LIF-Net: LiDAR-Camera Fusion for 3D Human Pose in Urban Scenes

Max A. Buettner[*†], Erik Schuetz[*†], Fabian B. Flohr[*]
[*]Munich University of Applied Sciences, Intelligent Vehicles Lab (IVLab), Munich, Germany
{max.buettner, erik.schuetz, fabian.flohr}@hm.edu

***Abstract*—Estimation of human pose and shape (3DHPS) in 3D is crucial to ensure the safety of vulnerable road users (VRUs) in autonomous driving (AD) scenarios, as it can serve as an additional feature for trajectory prediction and ego-motion planning in complex urban environments. To tackle this problem, we propose a novel 3DHPS model called LIF-Net, which utilizes RGB *and* LiDAR data to estimate VRUs as complete 3D bodies in the world coordinate system. LIF-Net utilizes a two-branch approach to encode both modalities separately into a latent representation. A cross-attention-based intermediate-fusion module learns to combine the two representations into a joint latent feature space, which is used by a Transformer decoder module to predict the parameters of a Skinned Multi-Person Linear Model (SMPL). We train and evaluate LIF-Net on the Waymo Open Dataset (WOD), containing challenging real-world scenarios with 2D and 3D keypoint annotations.**

Experimental results demonstrate the effectiveness and robustness of our approach compared to single-modality methods in challenging real-world AD scenarios, including poor lighting, occlusions, and varying bounding box and detection quality, while achieving an MPJPE of $122\,\mathrm{mm}$ and a PA-MPJPE of $76\,\mathrm{mm}$, which is an improvement in MPJPE of 35.5% over image-based methods and 9.1% over the LiDAR-based method. Code and model will be made publicly available at `lif-net`.

***Index Terms*—3D Human Pose and Shape Estimation, LiDAR-Camera Fusion, Autonomous Driving, SMPL**

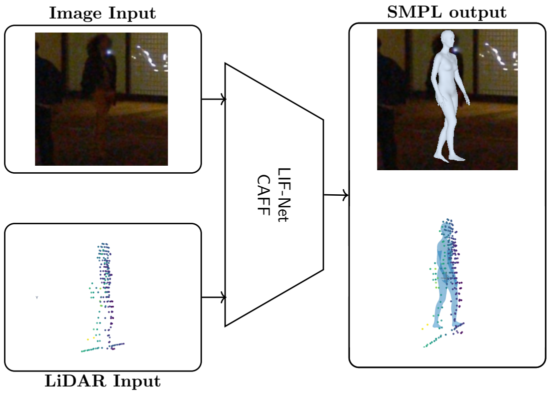

Fig. 1: **LIF-Net** uses multi-modal inputs to robustly and accurately estimate SMPL in both camera and LiDAR with cross-attention feature-fusion (CAFF), even if one sensor modality is weak (i.e., adverse lightning conditions in the image).

I. Introduction

Understanding human behavior in 3D is a foundational goal of computer vision, with critical applications in fields ranging from robotics to augmented reality. This task is especially crucial for the safety and reliability of autonomous vehicles (AVs), where a deep understanding of pedestrian pose and shape is essential for accurately predicting intent and navigating complex urban environments [1].

Despite increased research on multi-modal sensor fusion for AD [2]–[4], the dominant paradigm for 3D Human Pose and Shape (3DHPS) estimation has been monocular, which regresses parameters of a statistical body model like SMPL [5] directly from a single RGB image [6]. While these methods have achieved impressive results, they are fundamentally constrained by the ill-posed nature of lifting 2D observations into 3D space. This leads to an inherent vulnerability to depth ambiguities, where multiple 3D poses can correspond to the same 2D projection. This problem is exacerbated in real-world driving scenarios, which are characterized by severe partial occlusions, significant variations in subject distances, and challenging lighting conditions.

To overcome these limitations, on-board sensors like LiDAR offer direct 3D geometric information, providing an accurate signal for depth. However, research leveraging point clouds for 3DHPS is less developed. While some works have demonstrated feasibility with LiDAR-only methods [7]–[9], they are limited by the inherent sparsity of the data. The most promising direction is therefore multi-modal fusion, yet existing methods in this niche are also limited. The most closely related work, the LEIR model [10], relies on concatenation-based fusion and, crucially, was validated only in a controlled, indoor sports environment, not in the wild. This reveals a critical gap in the literature: a lack of robust, advanced fusion architectures for SMPL-based mesh recovery that have been validated in the complex, large-scale scenarios typical of autonomous driving.

In this work, we target this gap directly. We propose a novel, transformer-based architecture for multi-modal 3D human mesh recovery that can robustly process combined RGB and LiDAR inputs. A key aspect of our contribution is a method to generate accurate SMPL pseudo-ground truths for

[†] Authors contributed equally.

existing AD datasets by leveraging supervision from their 2D and 3D joint annotations. The core of our network is a cross-attention module that fuses extracted LiDAR features with RGB features by dynamically weighing each modality based on learned data to capture different situations, thus enhancing robustness and producing accurate mesh reconstructions, as illustrated in Figure 1 and further in Figure 3. We validate our approach by training and performing extensive evaluations on the large-scale Waymo Open Dataset.

In response to the challenges of single-modal estimation and the lack of robust fusion architectures for the autonomous driving domain, our main contributions are as follows:

- We introduce a novel intermediate-fusion architecture for the autonomous driving domain that uses cross-attention to address the complementary failure modes of RGB and LiDAR for robust, state-of-the-art SMPL estimation.
- We perform extensive evaluation of a multi-modal SMPL-based pose estimation method on the challenging, large-scale, and multi-modal Waymo Open Dataset.
- We show significant performance improvements compared to single-modality SMPL estimation approaches, reducing the pose error (MPJPE) by up to 35.5% over image-based and about 9% over LiDAR-based methods, demonstrating the effectiveness and robustness of our fusion strategy.

II. Related Works

A. Single-Modal 3D Human Mesh Recovery

Research in 3D Human Pose and Shape (3DHPS) estimation has been dominated by monocular methods, which began with the foundational works of the SMPL model [5] and regression-based Human Mesh Recovery (HMR) [6]. Early approaches relied on iterative optimization to fit the model to 2D evidence [11], but the field has since shifted towards end-to-end regression networks. These are often trained on a combination of datasets, leveraging in-the-loop optimization [12] or powerful backbones [13] to generate pseudo-ground truth annotations. More recently, transformer-based architectures have become state-of-the-art, using pose priors as queries [14], learned tokens [15], or explicit camera modeling [16] to achieve robust results. An emerging trend is to further condition these models on other modalities, such as natural language, where text prompts are used to guide the pose estimation and resolve ambiguities, as demonstrated by PromptHMR [17]. However, despite these advancements, all methods that rely primarily on a single RGB image are fundamentally constrained by the ill-posed nature of lifting 2D observations to 3D space, making them inherently vulnerable to depth ambiguity and occlusions.

To directly address this depth ambiguity, a smaller but growing body of work explores 3DHPS directly from 3D point clouds. Pioneering methods have adapted point cloud processing architectures like PointNet++ [18] to regress SMPL parameters from LiDAR scans, demonstrating feasibility in both indoor [19] and outdoor settings [7], [9]. Another approach, taken by LiDAR-HMR [8], introduces a sparse-to-dense reconstruction network to progressively refine the body mesh from the sparse input. While these approaches successfully leverage direct geometric information, they are limited by the inherent properties of the data: LiDAR point clouds are often extremely sparse, especially for distant subjects, and lack the rich semantic and textural cues from images that are vital for identifying fine-grained details. The respective limitations of each modality, i.e., depth ambiguity in RGB and geometric sparsity in LiDAR, are complementary, creating a strong motivation for multi-modal fusion that combines their respective strengths.

B. Multi-Modal Fusion for 3D Perception

Given the complementary nature of RGB and LiDAR, multi-modal fusion has become a key research direction for robust 3D perception. A common baseline for tasks like 3D object detection is to concatenate unimodal features and process them with an MLP [28]; however, this approach does not explicitly model the spatial relationships between feature sets. Consequently, state-of-the-art methods have shifted to attention-based mechanisms that can dynamically align and weigh features from different sensors [2], [3]. While these advanced fusion techniques are well-established for general perception, their application to 3D Human Pose and Shape (3DHPS) estimation remains less developed. Prior works in applying multimodal feature fusion to 3DHPS have progressed from regressing keypoints [29]–[31] to generating non-parametric human meshes [32], but the task of SMPL parameter estimation in the wild is rarely addressed. The most closely related work that does, the LEIR model [10], relies on concatenation-based fusion and was validated only in controlled, indoor environments. This reveals a clear gap for a more sophisticated, attention-based fusion architecture validated on a challenging, in-the-wild autonomous driving dataset.

C. Pose Estimation in the Autonomous Driving Domain

3DHPS in the autonomous driving (AD) domain presents unique challenges compared to controlled studio datasets, e.g., severe occlusions, large subject distances, and ego-motion. While prior work has addressed 3D keypoint estimation in this context [26], full parametric mesh recovery remains a significant problem, primarily due to the scarcity of representative, large-scale datasets. A survey of the data landscape (Table I) reveals an evolution from these early studio datasets [20], [21] towards more diverse in-the-wild [22], [23] and synthetic [24] collections for pre-training [14], [15]. However, datasets with the synchronized RGB and LiDAR data required for multi-modal research are scarce, and those that do exist [7], [9], [10] are typically captured in controlled, non-AD environments.

The Waymo Open Dataset (WOD) [27] is a notable exception, providing the necessary scale and diversity of complex urban scenarios to serve as a suitable benchmark for this task. This highlights a critical gap in the literature: a lack of robust, SMPL-based mesh recovery methods specifically

TABLE I: Dataset overview and comparison. The column "Frames" refers to the number of frames annotated with any type of keypoint or pose information. 2DKP, 3DKP, and SMPL refer to the dataset labeled with 2D keypoints, 3D keypoints, and SMPL parameters respectively. Multi. Subj. refers to the scene holding multiple or only a single annotated subject. ITW and AD refer to the dataset being in the wild (ITW), i.e., non-studio environment, or AD domain.

Dataset	**Data structure**								
	Frames	***Modality C***	***Modality L***	***2DKP***	***3DKP***	***SMPL***	***multiple subj.***	***ITW***	***AD***
Human3.6M [20]	3.6M	✓	✗	✓	✓	✗	✗	✗	✗
MPII [21]	40.5k	✓	✗	✓	✗	✗	✗	✓	✗
3DPW [22]	51k	✓	✗	✓	✓	✓	✓	✓	✗
EMDB [23]	105k	✓	✗	✓	✓	✓	✗	✗	✗
BEDLAM [24]	380k	✓	✗	✓	✓	✓	✓	✗	✗
LiDARHuman2.6M† [7]	18.4k†	✓	✓	✓	✓	✓	✗	✗	✗
SLOPER4D† [9]	32k†	✓	✓	✓	✓	✓	✗	✗	✗
RELI11D† [10]	3.6k†	✓	✓	✓	✓	✓	✗	✗	✗
FreeForm [25]	578k	✓	✓	✓	✓	✓	✓	✓	✗
PedX [26]	2.5k	✓	✓	✓	✓	✗	✓	✓	✓
WOD [27]	178k	✓	✓	✓	✓	✗	✓	✓	✓

† Refers to publicly available portion of the dataset, not total number reported in the original publication.

designed for and validated on a large-scale, multi-modal AD benchmark. Our work directly targets this gap, proposing a novel fusion architecture and demonstrating its effectiveness on this challenging domain.

III. Method

A. Preliminaries

a) SMPL: In this work, we utilize the SMPL model [5]. The model $\mathcal{M}$ is a function, $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \rightarrow \mathbf{V}$, that maps body pose ($\boldsymbol{\theta} \in \mathbb{R}^{72}$) and shape ($\boldsymbol{\beta} \in \mathbb{R}^{10}$) parameters to a high-resolution 3D vertex mesh $\mathbf{V} \in \mathbb{R}^{6890\times 3}$. These vertices can then be mapped to 3D keypoints via a pre-defined joint regressor.

b) HMR: Human mesh recovery (HMR) aims to reconstruct a person via the SMPL pose θ and shape β parameters by learning the predictor of $f(\mathcal{I}, \mathcal{P})$, when image $\mathcal{I}$, point cloud $\mathcal{P}$, or both are provided. The function also includes π, the camera translation, allowing the prediction of 3D meshes solely by image and resulting in $f(\mathcal{I}, \mathcal{P}) = (\theta, \beta, \pi)$.

c) Coordinate systems: We define three right-handed coordinate systems used throughout this work: (1) $\mathcal{F}_{\text{world}}$ [27]: a vehicle-centered frame; (2) $\mathcal{F}_{\text{cam}}$ [33]: a standard pinhole camera frame; and (3) $\mathcal{F}_{\text{SMPL}}$ [5]: a pelvis-centered body frame. All input 3D data, such as LiDAR points and ground-truth 3D joints, are initially provided in $\mathcal{F}_{\text{world}}$. Our entire processing pipeline, including the final SMPL mesh prediction, operates in the $\mathcal{F}_{\text{cam}}$ frame. We use the calibrated extrinsic and intrinsic parameters provided by the dataset to transform data from $\mathcal{F}_{\text{world}}$ to $\mathcal{F}_{\text{cam}}$.

B. Architecture

The overall architecture of our proposed network, illustrated in Figure 2, comprises four main components: a pre-trained RGB encoder, a LiDAR encoder, a feature fusion module, and a transformer decoder head initialized with weights pre-trained for the task of SMPL parameter regression. While the RGB encoder and decoder head leverage pre-trained weights, the LiDAR encoder and fusion module are trained from scratch to output the final pose ($\boldsymbol{\theta}$) and shape ($\boldsymbol{\beta}$) parameters for the SMPL model.

RGB Encoder. To extract rich semantic features, we employ a Vision Transformer Large (ViT-H/16) backbone [34], initialized with ViTPose [35] weights pre-trained for human pose estimation. This choice is motivated by the strength of vision transformers in modeling global context via self-attention [36], [37], making them highly effective for dynamic outdoor scenes. The encoder processes an input RGB crop $\mathbf{X}_{\text{rgb}} \in \mathbb{R}^{3\times 256\times 256}$ from ground-truth bounding boxes, tokenizes it into 256 patches, and passes them through 24 transformer blocks to produce the final feature map $f_{\text{img}} \in \mathbb{R}^{B\times 1280\times 16\times 12}$.

LiDAR Encoder. To encode geometric structure from sparse point clouds, we employ a PointNet++ [18] backbone that is trained from scratch on our target dataset. While other methods like VoxelNet [38] operate on voxelized representations, our choice is consistent with recent pose estimation literature [8], [10], [29] as PointNet++ directly processes raw points, excelling at capturing fine-grained detail by hierarchically learning local and global context. Before encoding, the point cloud for each subject is transformed into the sensor's coordinate frame and centered on the subject's root. We sample or pad this cloud to a fixed size of $N_P = 512$ points, where the padding operation randomly selects points, copies them, and applies a small normally distributed noise with $\sigma = 0.01$. The encoder processes the resulting input $\mathbf{X}_{\text{pc}} \in \mathbb{R}^{N_P\times 3}$ to produce a global feature vector $f_{\text{LiDAR}} \in \mathbb{R}^{B\times 1024}$ for the fusion module.

Feature Fusion. To robustly integrate visual and geometric information, we employ a cross-attention module with eight attention heads. We choose this over MLP-based fusion [28] or unimodal self-attention [39] to explicitly model inter-modal interactions. The core strategy is to use the dense feature map from the image encoder to query the global, geometric context provided by the LiDAR features, enhancing robustness against sensor degradation. The flattened image feature map forms the query tensors, Q_{rgb}. Concurrently, the global LiDAR

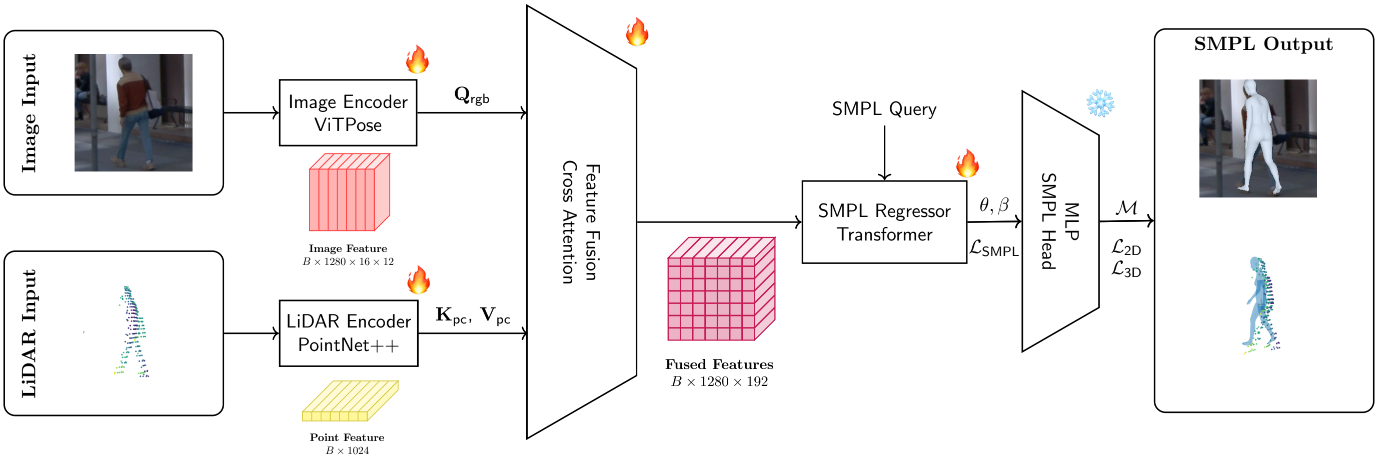

Fig. 2: An overview of our multi-modal architecture for 3D human pose and shape estimation. The network fuses features from a ViTPose image encoder and a PointNet++ LiDAR encoder via cross-attention using RGB features as queries and LiDAR features as keys and values. A transformer regressor predicts SMPL parameters ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$), supervised by 2D/3D keypoint ($\mathcal{L}_{\text{2D}}, \mathcal{L}_{\text{3D}}$) and parameter ($\mathcal{L}_{\text{SMPL}}$) losses. Modules marked with a fire icon are either trained or fine-tuned, while those marked with a snowflake icon are frozen.

feature is linearly projected to match the image feature's channel dimension, forming the keys and values K_{pc} and V_{pc}. Multi-head attention computes the fused representation $f_{\text{fusion}} \in \mathbb{R}^{B\times 192\times 1280}$, which is passed to the decoder head, according to:

$$f_{\text{fusion}} = \text{softmax}\left(\frac{Q_{\text{rgb}} K_{\text{pc}}^T}{\sqrt{d_k}}\right) V_{\text{pc}}, \tag{1}$$

where d_k is the key dimensionality.

Transformer Decoder Head. To regress the final SMPL parameters, we use a transformer decoder head which is initialized with pre-trained HMR2 weights [14] and fine-tuned on our multi-modal data using the fused feature map f_{fusion} as context. Following [14], an SMPL mean-parameter-initialized query token attends to this context, with its output passed through linear layers to predict a residual update for the final body pose ($\boldsymbol{\theta}$) and shape ($\boldsymbol{\beta}$).

SMPL Model. The final output of our network consists of the SMPL pose ($\boldsymbol{\theta}$) and shape ($\boldsymbol{\beta}$) parameters, which generate a 3D human mesh. As detailed in the preliminaries (Section III-A), we utilize the SMPL model [5], and its weights remain frozen during training.

C. Losses

Following state-of-the-art methods [6], [14], [15], our network is trained end-to-end by minimizing a total objective function, $\mathcal{L}_{\text{total}}$, which is a dynamically adapted, weighted sum of up to three components. Our training is primarily driven by the 3D keypoint loss, $\mathcal{L}_{\text{3D}}$, which provides direct spatial supervision using an L1 error between the predicted ($\hat{y}$) and ground-truth (y) 3D keypoints.

$$\mathcal{L}_{\text{3D}} = \|\hat{y} - y\|_1 \tag{2}$$

The second key supervision signal is the 2D keypoint loss, $\mathcal{L}_{\text{2D}}$, which anchors the 3D estimate in the image plane by projecting the predicted 3D joints $\hat{y}$ using the camera projection function π and penalizing the L1 error against the 2D ground truth y.

$$\mathcal{L}_{\text{2D}} = \|\pi(\hat{y}) - y\|_1 \tag{3}$$

To ensure plausible predictions, we regularize the output with an L2 loss on the SMPL parameters, $\mathcal{L}_{\text{SMPL}}$ (Eq. 4):

$$\mathcal{L}_{\text{SMPL}} = \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2^2 + \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 \tag{4}$$

$$\tag{5}$$

The final objective function is a weighted sum of these components (see Sec. IV-C for hyperparameter values):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{3D}}\mathcal{L}_{\text{3D}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \lambda_{\text{SMPL}}\mathcal{L}_{\text{SMPL}} \tag{6}$$

IV. Experiments

To validate our proposed multi-modal fusion architecture, we conduct a series of comprehensive experiments on the large-scale Waymo Open Dataset (WOD) [27]. We first detail our experimental setup, including the dataset preparation, evaluation metrics, and implementation details. We then present our main quantitative results, comparing our model's performance against state-of-the-art single-modal and multi-modal baselines. Following this, we perform extensive ablation studies to rigorously analyze the contribution of each key component of our model, particularly the impact of our cross-attention fusion mechanism. Finally, we provide qualitative results to visually demonstrate our model's performance and robustness in challenging, in-the-wild autonomous driving scenarios.

A. Dataset

Dataset Selection. Our work requires a large-scale, multi-modal dataset situated in the autonomous driving (AD) domain. As shown in our dataset survey in Table I, WOD [27] is the only publicly available dataset that simultaneously provides urban AD scenarios, multi-subject complexity, synchronized LiDAR-camera data, and human keypoint annotations in 2D and 3D. While it lacks the direct SMPL annotations needed for our primary task, its scale and in-the-wild nature make it the most suitable benchmark.

Dataset Statistics and Splits. The WOD dataset provides approximately 179k pedestrian samples, which we divide according to the official train/val split. As shown in Table II, these samples have varying annotation availability. Our experiments leverage the different subsets based on whether they have 2D keypoints ($\mathtt{WOD}_{\text{2D}}$), 3D keypoints ($\mathtt{WOD}_{\text{3D}}$), or both ($\mathtt{WOD}_{\text{2D3D}}$). All models are ultimately evaluated on the $\mathtt{WOD}^{\text{val}}_{\text{2D3D}}$ split, which contains samples with complete multi-modal annotations.

Pseudo-Ground Truth Generation. Since WOD lacks SMPL labels, we generate pseudo-ground truths for our training set. Instead of using traditional optimization-based fitting, which can yield noisy and implausible poses, we leverage a state-of-the-art regression model, TokenHMR [15]. The key advantage is that TokenHMR's discrete latent space acts as a powerful learned prior, ensuring all generated poses are anatomically realistic and providing a stable supervision signal. To further refine these labels, we replace the regressed global orientation with the orientation derived from the ground-truth 3D bounding boxes. For consistency across all joint annotations we apply the COCO convention.

B. Evaluation Metrics

Following common practice [14]–[16], we evaluate the performance of our method using two standard metrics for 3D human pose estimation, reported in millimeters (mm), where lower is better. Our core 3D evaluation relies on Mean Per Joint Position Error (MPJPE) and its Procrustes-Aligned variant (PA-MPJPE).

MPJPE measures the mean Euclidean distance between the N predicted 3D joints ($\hat{\mathbf{y}}$) and ground-truth 3D joints ($\mathbf{y}$) after aligning their root (pelvis) joints. It provides a measure of absolute pose accuracy, defined as:

$$\text{MPJPE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{y}_i - y_i \right\|_2 . \qquad (7)$$

Procrustes-Aligned MPJPE (PA-MPJPE) computes the error after finding the optimal scale factor $s \in \mathbb{R}$, rotation matrix R, and translation vector $t \in \mathbb{R}^3$ that best align the predicted pose with the ground truth. Therefore, PA-MPJPE is used to navigate to the source of potential errors within the pose itself, while MPJPE is a stricter metric better suited to measure true pose accuracy without such alignment. The metric is defined as:

$$\text{PA-MPJPE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \left\| sR\hat{y}_i + t - y_i \right\|_2 . \qquad (8)$$

TABLE II: Overview of dataset splits, listing the number of frames and counts of 2D-only, 3D-only, and combined 2D+3D samples with camera (C), LiDAR (L), 2D keypoints (KP), and 3D keypoints (KP). Symbols: ✓= fully available, × = not available, ∼ = partially available. The final "All" category represents the total accumulated count of all unique pedestrian samples across the different subsets.

	Split	Samples	C	L	2D KP	3D KP
2D	$\mathtt{WOD}^{\text{train}}_{\text{2D}}$	141,348	✓	~	✓	×
	$\mathtt{WOD}^{\text{val}}_{\text{2D}}$	26,619	✓	~	✓	×
	Total	167,967	✓	~	✓	×
3D	$\mathtt{WOD}^{\text{train}}_{\text{3D}}$	3,946	~	✓	×	✓
	$\mathtt{WOD}^{\text{val}}_{\text{3D}}$	1,063	~	✓	×	✓
	Total	5,009	~	✓	×	✓
2D3D	$\mathtt{WOD}^{\text{train}}_{\text{2D3D}}$	4,655	✓	✓	✓	✓
	$\mathtt{WOD}^{\text{val}}_{\text{2D3D}}$	905	✓	✓	✓	✓
	Total	5,560	✓	✓	✓	✓
All	$\mathtt{WOD}^{\text{train}}_{\text{all}}$	149,949	~	~	~	~
	$\mathtt{WOD}^{\text{val}}_{\text{all}}$	28,587	~	~	~	~
	$\mathtt{WOD}_{\text{all}}$	178,536	~	~	~	~

C. Training

Our model, LIF-Net, is trained on four NVIDIA H100 GPUs for 1000 epochs. We use the AdamW optimizer with an initial learning rate of 10^{-4} and a weight decay of 10^{-4}. We keep the learning rate constant throughout the training. We use a batch size of 64 per GPU. The training objective combines the 3D keypoint loss (L1), 2D keypoint loss (L1), and the SMPL parameter loss. The weights for these loss components were determined empirically through a series of experiments on the validation set, starting from values common in the literature [14], [15]. The used weights are set to $\lambda_{\text{3D}} = 5{*}10^{-2}$, $\lambda_{\text{2D}} = 5{*}10^{-3}$, and $\lambda_{\text{SMPL}} = 10^{-2}$. The λ_{SMPL} weight is further distributed internally to penalize the global orientation ($\lambda_{\text{global}} = 10^{-4}$), body pose ($\lambda_{\text{pose}} = 10^{-4}$), and shape parameters ($\lambda_{\text{betas}} = 5 * 10^{-5}$). For our ablation studies, the RGB-only baseline (LIF-Net$_{\text{RGB}}$) is trained with this same protocol. However, the LiDAR-only baseline (LIF-Net$_{\text{PC}}$) can only be trained with the $\mathcal{L}_{\text{SMPL}}$ loss on the $\mathtt{WOD}_{\text{2D3D}}$ split, as this is the only subset that has both the necessary LiDAR input and the image-derived pseudo-ground truth SMPL parameters.

D. Quantitative and Qualitative Evaluation

We present the primary quantitative results of our method, LIF-Net, on the WOD validation split as MPJPE and pa-MPJPE, compared to several other pose and joint estimation methods. Table III shows the quantitative results our model achieves on the WOD validation split as MPJPE and PA-MPJPE, compared to several other pose estimation methods, and improves the performance evaluated on MPJPE by 35.5% and 9% compared to the image-only and LiDAR-only models, respectively. The methods LiDARHMR [8], 4D-Humans [14], and TokenHMR [15] estimate complete SMPL parameter sets using LiDAR-only or camera-only approaches. 4D-Humans and TokenHMR initialized with weights provided by the authors have been fine-tuned on the used training set using the losses introduced to train LIF. LiDAR-HMR was not fine-tuned as we used the provided WaymoV2 checkpoint. One can clearly see that the placement of the meshes in 3D space benefits from our proposed fusion approach compared to the camera-only SMPL estimation methods, evident in the significantly reduced MPJPE metrics. LiDARHMR shows a superior performance in PA-MPJPE, which means that it learns the spatial relationships between joints better, but places the mesh worse than LIF-Net in 3D space, evident in its worse MPJPE value. Figure 3 illustrates representative qualitative results that highlight the diversity and complexity of the Waymo Open Dataset and demonstrate our model's robustness across challenging scenarios.

TABLE III: Quantitative comparison in MPJPE and PA-MPJPE of 3DHPS methods on $\texttt{WOD}^{\text{val}}_{\text{2D3D}}$. The results clearly show the impact of fusing camera and LiDAR modalities for pose estimation in MPJPE.

Method	**Modality**	**MPJPE** ↓ [mm]	**PA-MPJPE** ↓ [mm]
4D-Humans [14]	C	189	96
TokenHMR [15]	C	202	72
LiDARHMR [8]*	L	134	**63**
LIF-Net (ours)[‡]	C + L	**122**	76

[‡] **MPJPE vs. PA-MPJPE Trade-off:** Our fusion approach excels at absolute 3D positioning (MPJPE), which is critical for AD safety, by leveraging LiDAR localization cues. This results in a favorable trade-off with a minor degradation in the optimized PA-MPJPE.

E. Ablation Studies

To isolate the contribution of each component of our model, we performed several ablation studies. These experiments were designed to quantify the performance impact of our key architectural choices, particularly our fusion mechanism.

First, we evaluate our model with different input settings: image-only, LiDAR-only, and fused input, to assess the impact of modality choice. As shown in Table IV, combining both modalities improves performance, highlighting the benefit of multimodal fusion. One can see that the LiDAR input has a bigger influence on the performance of the model as we evaluate the placement of joints in 3D space.

Second, we applied different architectures for the fusion component of the network to evaluate the effectiveness of our proposed cross-attention-based feature fusion. Specifically, we compare two simple fusion strategies for combining LiDAR and camera features: cross-attention and an MLP. Results in Table V show that cross-attention outperforms the much simpler MLP alternatives, indicating the benefit of explicitly modeling inter-modal dependencies during fusion.

Third, we assess different training strategies for multi-modal input. Specifically, we show ablation studies on our proposed strategy using pseudo-global orientation and modeling the input in a camera coordinate system instead of a vehicle coordinate system. Table VI summarizes the results achieved in these experiments. One can clearly see that applying the pseudo-orientation derived from the detection bounding boxes helps stabilize the pose estimations in 3D space by improving the quality of the generated pseudo-groundtruth SMPL parameters. The effect of computing everything in a camera-based coordinate system, in contrast to that, is barely visible.

Finally, we evaluate the performance versus efficiency of our model using four different ViT backbone sizes (Table VII). The results confirm a clear trade-off: while the lightweight ViT-S model is 13% faster than the ViT-H model, its accuracy is lower, with 14% and 24% increases in MPJPE and PA-MPJPE, respectively. This provides a range of options for balancing speed and performance depending on the application requirements.

TABLE IV: Ablation I: Training on different input modalities, evaluated on $\texttt{WOD}^{\text{val}}_{\text{2D3D}}$.

Input Modality	**MPJPE** ↓ [mm]	**PA-MPJPE** ↓ [mm]
Image only	180	152
LiDAR only	128	75
LiDAR + Image	122	76

TABLE V: Ablation II: Comparison of fusion strategies for multi-modal input, evaluated on $\texttt{WOD}^{\text{val}}_{\text{2D3D}}$.

Fusion Strategy	**MPJPE** ↓ [mm]	**PA-MPJPE** ↓ [mm]
Cross-Attention	122	76
MLP	152	89

TABLE VI: Ablation III: Comparison of training strategies for multi-modal input, evaluated on $\texttt{WOD}^{\text{val}}_{\text{2D3D}}$. We investigate the effects of pseudo global orientation and projecting into a camera coordinate system.

Orient	**Cam**	**MPJPE** ↓ [mm]	**PA-MPJPE** ↓ [mm]
-	-	130	75
✓	✓	128	76
✓	-	122	76
-	✓	128	75

V. Conclusion

This paper represents a novel LiDAR and image fusion architecture for 3D human mesh recovery by utilizing cross-attention feature-fusion. With this proposed method, we address the significant challenge of 3D human mesh recovery in autonomous driving scenarios, where unimodal approaches are fundamentally limited. Our extensive experiments on the Waymo Open Dataset demonstrate that fusing both LiDAR

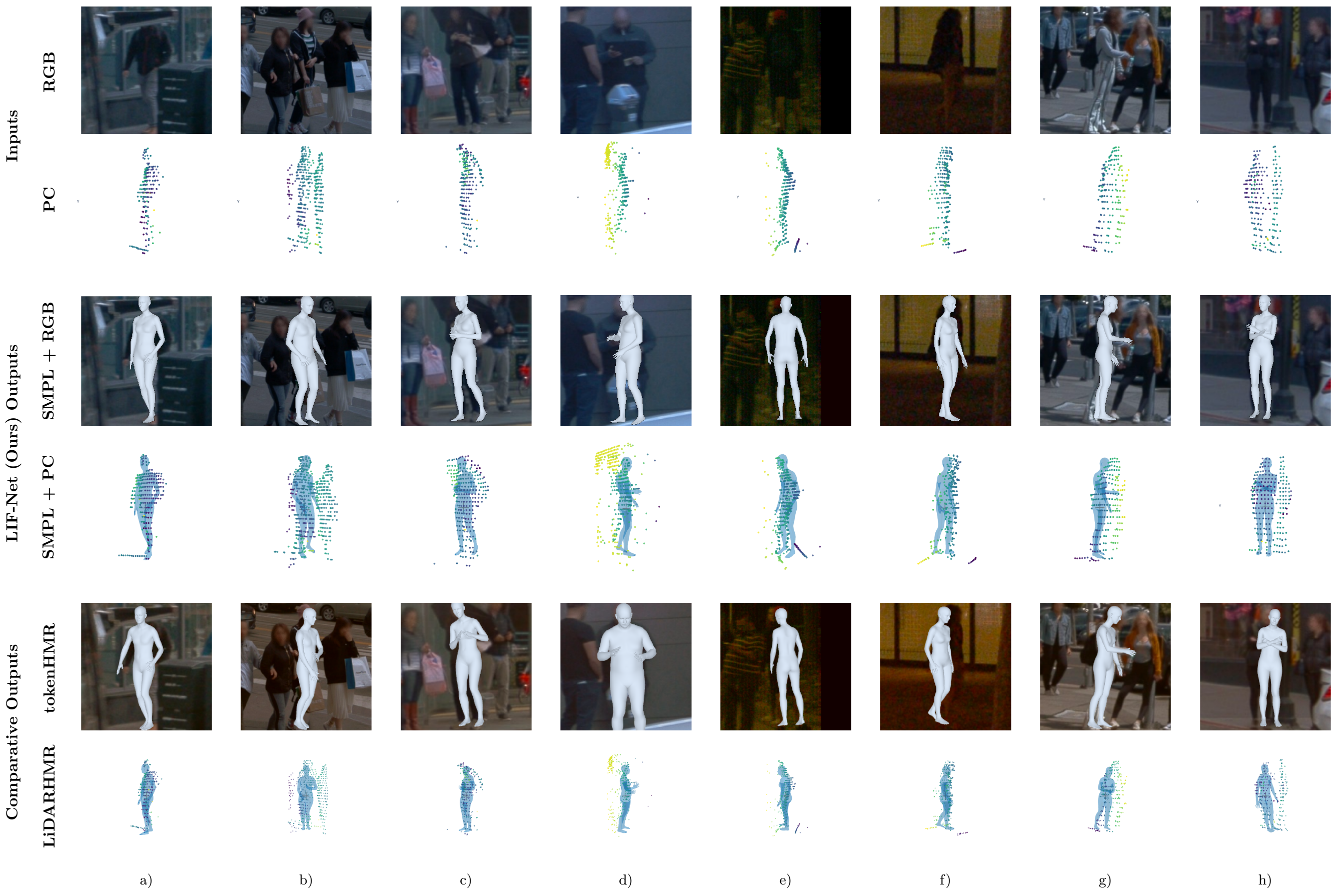

Fig. 3: We show multiple qualitative samples, highlighting both the diversity and complexity of WOD, and demonstrate our model's robustness to common failure modes in autonomous driving detections. The figure is structured as follows. Row one and two: Input of image and LiDAR modality. Row three and four: **LIF-Net outputs** overlaying the image and LiDAR inputs. Row five and six: **TokenHMR** and **LiDAR-HMR outputs** overlaying their respective input. The samples hold the following cases. *Sample a, b. Occlusion, multi-subject.* Bottom-up 3D joint estimators often fail to predict occluded parts, such as hidden heads or limbs overlapping with nearby pedestrians. A top-down SMPL approach mitigates this by leveraging body priors to reconstruct missing or occluded regions consistently. *Sample c, d. Large objects, annotation mismatch.* Despite low visual contrast between a black umbrella and clothing, all models can estimate the pose correctly. In contrast, inconsistent 2D/3D bounding boxes in the other sample cause fusion errors due to mismatched pedestrian regions, while single modality models perform well. *Sample e, f. Poor Lightning.* Image model accuracy drops significantly under adverse lighting, while fusion and LiDAR models remain robust. *Sample g, h. Point cloud noise.* Contaminated point clouds impair LiDAR estimation, but fusion and image models consistently maintain pose accuracy. Samples e, f, g, h highlight the robustness of our **LIF-Net**.

and RGB modalities consistently improves performance and robustness, leading to state-of-the-art results in these challenging real-world conditions. Specifically, our model achieves an improvement of approximately 35.5% over image-based methods and 9.1% over the LiDAR-based method in terms of MPJPE.

While our model shows strong performance, we acknowledge two key limitations: its reliance on the alignment between multi-modal ground-truth bounding boxes for subject localization, and its validation on a single, albeit large-scale, dataset. Future work will therefore focus on integrating a dedicated pedestrian detection module and on validating our approach across other multi-modal datasets to further assess its generalization capabilities.

Acknowledgment

This research was conducted within the project “Solutions and Technologies for Automated Driving in Town: An Urban Mobility Project”, funded by the Federal Ministry for

TABLE VII: Ablation IV: Comparison of model sizes through their accumulated parameters, with Cross-Attention as fusion strategy for multi-modal input, evaluated on $\texttt{WOD}^{\text{val}}_{\text{2D3D}}$. **IT**: Inference Time (ms) is the time taken for a forward pass.

Image Backbone	**ΣParams**	**IT** [ms]	**MPJPE** [mm]	**PA-MPJPE** [mm]
ViT-H	679M	95	122	76
ViT-L	353M	86	126	80
ViT-B	135M	84	129	85
ViT-S	70M	83	139	94

Economic Affairs and Climate Action (BMWK), based on a decision by the German Bundestag, grant no. 19A22006N.

References

[1] V. Kress, F. Jeske, S. Zernetsch, K. Doll, and B. Sick, "Pose and semantic map based probabilistic forecast of vulnerable road users' trajectories," *IEEE Intell. Veh*, vol. 8, pp. 2592–2603, 2023.

[2] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. ICRA*, 2023, pp. 2774–2781.

[3] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "TransFusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proc. IEEE CVPR*, 2022, pp. 1090–1099.

[4] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, A. Yuille, and M. Tan, "DeepFusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proc. IEEE CVPR*, Jun. 2022, pp. 17 161–17 170.

[5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[6] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE CVPR*, 2018, pp. 7122–7131.

[7] J. Li, J. Zhang, Z. Wang, S. Shen, C. Wen, Y. Ma, L. Xu, J. Yu, and C. Wang, "LiDARCap: Long-range marker-less 3d human motion capture with lidar point clouds," in *Proc. IEEE CVPR*, 2022, pp. 20 502–20 512.

[8] B. Fan, W. Zheng, J. Feng, and J. Zhou, "LiDAR-HMR: 3d human mesh recovery from lidar," *arXiv preprint arXiv:2311.11971*, 2023.

[9] Y. Dai, Y. Lin, X. Lin, C. Wen, L. Xu, H. Yi, S. Shen, Y. Ma, and C. Wang, "SLOPER4D: A scene-aware dataset for global 4d human pose estimation in urban environments," in *Proc. IEEE CVPR*, 2023, pp. 682–692.

[10] M. Yan, Y. Zhang, S. Cai, S. Fan, X. Lin, Y. Dai, S. Shen, C. Wen, L. Xu, Y. Ma *et al.*, "RELI11D: A comprehensive multimodal human motion dataset and method," in *Proc. IEEE CVPR*, 2024, pp. 2250–2262.

[11] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proc. IEEE CVPR*, 2019, pp. 10 975–10 985.

[12] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *Proc. IEEE ICCV*, 2019, pp. 2252–2261.

[13] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "Cliff: Carrying location information in full frames into human pose and shape estimation," in *Proc. of the ECCV*. Springer, 2022, pp. 590–606.

[14] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and tracking humans with transformers," in *Proc. IEEE ICCV*, 2023, pp. 14 783–14 794.

[15] S. K. Dwivedi, Y. Sun, P. Patel, Y. Feng, and M. J. Black, "TokenHMR: Advancing human mesh recovery with a tokenized pose representation," in *Proc. IEEE CVPR*, 2024, pp. 1323–1333.

[16] P. Patel and M. J. Black, "CameraHMR: Aligning people with perspective," *IEEE 3DV*, 2024.

[17] Y. Wang, Y. Sun, P. Patel, K. Daniilidis, M. J. Black, and M. Kocabas, "Prompthmr: Promptable human mesh recovery," in *Proc. IEEE CVPR*, 2025, pp. 1148–1159.

[18] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Adv. in NeurIPS*, 2017.

[19] Y. Dai, Y. Lin, C. Wen, S. Shen, L. Xu, J. Yu, Y. Ma, and C. Wang, "HSC4D: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar," in *Proc. IEEE CVPR*, 2022, pp. 6792–6802.

[20] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE TPAMI*, no. 7, pp. 1325–1339, 2013.

[21] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE CVPR*, 2014, pp. 3686–3693.

[22] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *Proc. of the ECCV*, 2018, pp. 601–617.

[23] M. Kaufmann, J. Song, C. Guo, K. Shen, T. Jiang, C. Tang, J. J. Zárate, and O. Hilliges, "EMDB: The electromagnetic database of global 3d human pose and shape in the wild," in *Proc. IEEE ICCV*, 2023, pp. 14 632–14 643.

[24] M. J. Black, P. Patel, J. Tesch, and J. Yang, "BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion," in *Proc. IEEE CVPR*, 2023, pp. 8726–8737.

[25] Y. Ren, X. Han, C. Zhao, J. Wang, L. Xu, J. Yu, and Y. Ma, "Livehps: Lidar-based scene-level human pose and shape estimation in free environment," in *Proc. IEEE CVPR*, 2024, pp. 1281–1291.

[26] W. Kim, M. S. Ramanagopal, C. Barto, M.-Y. Yu, K. Rosaen, N. Goumas, R. Vasudevan, and M. Johnson-Roberson, "PedX: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections," *IEEE RA-L*, no. 2, pp. 1940–1947, 2019.

[27] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo Open Dataset," in *Proc. IEEE CVPR*, 2020, pp. 2446–2454.

[28] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," in *Proc. IEEE CVPR*, 2018, pp. 244–253.

[29] P. Bauer, A. Bouazizi, U. Kressel, and F. Flohr, "Weakly supervised multi-modal 3d human body pose estimation for autonomous driving," in *IEEE Intell. Veh*, 2023, pp. 1–7.

[30] P. Cong, Y. Xu, Y. Ren, J. Zhang, L. Xu, J. Wang, J. Yu, and Y. Ma, "Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar," in *Proc. of the AAAI*, 2023, pp. 461–469.

[31] M. Fürst, S. T. Gupta, R. Schuster, O. Wasenmüller, and D. Stricker, "Hperl: 3d human pose estimation from rgb and lidar," in *Proc. ICPR*, 2021, pp. 7321–7327.

[32] A. Chen, X. Wang, K. Shi, S. Zhu, B. Fang, Y. Chen, J. Chen, Y. Huo, and Q. Ye, "Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions," in *Proc. ICRA*. IEEE, 2023, pp. 2752–2758.

[33] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, 2000.

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

[35] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Adv. in NeurIPS*, pp. 38 571–38 584, 2022.

[36] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao *et al.*, "Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion," in *Proc. IEEE CVPR*, 2023, pp. 17 524–17 534.

[37] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal attention for long-range interactions in vision transformers," in *Adv. in NeurIPS*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Curran Associates, Inc., 2021, pp. 30 008–30 022.

[38] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proc. IEEE CVPR*, 2018, pp. 4490–4499.

[39] J. Mao, M. Niu, H. Bai, X. Liang, H. Xu, and C. Xu, "Pyramid r-cnn: Towards better performance and adaptability for 3d object detection," in *Proc. IEEE CVPR*, 2021, pp. 2723–2732.