occlusion case. Human-M3 (Fan et al. 2023a) proposes a camera-LiDAR fusion method, but it relies on accurate calibrations causing the system must be fixed and pre-deployment. We present the first calibration-free method based on the single LiDAR and expandable cameras, which can work in open environments.

Methodology

Our goal is to capture human motion based on a single LiDAR with expandable moving cameras in the open environment. An overview of our pipeline is shown in Fig. 2. We take the sequential point clouds and multi-view videos as input. There are three main modules in our network, including pose-aware cross-sensor matching(PCM), sensor-expandable pose optimizer(SPO), and the SMPL solver. Firstly, we use PCM to get the matching pairs between point cloud and the images. Then, based on the matching pairs and predicted human key points, we use an optimization framework to generate the alignment matrix, which can unify the cross-sensor data into the unified world coordinate. Next, the SPO takes expandable unified data as the input and optimizes the 3D key points by multi-sensor fusion. Finally, we use the SMPL solver to estimate the SMPL parameters from refined 3D key points.

Preliminaries

Our framework takes sequential multi-person's point clouds and videos as input and predicts all person's SMPL parameters. For PCM, we define $x_{3d}(n,t)$ and $x_{2d}(n,t)$ to represent 3D and 2D key points; $x_{bp}^{3d}(n,t)$ and $x_{bp}^{2d}(n,t)$ as body poses extracted by pretrained model; $C(n,t)$ as matching pairs between LiDAR and camera for $n \in N$ individuals in the scene. For SPO, we define x_{pc} as normalized 3D point clouds; x_{3d} as 3D key points which is root-centered; x_{2d} as normalized 2D key point. The predicted 3D joints $\hat{J}_{3d}$ and 2D key points $\hat{J}_{2d}$ are supervised by the ground truth J_{3d}^{GT} and J_{2d}^{GT} respectively. We normalize J_{3d}^{GT} and J_{2d}^{GT} using the same normalization process applied to the input data following LiveHPS and WHAM. For the motion, We define $\theta^{GT}(n,t)$, $\beta^{GT}(n)$, and $T^{GT}(n,t)$ as the ground truth SMPL parameters, $N_J = 24$ and $N_V = 6890$ represents the number of human joint and mesh vertex.

Pose-aware Cross-sensor Matching

Human matching among multi-sensors is a prerequisite for our calibration-free method. We define the matching target is to get the matching pairs $C(t) = \{(i,j), i \in N, j \in M\}$, where N and M are the sets of indices for the human appears in the LiDAR and camera respectively. And we use $\mathbf{C}_I$ to represent identity matching. The body pose, inherently independent of coordinate systems, is ideal for cross-sensor matching applications. Yet, relying solely on body pose might not capture global information accurately. For instance, when individuals perform synchronized activities, body pose alone may not provide precise matches. To address this, we introduce **Pose-aware Cross-sensor Matching(PCM)**, which integrates both global key point and local body pose data, as outlined in Alg.1.

Algorithm 1: Pose-aware Cross-sensor Matching

Input: $\{\mathbf{x}_{bp}^{3d}, \mathbf{x}_{bp}^{2d}, \mathbf{x}_{3d}, \mathbf{x}_{2d}, K, n_{iter}\}$
Output: $\mathbf{C}$
$\mathbf{C}_{init}, \mathbf{M}_{init} \leftarrow$ Hungarian(Sim$(x_{bp}^{3d}, x_{bp}^{2d})$)
$n_{3d} \leftarrow$ len$(\mathbf{x}_{3d})$; $n_{2d} \leftarrow$ len$(\mathbf{x}_{2d})$;
$v \leftarrow$ Variance$(\mathbf{M}_{init}^{transl})$
if $v > \delta$ **then**
  $\mathbf{Q} \leftarrow$ zeros(n_{3d}, n_{2d})
  for $t \leftarrow 1$ ***to*** T **do**
    $\mathbf{C}_{opt}, M, cost \leftarrow$ OptMatch$(\mathbf{x}_{3d}(t), \mathbf{x}_{2d}(t), \mathbf{x}_{bp}^{3d}(t), \mathbf{x}_{bp}^{2d}(t), K_t, n_{iter})$
    foreach (i,j) *in* $\mathbf{C}_{opt}$ **do**
      $\mathbf{Q}[i,j] \leftarrow \mathbf{Q}[i,j] +$ cost
    end
  end
  $\mathbf{C} \leftarrow$ Hungarian$(\mathbf{Q})$
  return $\mathbf{C}$
end
else
  return $\mathbf{C}_{init}$
end

Body Pose-based Matching We use pre-trained LiveHPS to predict 3D key points $x_{3d}(n,t)$ and human body pose $x_{bp}^{3d}(n,t)$, RTMPose to predict 2D key points $x_{2d}(n,t)$ and WHAM to predict human body pose $x_{bp}^{2d}(n,t)$, which are then employed to predict the matching pairs $C(t) = \{(i,j), i \in N, j \in M\}$ between LiDAR and cameras. Then, we use sequential body poses $\mathbf{x}_{bp}^{3d}(n,t)$ and $\mathbf{x}_{bp}^{2d}(n,t)$ to calculate the body pose similarity for cost so that we can set up the Hungarian algorithm's cost matrix $\mathbf{Q}$, identifying a preliminary match C_{init} and a corresponding calibration matrix M_{init}. The similarity Sim$(x_{bp}^{3d}(n), x_{bp}^{2d}(m))$ between the 3D body pose of person index n and the 2D body pose of person index m is calculated as follows:

$$\text{sim}(x_{bp}^{3d}(n), x_{bp}^{2d}(m)) = \frac{1}{T}\sum_{t=1}^{T} \frac{x_{bp}^{3d}(n,t) \cdot x_{bp}^{2d}(m,t)}{\|x_{bp}^{3d}(n,t)\| \|x_{bp}^{2d}(m,t)\|}. \tag{1}$$

The matching cost $\mathbf{Q}$ is updated across all frames to secure the globally optimal matches using the Hungarian algorithm. Due to occlusions or sensor limitations, not all detected persons are matched; those with large re-projection errors are excluded to ensure optimal pairing. Additionally, to correct calibration errors potentially caused by the symmetric nature of human anatomy and to ensure the calibration reflects true orientations, we integrate global pose data. This approach also involves verifying matches by analyzing the variance v of the calibration's estimated translation over multiple frames. High variance re-projection errors are smoothed using a temporal window.

Key Point-based Optimization Matching To further exploit the information of points in the scene and achieve more accurate global matching results, we propose Key-Point based Optimization Matching, enhance the matching and calibration accuracy using 2D and 3D key points $\mathbf{x}_{2d}$ and $\mathbf{x}_{3d}$

Given the candidate match $C \in \mathbf{C}$, we can use the Perspective-n-Point (PnP) algorithm to estimate the camera pose $M_C = \text{PnP}(x_{3d}, x_{2d}, C)$ that best aligns the paired 2D key points in the image with their corresponding 3D key points in world coordinates. Based on this estimated camera pose and the camera intrinsic matrix K, we define a projection matrix:

$$P_C = K * M_C, \tag{2}$$

which captures the transformation from 3D world coordinates to 2D image coordinates.

Given the initial calibration matrix M_C, we define a function $E_{proj}(x_{3d}, x_{2d}, C)$ that calculates the re-projection error associated with that matrix:

$$E_{\text{proj}}(x_{3d}, x_{2d}, C) = \| P_C x_{3d}(C) - x_{2d}(C) \|_2 . \tag{3}$$

Thus, the problem of finding the optimal match can be formulated as:

$$C^* = argmin_{C \in \mathbf{C}}\{E_{proj}(x_{3d}, x_{2d}, C)\}. \tag{4}$$

Based on this, we can iteratively optimize the matching and calibration matrix according to the initial calibration matrix or initial matching. For the optimization process, we initialize proposal matches $\mathbf{C}_{proposal}$ with every 2D-3D pair first and optimize the proposal matches iteratively. We selected the set with the minimum weighted re-projection error and body pose error as the globally optimal matching. The weighted re-projection error and body pose error are defined below:

$$\begin{aligned} E_{proj}(x_{3d}, x_{2d}, x_{bp}^{3d}, x_{bp}^{2d}, C) = E_{proj}(x_{3d}, x_{2d}, C) \\ + \lambda_0 E_{proj}^{bp}(x_{bp}^{3d}, x_{bp}^{2d}, C), \end{aligned} \tag{5}$$

where the body pose matching error $E_{proj}^{bp}(x_{bp}^{3d}, x_{bp}^{2d}, C)$ is defined as below:

$$E_{proj}^{2d}(x_{bp}^{3d}, x_{bp}^{2d}, C) = \| J_{proj}(x_{bp}^{3d}, C) - J_{proj}(x_{bp}^{2d}, C) \|_2, \tag{6}$$

where $J_{proj}(x_{bp}^{3d}, C)$ and $J_{proj}(x_{bp}^{2d}, C)$ are the joint of SMPL with mean shape and body pose x_{bp}^{3d}, we project it to image plane with camera intrinsic matrix K. The detailed optimization process is presented in Alg. 2.

Our proposed Body Pose-aware Global Matching method leverages the distinctive qualities of body poses to bridge the sensorial gaps in multi-sensor environments. This method not only utilizes local body pose data but also incorporates global key point information, providing a robust framework for achieving accurate cross-sensor matching. By integrating both local and global data points, our approach significantly enhances the precision and reliability of the matching process, even in complex scenarios where traditional methods based solely on local data might fail.

Algorithm 2: OptMatch

Input: $\mathbf{x}_{bp}^{3d}$, $\mathbf{x}_{bp}^{2d}$, $\mathbf{x}_{3d}$, $\mathbf{x}_{2d}$, $\mathbf{K}$, n_{iter}
Output: $\mathbf{C}_{opt}$, M, cost_{max}
$n_{3d} \leftarrow \text{len}(\mathbf{x}_{3d})$; $n_{2d} \leftarrow \text{len}(\mathbf{x}_{2d})$;
$\mathbf{C}_{\text{proposal}} \leftarrow \emptyset$; $\text{cost}_{max} \leftarrow 0$; $\mathbf{C}_{opt} \leftarrow \emptyset$;
for $i \leftarrow 1$ ***to*** n_{3d} **do**
  for $j \leftarrow 1$ ***to*** n_{2d} **do**
    $\mathbf{Q} \leftarrow -E_{proj}(\mathbf{x}_{3d}[i], \mathbf{x}_{2d}[j], \mathbf{C}_I)$
    $\mathbf{C} \leftarrow \text{Hungarian}(\mathbf{Q})$
    $\mathbf{C}_{proposal} \leftarrow \mathbf{C}_{proposal} \cup \{\mathbf{C}\}$
  end
end
foreach $\mathbf{C}$ ***in*** $\mathbf{C}_{proposal}$ **do**
  for $i_{iter} \leftarrow 1$ ***to*** n_{iter} **do**
    $\mathbf{Q} \leftarrow -E_{proj}(\mathbf{x}_{3d}, \mathbf{x}_{2d}, \mathbf{x}_{bp}^{3d}, \mathbf{x}_{bp}^{2d}, \mathbf{C})$
    $\text{cost} \leftarrow \text{Sum}(\mathbf{Q})$
    if $cost > cost_{max}$ **then**
      $\text{cost}_{max} \leftarrow cost$; $\mathbf{C}_{opt} \leftarrow \mathbf{C}$
    end
    $\mathbf{C} \leftarrow \text{Hungarian}(\mathbf{Q})$
  end
end
$M \leftarrow \text{solvePnP}(\mathbf{x}_{3d}, \mathbf{x}_{2d}, \mathbf{C}_{opt})$
return $\mathbf{C}_{opt}$, M, cost_{max}

Sensor-expandable Pose Optimizer

Each individual's 3D and 2D data are matched and used to compute a preliminary calibration matrix. We avoid network overfitting from implicit multi-modal calibration by transforming the data from the LiDAR coordinate system to the camera coordinate system using this matrix. Then we introduce the **Sensor-expandable Pose Optimizer(SPO)** for optimizing the human 3D joints by effectively integrating data from each sensor.

The SPO's inputs are processed by three specialized MLP encoders: a 3D motion encoder E_{3d}, a 2D motion encoder E_{2d}, and a 3D point cloud encoder E_{pc}. These encoders respectively transform 3D key points x_{3d}, 2D key points x_{2d}, and 3D point clouds x_{pc} into feature representations ϕ_{3d}, ϕ_{2d}, and ϕ_{pc}.

Feature Integrator Lifting from 2D key points to 3D is inherently ambiguous. Therefore, we also utilize the static image features ϕ_{feat} extracted by a model, pre-trained on human mesh recovery task, as input, and integrate them with 2D feature integrator network I_{2d} to enhance the 2D motion features ϕ'_{2d}. The feature integrator employs residual connection layer with which the network can more easily learn the complex relationships between these features, resulting in a more expressive and comprehensive feature representation. Following WHAM, we pre-train the 2D motion encoder in SURREAL dataset with arbitrary camera motion. Similarly, point clouds and 3D key points features are integrated in the same way and get the enhanced feature ϕ'_{3d}.

Cross Modal Interaction Considering human motions are coherent over time, we employ temporal interactions to leverage the sequential nature of the data to capture temporal dependencies and correlations, thereby improving the overall results. We use multi-head self-attention layer to capture this temporal relationship and get the encoded feature f_{3d} and f_{2d}. Subsequently, we interact the features from the arbitrary two modalities(or two sensor features) by designing a bidirectional cross-attention module and get the f_{3d} and f_{2d} from two distinct modalities.

Finally, we input all the 3D and 2D features into the Motion Decoder D_{3d} and D_{2d}, where it predicts the 3D key points $\hat{J_{3d}}$ and 2D key points $\hat{J_{2d}}$ in their respective camera coordinate systems. To refine the results of 3D and 2D key points based on their projection correspondence, we add a regularization term for the re-projection loss. The loss functions of our method are defined below:

$$\begin{aligned} \mathcal{L}_{3d} =& \| \hat{J_{3d}} - J_{3d}^{GT} \|_2^2, \\ \mathcal{L}_{2d} =& \| \hat{J_{2d}} - J_{2d}^{GT} \|_2^2, \\ \mathcal{L}_{proj} =& \| Porj(\hat{J_{3d}}, K) - J_{2d}^{GT} \|_2^2, \\ \mathcal{L}_{total} =& \lambda_1 \mathcal{L}_{3d} + \lambda_2 \mathcal{L}_{2d} + \lambda_3 \mathcal{L}_{proj}. \end{aligned} \tag{7}$$

SMPL Solver

In the last stage, we transform the 3D skeleton key points obtained from the previous stage into the world coordinate system and use a temporal attention-based network to predict the parameters of the SMPL. The loss function of SMPL Solver is as follows:

$$\begin{aligned} \mathcal{L}_{smpl} =& \lambda_4 \mathcal{L}_{mse}(\beta) + \lambda_5 \mathcal{L}_{mse}(\theta) \\ &+ \lambda_6 \mathcal{L}_{mse}(J_{smpl}) + \lambda_7 \mathcal{L}_{mse}(V_{smpl}), \end{aligned} \tag{8}$$

where J_{smpl} and V_{smpl} are generated by SMPL:

$$J_{smpl}, V_{smpl} = \text{SMPL}(\beta, \theta, \tau). \tag{9}$$

Besides, we utilize an attention-based network for predicting the translation $T(n,t)$, within which we predict the distance $Tr(n,t)$ from the centroid of the point cloud $\overline{x}_{pc}(n,t)$ to the root node, mirroring the methodology employed in LiveHPS(Ren et al. 2024). The translation loss is defined as below:

$$\mathcal{L}_{mse}(Tr) = = \| \hat{Tr} - Tr^{GT} \|_2^2, \tag{10}$$

where $Tr^{GT} = T^{GT} - \overline{x}_{pc}^{GT}$.

Experiment

In this section, we compare our FreeCap with current LiDAR-based SOTA method LiveHPS (Ren et al. 2024) and

camera-based SOTA method WHAM (Shin et al. 2024) in Human-M3 (Fan et al. 2023a) and FreeMotion (Ren et al. 2024). In this experiment, we utilize only the indoor portion of the FreeMotion dataset, which includes multi-view camera data. The extensive experiments demonstrates the effectiveness and robustness of our multi-modal fusion strategy. Furthermore, we also present comprehensive ablation studies to evaluate the necessity of our network modules and fusion method, and the efficiency and generalization of our matching strategy. Finally, we also discuss the universality of our freecap in various settings. Following LiveHPS, our evaluation metrics include J/V Err(PS/PST)(mm), Ang Err($degree$), Accel Err(m/s^2) and SUCD(mm).

Implementation Details

We build our framework on PyTorch 2.0.0 and CUDA 11.8 and run the whole process on a server equipped with an Intel(R) Xeon(R) 444 E5-2678 CPU and 8 NVIDIA RTX3090 GPUs. For PCM, we set δ to 100, λ_0 to 0.1 and n_{iter} to 2 to get a stable matching. During the training of SPO, we train the network over 500 epochs with batch size of 32 and sequence length of 32, using an initial learning rate of 10^{-4}, and AdamW optimizer with weight decay of 10^{-4}. We set $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.01$ throughout our experiment. As for SMPL solver, we set $\lambda_4 = 1$, $\lambda_5 = 0.2$, $\lambda_6 = 10$ and $\lambda_7 = 1$. For the experiment, we consistently employed the SURREAL (Varol et al. 2017) dataset for pre-training throughout our experimental procedure, mirroring the approach adopted by LiveHPS. Following this, we pre-train the WHAM method on the SURREAL dataset and subsequently refine it through fine-tuning on both the Freemotion (Ren et al. 2024) and Human-M3(Fan et al. 2023a) datasets.

Comparison

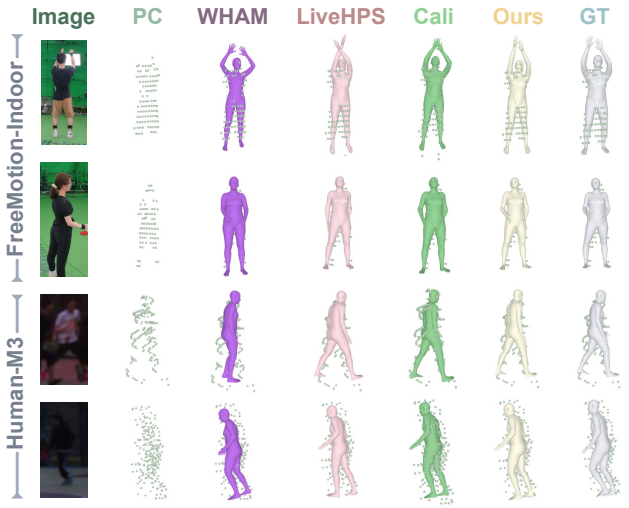

Figure 3: Qualitative comparisons. We show the global human mesh with point cloud, the point cloud matches the result better, representing more accurate estimation.

We evaluate our Freecap with SOTA method LiveHPS, WHAM and our designed calibration-based method in the testing dataset of FreeMotion and Human-M3 to demonstrate the effectiveness in multi-person large-scale scenes and the generalization in different novel views of cameras, here we only use single LiDAR and single camera as the input of FreeCap, the Human-M3 only provides the merged point cloud from all LiDARs. Our method achieves SOTA as shown in Tab. 1, it is worth noting that WHAM estimates the translation based on the provided location in the first frame. It is unfair for other methods, which are information independent of the starting frame, so we do not show the WHAM's global evaluations. FreeCap achieves significant improvement, especially in J/V Err(PS) and Ang Err, which reflect the accuracy of human local pose, other global evaluations also improved based on the better local pose estimation. Since there is no calibration-based LiDAR-camera fusion method, we also design a simple calibration-based method based on the transformer framework, while FreeCap is efficient with more expandable sensor settings. To evaluate the generalization of our calibration-free strategy, we also evaluate WHAM and FreeCap in the testing dataset with the novel view of camera. FreeCap can maintain performance stability when input data from unknown perspectives, while the performance of WHAM deteriorates significantly. The visual comparisons presented in Fig. 3 further underscore the superiority of our method in both pose and shape estimations, such as the first line, the LiveHPS estimates the incorrect human upper limb pose because of the occlusion in point cloud and WHAM estimates the incorrect human body shape because lack of depth information, while our fusion method utilizes the information from image to optimize the limb pose.

Ablation Study

We evaluate the superiority of each module in Human-M3 and FreeMotion. To prove the efficiency of our matching algorithm, we evaluate the matching accuracy and the frames per second(FPS) when inference in Human-M3. Finally, we also evaluate our sensor-expandable network framework in FreeMotion by adding more cameras.

Network Architecture. As Tab. 2 shown, the PCM module without matching refinement results in incorrect human matching pairs because of the similar body pose in the scenes, incorrect matching results in a significant effect for our calibration-free multi-modal fusion. In our SPO, we utilize the temporal interaction and the multi-sensor interaction, the result proves the significant improvements in both interactions. The more direct fusion strategy is to use the body pose estimated from camera and LiDAR for fusion, since the body pose is perspective-independent intermediate results, but it lost the raw data information. Moreover, FreeCap can predict the calibration matrix for calibration-based fusion, but the method cannot dynamically learn the useful information in different sensors.

Matching Strategy. To further demonstrate the accuracy and efficiency of our matching method, we conduct a detailed ablation study on the PCM, as shown in Tab. 3. The "KPs" represents we calculate all 2D key points and all 3D key points in every possible matching pair, which requires

	Human-M3					FreeMotion				
	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	SUCD↓	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	SUCD↓
WHAM	69.42/83.35	-	10.13	9.20	-	82.83/97.55	-	12.24	4.51	-
WHAM(NV)	85.50/101.44	-	10.42	9.25	-	108.16/124.52	-	19.89	4.49	-
LiveHPS	57.81/71.27	97.11/103.10	10.44	12.58	6.75	59.30/73.12	100.81/109.09	13.10	6.18	4.97
Cali-based	58.18/71.24	98.37/104.76	10.06	9.69	6.70	57.33/69.37	99.36/106.24	12.53	5.99	5.00
Ours	**55.45/68.52**	**96.47/102.67**	**9.14**	**9.60**	**6.66**	**53.31/65.50**	**95.97/102.91**	**11.14**	**5.97**	**4.82**
Ours(NV)	56.24/69.25	96.45/102.46	9.17	9.59	6.58	57.37/69.61	99.20/106.15	11.74	6.08	4.96

Table 1: Comparison with state-of-the-art methods on various datasets. Notably, the "Cali-based" method is our designed LiDAR-camera fusion method based on calibration information. "NV" represents the novel view of the camera in the test dataset.

	Human-M3					FreeMotion				
	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	SUCD↓	J/V Err(PS)↓	J/V Err(PST)↓	Ang Err↓	Accel Err↓	SUCD↓
w/o MT-Refine	59.01/ 73.07	103.44/ 110.03	9.29	9.94	8.00	-	-	-	-	-
w/o Temp-Int	59.55/73.78	103.79/ 110.60	9.35	9.71	8.06	55.48/68.09	97.54/104.81	11.47	6.20	11.79
w/o Sensor-Int	58.33/72.16	98.38/ 105.22	9.96	9.56	6.66	59.84/73.40	100.87/108.90	12.91	6.08	11.74
Body Pose	58.84/75.81	97.90/107.46	13.10	13.10	6.12	56.97/70.48	98.43/106.64	12.65	6.55	5.08
Calibration	59.46/72.77	98.70/105.07	10.13	9.66	6.62	58.57/71.21	100.17/107.38	12.23	6.15	4.98
Ours	**55.45/68.52**	**96.47/102.67**	**9.14**	**9.60**	**6.66**	**53.31/65.50**	**95.97/102.91**	**11.14**	**5.97**	**4.82**

Table 2: Ablation study for our network modules and multi-modal fusion methods.

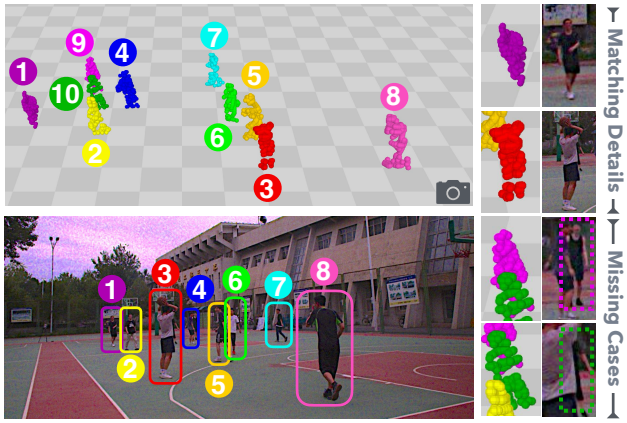

Figure 4: Visualization of our matching results in Human-M3. The view of camera and LiDAR is different, while the location of camera is labeled by the camera logo. We zoom in on some cases on the right.

the time complex is $\mathcal{O}(n^n)$. The "KP" represents we random select 3D key points of one person and calculate with all 2D key points, the time complex is $\mathcal{O}(n)$ but the accuracy decrease. "Pose" represents we only use the estimated body pose for single frame matching and "P & T" represents we use the sequential body pose for sequence matching. "P & K" represents our matching method for single frame matching without temporal information. Our method achieves the highest accuracy while balancing efficiency. Fig. 4 demonstrates our method's performance in extremely

	KPs	KP	Pose	P & K	P & T	**P & T & K**
Acc.(%)↑	89.88	77.07	81.59	85.87	96.01	**98.10**
FPS↑	0.08	12.06	127.47	3.40	309.28	**159.49**

Table 3: Ablation study for our matching method in Human-M3.

complex scenes with ten individuals, our method works well even if some persons in the camera are missing.

Conclusion

In this paper, we introduce a novel hybrid MoCap system that works in open environments without calibration, leveraging single LiDAR and expandable cameras. Our approach integrates the strengths of LiDAR and camera-based methods, utilizing cross-modal matching to effectively align and predict human poses in multi-person scenes. By combining 2D and 3D key points and optimizing the calibration matrix, our method adjusts matching conditions dynamically, ensuring accurate and robust motion capture. The extensive experiments demonstrate that our method achieves SOTA, offering a flexible and expandable solution for complex motion capture scenarios.

Acknowledgments

This work was supported by NSFC (No.62206173), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo).

References

Belagiannis, V.; and Zisserman, A. 2016. Recurrent Human Pose Estimation. ArXiv:1605.02914.

Cai, Z.; Pan, L.; Wei, C.; Yin, W.; Hong, F.; Zhang, M.; Loy, C. C.; Yang, L.; and Liu, Z. 2023. PointHPS: Cascaded 3D Human Pose and Shape Estimation from Point Clouds. *arXiv preprint arXiv:2308.14492*.

Cong, P.; Xu, Y.; Ren, Y.; Zhang, J.; Xu, L.; Wang, J.; Yu, J.; and Ma, Y. 2023. Weakly Supervised 3D Multi-Person Pose Estimation for Large-Scale Scenes Based on Monocular Camera and Single LiDAR. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1): 461–469.

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. *arXiv preprint arXiv:2204.01026*.

Fan, B.; Wang, S.; Guo, W.; Zheng, W.; Feng, J.; and Zhou, J. 2023a. Human-m3: A multi-view multi-modal dataset for 3d human pose estimation in outdoor scenes. *arXiv preprint arXiv:2308.00628*.

Fan, B.; Zheng, W.; Feng, J.; and Zhou, J. 2023b. LiDAR-HMR: 3D Human Mesh Recovery from LiDAR. *arXiv preprint arXiv:2311.11971*.

Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; and Chen, K. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end Recovery of Human Shape and Pose. In *CVPR*.

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

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*.

Liu, Y.; Gall, J.; Stoll, C.; Dai, Q.; Seidel, H.-P.; and Theobalt, C. 2013. Markerless motion capture of multiple characters using multiview image segmentation. *TPAMI*, 35(11): 2720–2735.

Loper, M.; Mahmood, N.; and Black, M. J. 2014. MoSh: Motion and shape capture from sparse markers. 33(6): 220:1–13.

Lu, Y.; Jiang, Q.; Chen, R.; Hou, Y.; Zhu, X.; and Ma, Y. 2023. See more and know more: Zero-shot point cloud segmentation via multi-modal visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21674–21684.

Malleson, C.; Collomosse, J.; and Hilton, A. 2019. Real-time multi-person motion capture from multi-view video and IMUs. *IJCV*, 1–18.

Malleson, C.; Collomosse, J.; and Hilton, A. 2020. Real-time multi-person motion capture from multi-view video and IMUs. *IJCV*, 128(6): 1594–1611.

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 *3DV*, 506–516. IEEE.

Miezal, M.; Taetz, B.; and Bleser, G. 2017. Real-time inertial lower body kinematics and ground contact estimation at anatomical foot points for agile human locomotion. In *ICRA*, 3256–3263.

Park, S. I.; and Hodgins, J. K. 2008. Data-driven Modeling of Skin and Muscle Deformation. 27(3): 96:1–6.

Rajasegaran, J.; Pavlakos, G.; Kanazawa, A.; and Malik, J. 2022. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2740–2749.

Raskar, R.; Nii, H.; Dedecker, B.; Hashimoto, Y.; Summet, J.; Moore, D.; Zhao, Y.; Westhues, J.; Dietz, P.; Barnwell, J.; et al. 2007. Prakash: lighting aware motion capture using photosensing markers and multiplexed illuminators. In *TOG*, volume 26, 36. ACM.

Ren, Y.; Han, X.; Yao, Y.; Long, X.; Sun, Y.; and Ma, Y. 2025. LiveHPS++: Robust and Coherent Motion Capture in Dynamic Free Environment. In *European Conference on Computer Vision*, 127–144. Springer.

Ren, Y.; Han, X.; Zhao, C.; Wang, J.; Xu, L.; Yu, J.; and Ma, Y. 2024. LiveHPS: LiDAR-based Scene-level Human Pose and Shape Estimation in Free Environment. *arXiv preprint arXiv:2402.17171*.

Ren, Y.; Zhao, C.; He, Y.; Cong, P.; Liang, H.; Yu, J.; Xu, L.; and Ma, Y. 2023. LiDAR-aid Inertial Poser: Large-scale Human Motion Capture by Sparse Inertial and LiDAR Sensors. *TVCG*.

Rhodin, H.; Richardt, C.; Casas, D.; Insafutdinov, E.; Shafiei, M.; Seidel, H.-P.; Schiele, B.; and Theobalt, C. 2016. EgoCap: Egocentric Marker-less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph.*, 35(6): 162:1–162:11.

Sapp, B.; and Taskar, B. 2013. MODEC: Multimodal Decomposable Models for Human Pose Estimation. In *CVPR*.

Shin, S.; Kim, J.; Halilaj, E.; and Black, M. J. 2024. WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Song, M.-H.; and Godøy, R. I. 2016. How Fast Is Your Body Motion? Determining a Sufficient Frame Rate for an Optical Motion Tracking System Using Passive Markers. *PLOS ONE*, 11(3).

Sridhar, S.; Mueller, F.; Oulasvirta, A.; and Theobalt, C. 2015. Fast and Robust Hand Tracking Using Detection-Guided Optimization. In *CVPR*.

Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M.; Laptev, I.; and Schmid, C. 2017. Learning from synthetic humans. In *CVPR*.

Wei, X.; Zhang, P.; and Chai, J. 2012. Accurate Realtime Full-body Motion Capture Using a Single Depth Camera. *SIGGRAPH Asia*, 31(6): 188:1–12.

Xu, L.; Su, Z.; Han, L.; Yu, T.; Liu, Y.; and FANG, L. 2019. UnstructuredFusion: Realtime 4D Geometry and Texture Reconstruction using CommercialRGBD Cameras. *TPAMI*, 1–1.

Xu, Y.; Cong, P.; Yao, Y.; Chen, R.; Hou, Y.; Zhu, X.; He, X.; Yu, J.; and Ma, Y. 2023. Human-centric scene understanding for 3d large-scale scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20349–20359.

Yi, X.; Zhou, Y.; Habermann, M.; Shimada, S.; Golyanik, V.; Theobalt, C.; and Xu, F. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. In *CVPR*.

Yi, X.; Zhou, Y.; and Xu, F. 2021. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13.

Yu, T.; Guo, K.; Xu, F.; Dong, Y.; Su, Z.; Zhao, J.; Li, J.; Dai, Q.; and Liu, Y. 2017. BodyFusion: Real-time Capture of Human Motion and Surface Geometry Using a Single Depth Camera. In *ICCV*. ACM.

Yu, T.; Zheng, Z.; Guo, K.; Zhao, J.; Dai, Q.; Li, H.; Pons-Moll, G.; and Liu, Y. 2019. DoubleFusion: Real-time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor. *TPAMI*.

Zhang, Z.; Wang, C.; Qin, W.; and Zeng, W. 2020. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *CVPR*, 2200–2209.

Zheng, Z.; Yu, T.; Li, H.; Guo, K.; Dai, Q.; Fang, L.; and Liu, Y. 2018. HybridFusion: Real-time Performance Capture Using a Single Depth Sensor and Sparse IMUs. In *ECCV*.

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. In *ECCV*, 581–597. Springer.

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*.