LIF-Net

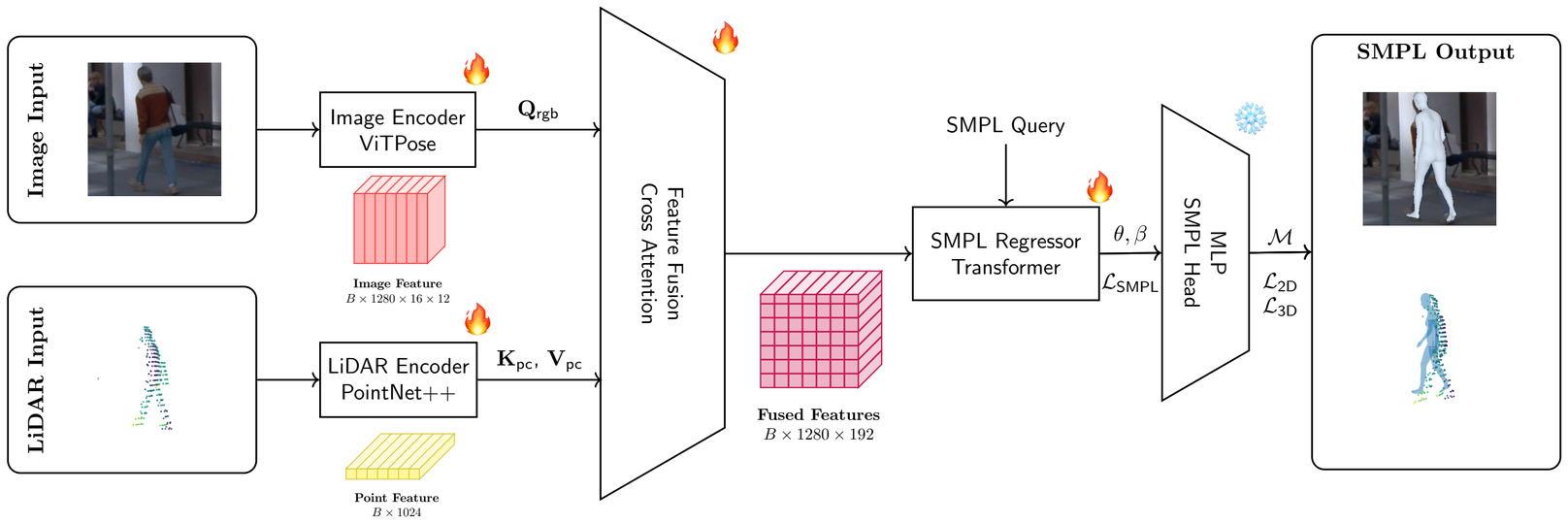
LiDAR–Camera Fusion for 3D Human Pose in Urban Scenes

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 | †Equal contribution | Project Page: https://iv.ee.hm.edu/lif-net/

Architecture Overview



A ViT-H/16 (ViTPose) RGB encoder and a PointNet++ LiDAR encoder feed a cross-attention fusion module; an HMR2-initialised transformer decoder regresses SMPL $(\boldsymbol{\theta}, \boldsymbol{\beta})$, supervised by 2D/3D keypoint and parameter losses. Modules marked with a fire icon are either trained or fine-tuned, while those marked with a snowflake icon are frozen.

−35.5%
MPJPE vs. image-only

−9.1%
MPJPE vs. LiDAR-only

C + L
cross-attention fusion

Motivation

- 3D human pose & shape (3DHPS) of vulnerable road users is a key cue for intent prediction and motion planning in autonomous driving.
- **Monocular RGB** lifts 2D→3D: ill-posed, so it suffers depth ambiguity and fails under occlusion and poor lighting.
- **LiDAR** gives direct geometry but is sparse and lacks semantic cues.
- The two failure modes are *complementary* — yet prior SMPL fusion (RELI11D / LEIR) was validated only indoors. No robust fusion exists for the in-the-wild AD domain.

Contributions

1. A novel **intermediate cross-attention fusion** architecture that exploits the complementary RGB / LiDAR failure modes for robust SMPL estimation.
2. Extensive multi-modal SMPL evaluation on the large-scale, in-the-wild **Waymo Open Dataset**.
3. State-of-the-art accuracy: **−35.5%** MPJPE vs. image-only and **−9.1%** vs. LiDAR-only.

Pseudo-Ground-Truth SMPL

WOD has no SMPL labels. We regress pseudo-GT with **TokenHMR** (discrete latent prior ⇒ anatomically plausible poses) and replace the global orientation with the one derived from the 3D bounding box, stabilising 3D placement.

Method

RGB encoder. ViT-H/16 (ViTPose), 256×256 crop → $f_{\text{img}} \in \mathbb{R}^{B\times 1280\times 16\times 12}$.

LiDAR encoder. PointNet++, fixed N_P=512 points → $f_{\text{LiDAR}} \in \mathbb{R}^{B\times 1024}$.

Decoder. HMR2-initialised transformer; an SMPL mean-init query attends to the fused context and predicts a residual on $(\boldsymbol{\theta}, \boldsymbol{\beta})$; SMPL stays frozen.

Cross-Attention Feature Fusion

Image features *query* the global LiDAR context (8 heads): RGB→ Q, LiDAR→ K, V. This dynamically reweights modalities and adds robustness to single-sensor degradation.

$$f_{\text{fusion}} = \text{softmax}\left(\frac{Q_{\text{rgb}} K_{\text{pc}}^{\top}}{\sqrt{d_k}}\right) V_{\text{pc}} \in \mathbb{R}^{B\times 192\times 1280}$$

Training Objective

End-to-end weighted sum of keypoint and parameter losses:

$$\mathcal{L}_{\text{3D}} = \|\hat{y} - y\|_1, \qquad \mathcal{L}_{\text{2D}} = \|\pi(\hat{y}) - y\|_1,$$
$$\mathcal{L}_{\text{SMPL}} = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2,$$
$$\mathcal{L}_{\text{tot}} = \lambda_{\text{3D}}\mathcal{L}_{\text{3D}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \lambda_{\text{SMPL}}\mathcal{L}_{\text{SMPL}}$$

Setup: 4× NVIDIA H100, 1000 epochs, AdamW, lr 10^{-4}, batch 64/GPU; ∼179k WOD pedestrian samples.

Results - Waymo Open Dataset

Method	Mod.	MPJPE↓	PA-MPJPE↓
4D-Humans	C	189	96
TokenHMR	C	202	72
LiDAR-HMR	L	134	**63**
LIF-Net (ours)	C+L	**122**	76

Values in mm, lower is better. Fusion wins on **absolute 3D localisation (MPJPE)** - the safety-critical metric - trading a small PA-MPJPE margin.

Ablations

I — Input modality

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

II — Fusion strategy

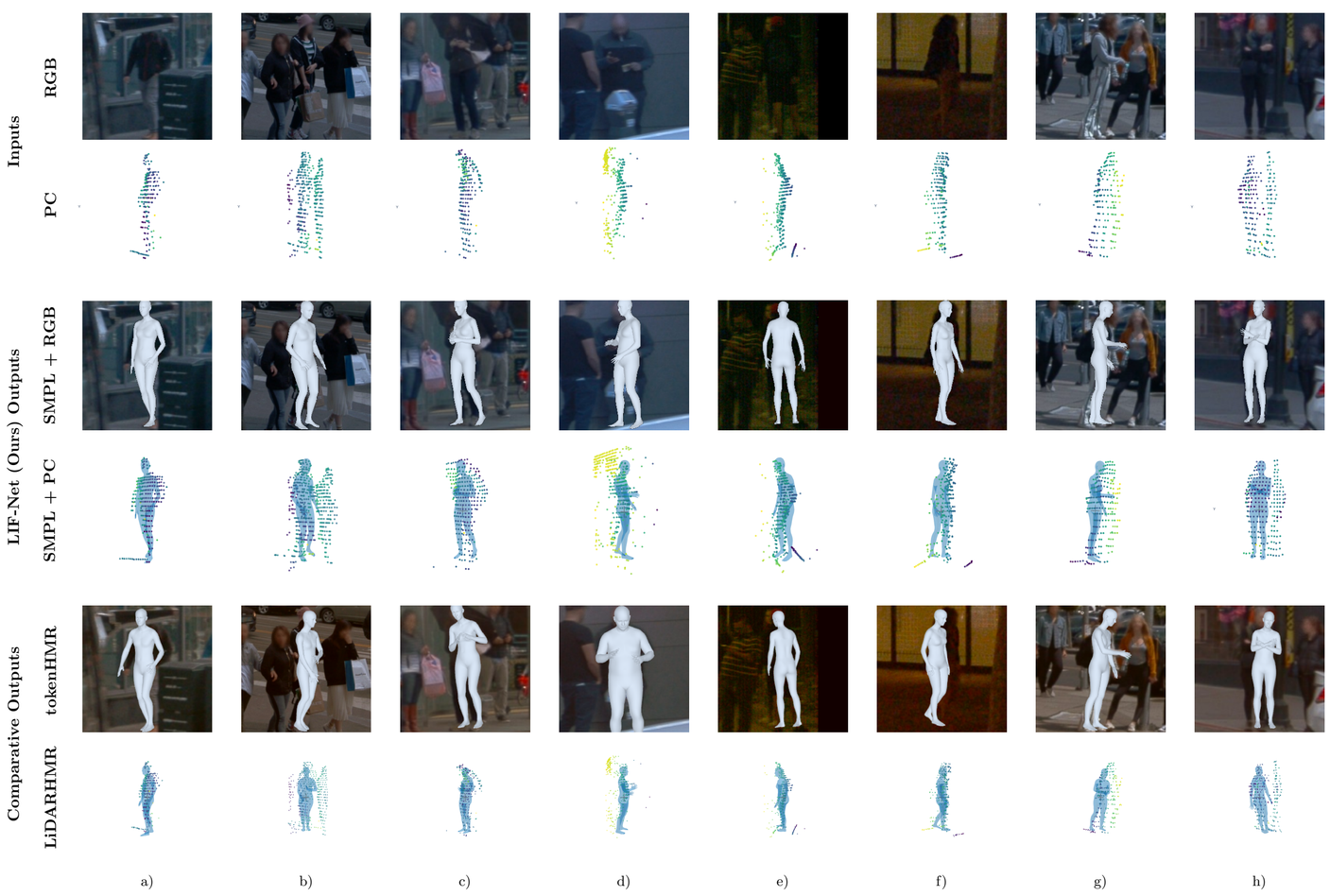
Strategy	MPJPE↓	PA-MPJPE↓
Cross-Attention	**122**	**76**
MLP	152	89

Backbone trade-off: ViT-S is 13% faster than ViT-H but +14% / +24% error (MPJPE / PA-MPJPE).

Conclusion

- Cross-attention RGB+LiDAR fusion yields **SOTA SMPL recovery** in AD, robust to occlusion, poor lighting and point-cloud noise.
- **Limits:** relies on aligned multi-modal bounding boxes; single dataset.
- **Future:** dedicated pedestrian detector; cross-dataset generalisation.

Qualitative Results - Robustness to AD Failure Modes



Occlusion & multi-subject (a,b) | low contrast / bbox mismatch (c,d) | poor lighting (e,f) | point-cloud noise (g,h). Fusion stays accurate where the weak modality fails.

Selected References

[1] Loper et al., SMPL, SIGGRAPH Asia'15
[2] Kanazawa et al., HMR, CVPR'18
[3] Goel et al., 4D-Humans (HMR2.0), ICCV'23
[4] Dwivedi et al., TokenHMR, CVPR'24
[5] Li et al., LiDARCap, CVPR'22
[6] Fan et al., LiDAR-HMR'23
[7] Cong et al., RELI11D, CVPR'24
[8] Qi et al., PointNet++, NeurIPS'17
[9] Xu et al., ViTPose, NeurIPS'22
[10] Sun et al., Waymo Open Dataset, CVPR'20

Acknowledgment

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