RGB-LiDAR Fusion for 3D Human Pose Estimation

Shuqi Fan[1], XinCheng Lin[1], Youtong Shi[1], Chaoyan Zhang[1], Chenlu Lin[2]

[1]School of Informatics, Xiamen University
[2]Institute of Artificial Intelligence, Xiamen University
{23020231154181, 23020231154148, 23020231154157, 23320231154457, 36920231153214}@stu.xmu.edu.cn

Abstract

3D human pose estimation is a fundamental research area in computer vision with numerous applications. Traditional methods for human pose estimation primarily rely on RGB images, which have inherent limitations such as occlusions and lighting variations and cannot accurately complete large-scale and long-distance tasks. Complementary, LiDAR has great advantages in the field of 3D human pose estimation due to its small environmental impact and long capture distance, providing additional depth information that can enhance the accuracy and robustness of human pose estimation. However, combining RGB and LiDAR data for robust 3D human pose estimation is still an open challenge. This paper proposes a novel RGB-LiDAR fusion approach for human pose estimation that fuses information from both modalities to achieve improved accuracy and robustness. While existing methods combining RGB and LiDAR data focus on estimating joint positions, our proposed approach aims to estimate anatomically plausible human body meshes for more realistic and natural-looking results.

Introduction

Human pose estimation has been a prominent research topic in computer vision due to its wide range of applications, including action recognition, human-computer interaction, and virtual reality. Traditional methods primarily rely on RGB images to estimate the 2D or 3D human pose. DeepPose(Toshev and Szegedy 2014)applied CNNs in a cascaded regressor for 2D human pose estimation, whereas Tompson et al. (Tompson et al. 2015)predicted heatmaps for the joints instead of direct regression. VIBE(Kocabas, Athanasiou, and Black 2020)has designed an action discriminator consisting of GRU layer and self attention layer to determine whether the generated pose sequence is real. HuMoR(Rempe et al. 2021)proposed a generation model in the form of a CVAE for learning the distribution of potential pose transitions in motion sequences. However, RGB images alone often suffer from inherent limitations such as occlusions, lighting variations, ambiguity in depth perception, small capture range and short capture distance.

Recently, LiDAR has been applied to 3D human pose estimation due to its advantages of large capture range, long capture distance, and minimal environmental interference. A LiDAR sensor provides accurate depth information of a large-scale scene with a large effective range by emitting laser beams and measuring their time-of-flight or phase shift upon reflection, which can complement RGB data and enhance the accuracy and robustness of human pose estimation. The depth data can provide precise geometric measurements of the scene, enabling accurate localization of human body joints. These properties potentially allow capturing human motions under the long range setting in general lighting conditions, without suffering from the degraded artifacts of visual sensors.

By combining RGB and LiDAR data, we can leverage the strengths of both modalities: the rich texture and appearance information from RGB images and the accurate depth measurements from LiDAR. However, combining multi-modal information is not trivial. There are some works that combine RGB and LiDAR for human pose estimation, such as HPERL(Fürst et al. 2021), an end-to-end architecture for multi-person 3D pose estimation that fuses RGB images and LiDAR point clouds for superior precision. However, existing methods primarily focus on estimating joint positions instead of human body meshes which are more anatomically plausible, realistic and natural-looking. In this paper, we propose a novel RGB-LiDAR fusion approach for human pose estimation. The estimated result will be in the form of human body meshes. We fuses the information from both modalities to achieve improved accuracy and robustness.

Related work

3D Human Pose Estimation. There are considerable amount of prior works in 3D human pose estimation. There are two main classes of human pose estimation methods, the first of which is model based(Zanfir, Marinoiu, and Sminchisescu 2018; Kolotouros et al. 2019)and relies on statistical human body mesh models like SMPL(Loper et al. 2023). These methods do not estimate 3D pose directly, but instead regress the parameters of a statistical body model, which has built-in anatomical and kinematic constraints. This leads to more natural predictions, with body shape usually estimated as well, even for poses not encountered during training. The second class of methods is skeletonbased(Varol et al. 2018; Sun et al. 2018), where the 3D pose is represented by 3D joint positions and these are to be regressed or detected directly from the input. These methods have the advantage of usually being more accurate and faster, but they are not guaranteed to produce anatomically correct human skeletons (e.g. the left arm may be reconstructed with different length than the right arm). Our proposed model falls in the former category, as our goal is to offer anatomically plausible estimations which are more realistic and natural-looking.

RGB-Based HPE. RGB-based human pose estimation methods can be divided into optimization-based methods and regression-based methods. SMPLify(Bogo et al. 2016)is an optimization based pose estimation algorithm that utilizes neural networks to obtain 2D keypoints information from images, and generates a 3D human body model based on the human body parameterization model SMPL. By minimizing the objective function, the error between the projected 2D keypoints of the 3D human body model and the detected 2D keypoints is penalized, and the human body model parameters are iteratively optimized. HMR(Kanazawa et al. 2018)is a regression based attitude estimation algorithm that directly learns the mapping from 2D image pixels to 3D model parameters through deep neural networks. SPIN(Kolotouros et al. 2019)combines iterative optimization based methods with network regression based methods. In response to the occurrence of occlusion in images. PARE(Kocabas et al. 2021)visualizes the impact of local occlusion on the global pose. Through a partially guided attention mechanism, the visibility information of individual body parts is utilized, while the information of adjacent body parts is used to predict the occluded parts.

LiDAR-Based HPE. Due to the lack of depth information in image-based methods, traditional keypoint projection approaches often suffer from significant errors. Additionally, image-based methods are susceptible to variations in lighting conditions, leading to degraded performance in low-light environments. To address these problems, LiDAR is utilized to scan the human body and the surrounding scene, generating point clouds data that contains shape, pose, and depth information of the human body to estimate poses. However, due to the lack of ground-truth 3D human pose annotations paired with LiDAR data, there has not been a lot of works on 3D human pose estimation from LiDAR information. PedX consists of 5, 000 pairs of stereo images and LiDAR point clouds for pedestrian poses. The Waymo Open Dataset(Sun et al. 2020)has a similar amount of 3D annotations as PedX, but it features many more different environments. LiDARHuman26M(Li et al. 2022) consists of LiDAR point clouds, RGB videos, and IMU data. With the existence of these datasets.With the existence of these datasets, LiDARCap (Li et al. 2022) propose a three-stage pipeline consisting of a temporal encoder, an inverse kinematics solver, and an SMPL optimizer to improve pose estimation performance. GC-KPL(Weng et al. 2023)learns human 3D keypoints for in-the-wild point clouds without any manual keypoint annotations. LPFormer(Ye et al. 2023)is a complete two-stage top-down 3D human pose estimation framework that uses only LiDAR point cloud as input. LiDAR-based methods exhibits robustness against lighting variations, but their accuracy heavily relies on the precision of the captured point cloud data. Therefore, a better approach is to combine these two modalities for simulating human body poses.

Multi-Modal HPE Based on Images and Point Clouds. HPERL model trains on 2D groundtruth pose annotations and uses a reprojection loss for the 3D pose regression task. A multi-modal model(Zheng et al. 2022)uses 2D labels on RGB images as weak supervision, and creates pseudo ground-truth 3D joint positions from the projection of annotated 2D joints. HUM3DIL leverages RGB information with LiDAR points, by computing pixel-aligned multi-modal features with the 3D positions of the LiDAR signal. However, these methods primarily focus on estimating joint positions instead of human body meshes.

Proposed Solution

We propose a multi-modality baseline for human motion estimation. Given the synchronized LiDAR point clouds and RGB images that are captured by multiple sensors, the task of the baseline is to predict the 3D pose of the human in the world coordinate system.

As is depicted in Figure 1, our method consists of three major modules: feature extractors, the multimodal cross-attention model (MMCA), and SMPL-based inverse kinematics solver. For each modality, feature extractors are used to extract its features, which are fused through the MMCA modules to fully utilize the 3D geometric information of point clouds and the appearance information of RGB images. In the end, the fused features are fed into SMPL solver to obtain the estimated poses.

Feature Extraction

RGB Feature Extraction. To extract the corresponding feature for each RGB frame, we first project the point cloud of the body onto the image, which could determine the boundary of the point cloud in this frame. Then, the bounding box that corresponds to the human body can be obtained. We crop the image from the bounding box and feed the image into a RGB encoder (DINOv2 (Oquab et al. 2023)). The feature for RGB modality, $R_{f_{i-N}}$ is obtained.

LiDAR Feature Extraction. For the human body point clouds P^{W}_{i-N}, we extract its features $P_{f3d_{i-N}}$ by feeding the point clouds into a PointNet++ (Qi et al. 2017) and a GRU network.

MMCA

To automatically learn correspondence among two modalities and eliminate calibration sensitivity, we uses the cross-attention strategy. The LiDAR point clouds and the RGB images are fused using the multimodal cross-attention model (MMCA). We aims to effectively integrate the geometry information with appearance information, which allows a comprehensive integration of the features from different modalities.

The design of the MMCA is depicted in Figure 1. The LiDAR features $F_{3D_{i-N}}$ and the RGB features $F_{2D_{i-N}}$ are processed through a series of transformer encoders (Vaswani et al. 2017) and self-attention mechanisms. MMCA employs

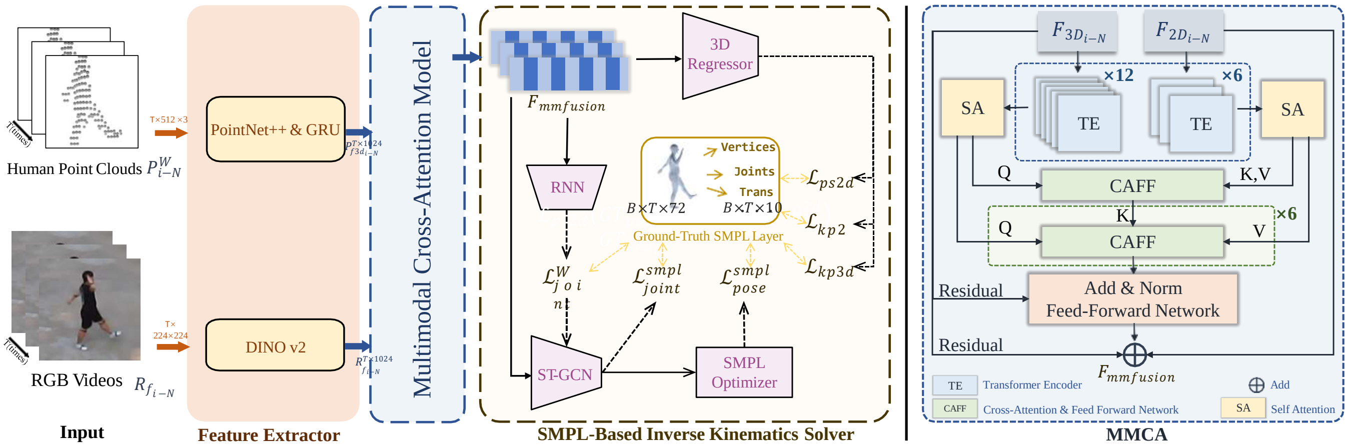

Figure 1: **Overview of our method (Left) and Multimodal Cross-Attention Unit (Right).** Orange arrows represent different modalities of data input. Dark blue arrows represent the inputs and outputs data flows of the MMCA model. Dotted arrows represent the predicted data and calculation loss with ground truth.

a 2-layer cross-attention structure, using the fused keys as intermediaries to match and align two sources of information. In the first layer, the features from the LiDAR act as queries, while the features from the RGB serve as keys and values. In the second layer, the output from the last layer serves as the keys; the LiDAR feature and RGB feature serve as query and value, respectively. The output features are obtained by element-wise addition of the input features and the results of the cross-attention structures.

MMCA is a flexible method which can use different combinations of modalities as input. For single-modality input, the extracted input features are fed into the MMCA whose cross-attention module is replaced with a self-attention module. For two-modality input, features from two modalities are fed into the MMCA.

SMPL-Based Inverse Motion Solver

The fused features $F_{mmfusion}$ obtained from MMCA, are used in this solver, which consists of three branches. In the first branch, the extracted features are inputted into a 3D regressor, which is responsible for estimating the 3D joints and camera intrinsic parameters. To guide the training and ensure accurate estimation, three loss functions are employed in this branch. The first loss function, $\mathcal{L}_{ps2d}$, serves as a projection loss, which ensures that the 2D appearance of the SMPL model aligns with the human body in pixel coordinates. By minimizing the discrepancy between the projected 2D model and the observed human body in the image, this loss function aids in achieving accurate alignment and pixel-level correspondence. This loss is defined as:

$$\mathcal{L}_{ps2d} = \mathcal{L}_{shape2d} + \mathcal{L}_{pose2d} \tag{1}$$

$\mathcal{L}_{ps2d}$ consists two terms that specifically target the pose and shape parameters of the SMPL model. The shape term β is the 10-dimensional shape parameter of the SMPL model. The pose term, θ, is a $N \times 3 \times 3$ rotation matrix, where N is 24 and represents the number of joint points. $L_{shape2d}$ and L_{pose2d} are defined as follows.

$$\mathcal{L}_{shape2d} = \frac{1}{10}\sum_{i=1}^{10}(\beta_{pred_i} - \beta_{gt_i})^2 \tag{2}$$

$$\mathcal{L}_{pose2d} = \frac{1}{N\times3\times3}\sum_{i=1}^{N}\sum_{j=1}^{3}\sum_{k=1}^{3}(\theta_{pred_{ijk}} - \theta_{gt_{ijk}})^2 \tag{3}$$

where β_{pred} and β_{gt} are the predicted and ground-truth shapes, respectively. θ_{pred} and θ_{gt} are the predicted and ground-truth poses, respectively.

The second loss function, $\mathcal{L}_{kp2d}$, is used to constrain the 2D joints of the human body. By comparing the estimated joints $KP2d_{pred}$ with the ground truth annotations $KP2d_{gt}$, this loss function encourages the regressor to accurately capture the spatial relationships and positions of the joints in the 2D image space.

$$\mathcal{L}_{kp2d} = \frac{1}{N\times2}\sum_{i=1}^{N}\sum_{j=1}^{2}(KP2d_{pred_{ij}} - KP2d_{gt_{ij}})^2 \tag{4}$$

The 3D joints predicted by the 3D regressor are constrained by the loss $\mathcal{L}_{kp3d}$, which ensures that the regressor accurately captures the spatial relationships and positions of the joints by comparing predicted joints $KP3d_{pred}$ with the ground truth annotations $KP3d_{gt}$.

$$\mathcal{L}_{kp3d} = \frac{1}{N\times3}\sum_{i=1}^{N}\sum_{j=1}^{3}(KP3d_{pred_{ij}} - KP3d_{gt_{ij}})^2 \tag{5}$$

In the second branch of the solver, the extracted features are fed into a RNN network, which is designed to generate the 3D human joints in the world coordinate system. To guide the training process and ensure appropriate joint prediction, we employ $\mathcal{L}^{W}_{joint}$ to encourage alignment between the predicted 3D joints Jt_{pred} and the ground truth labels Jt_{gt}.

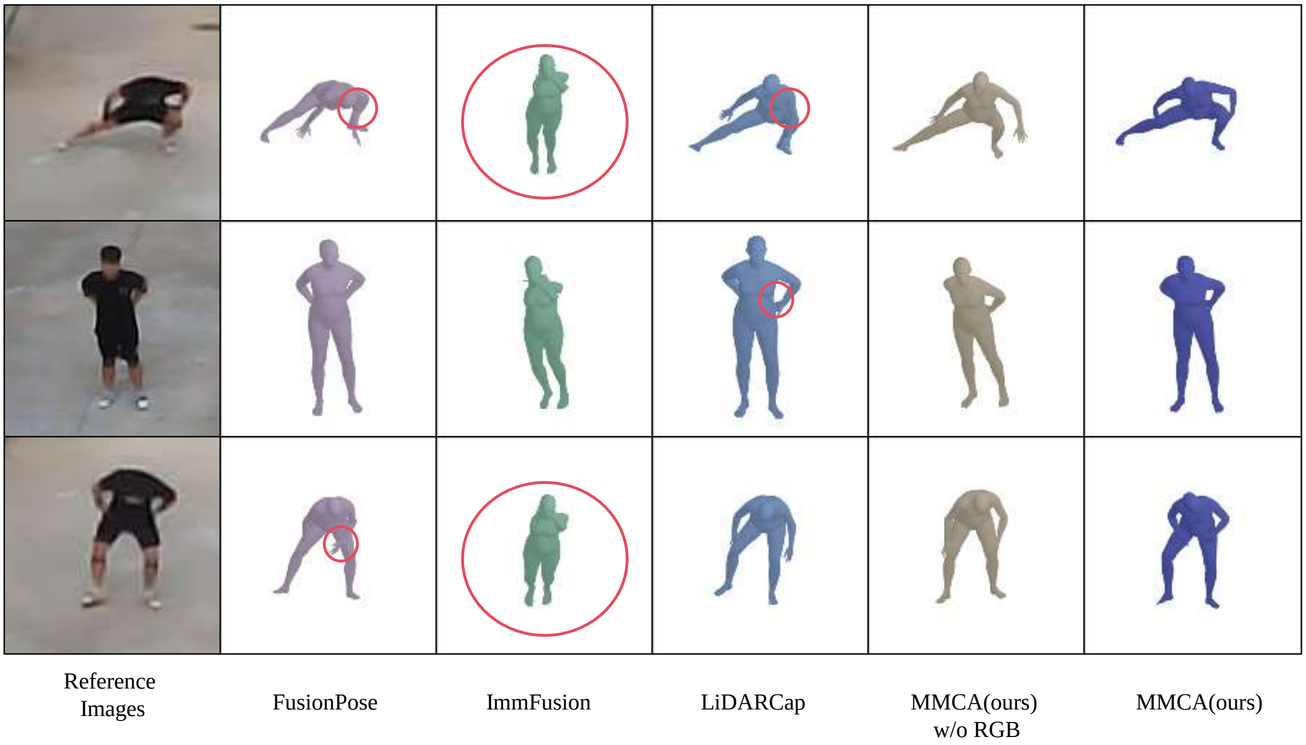

Figure 2: This figure presents the visualization results of different methods on LiDARHuman26M. The leftmost column is the reference image, while the remaining five columns depict the results of FusionPose, ImmFusion, LiDARCap, MMCA without RGB, and MMCA with full modalities, respectively. It can be observed that other methods exhibit significant estimation errors, particularly in capturing hand pose of individuals.

$$\mathcal{L}^{W}_{joint} = \frac{1}{N \times 3}\sum_{i=1}^{N}\sum_{j=1}^{3}(Jt_{pred_{ij}} - Jt_{gt_{ij}})^2 \tag{6}$$

The third branch of our approach employs ST-GCN (Yan, Xiong, and Lin 2018), where the fused features from the previous branches are utilized to predict the 3D human joints. To ensure accurate joint orientation, we apply $\mathcal{L}^{smpl}_{joint}$ to encourage alignment between the predicted joint orientations Jt_{pred} and the ground truth orientations Jt_{gt}.

$$\mathcal{L}^{smpl}_{joint} = \frac{1}{N \times 3}\sum_{i=1}^{N}\sum_{j=1}^{3}(Jt^{smpl}_{pred_{ij}} - Jt^{smpl}_{gt_{ij}})^2 \tag{7}$$

The outputs of this branch, which represent the predicted 3D joints, are then passed through a SMPL optimizer that converts the joint positions into human poses in axis-angle form. And the loss $\mathcal{L}^{smpl}_{pose}$ is employed to enforce alignment between predicted pose θ^{smpl}_{pred} with the ground truth poses θ^{smpl}_{gt}.

$$\mathcal{L}^{smpl}_{pose} = \frac{1}{N \times 3 \times 3}\sum_{i=1}^{N}\sum_{j=1}^{3}\sum_{k=1}^{3}(\theta^{smpl}_{pred_{ijk}} - \theta^{smpl}_{gt_{ijk}})^2 \tag{8}$$

All the aforementioned losses play crucial roles in our method to achieve accurate estimates of human pose.

Experiments

Implementation Details. We use two RGB-Based methods to estimate human posture. First, we test HMR. we employs a pre-trained network to regress pose, shape and camera parameters. Second, we use VIBE, which first extracts the spatial features of the image through Resnet50, and then processs the sequence through GRU to learn its temporal features, finally obtain 82 SMPL parameters through the regression layer.

For LiDAR-Based method, we conduct experiments using LiDARCap (Li et al. 2022). This method includes the extraction of point cloud features and the solution of human posture. The first step is to process the point cloud through PointNet++(Qi et al. 2017) and extract the global descriptor. Then, a temporal encoder implemented using GRU is used to fuse the temporal information in consecutive frames. Next, an MLP decoder is used to predict the positions of human joints based on the fused features. The predicted joint positions are combined with the 1024-dimensional features and fed into ST-GCN, which computes the predicted pose parameters. Finally, joint positions are calculated in the SMPL optimizer, resulting in predicted human body information, including pose and other relevant details.

Input Modality	Method	ACCEL↓	MPJPE↓	PA-MPJPE↓	PVE↓	PCK0.3↑
RGB	HMR (Li et al. 2022)	220.07	224.86	130.71	284.15	0.49
	VIBE (Kocabas et al. 2020)	120.49	154.61	108.19	191.55	0.82
LiDAR	LiDARCap (Li et al. 2022)	45.89	80.08	67.50	102.24	0.85
	MMCA(Ours)	**45.60**	**79.00**	**67.45**	**100.87**	**0.85**
LiDAR+RGB	ImmFusion (Chen et al. 2023)	46.45	96.93	81.16	107.29	0.75
	FusionPose (Cong et al. 2023)	44.51	78.18	66.70	99.66	0.85
	MMCA(Ours)	**44.52**	**75.09**	**62.94**	**95.96**	**0.87**

Table 1: Performance evaluation of MMCA in the LiDARHuman26M dataset (Li et al. 2022). Unit: mm

We use two methods in experiments to predict human posture based on point clouds and RGB images. The first is ImmFusion (Chen et al. 2023), which can be divided into three main parts. In the first part, images and point clouds are passed through the modal masking module. In the second part, PointNet++(Qi et al. 2017) is used to extract features from the point cloud. For images, HrNet is used to extract local mesh features, which are then converted into global features using CNN. At the same time, the local grid features are adjusted by MLP to match the size of the local cluster features of the point cloud. The global features of the two modalities are fused together using a small Transformer module, and the template's vertices and joints are also incorporated into this fusion process. The last part combines the features to obtain the fusion features through the Fusion Transformer Module. Finally, the fused features are input into SMPL to obtain human pose.

The second is FusionPose (Cong et al. 2023), which is a method that combines 3D point clouds and 2D perspective RGB images for human pose prediction. In their proposed IPAFusion approach, a fusion technique is introduced to effectively combine the two modes. To extract features from point clouds, they adopt PointNet, which provides global feature representation. This global feature is then combined with the original feature to obtain the final feature. Similarly, for image modality, HrNet is selected to extract features. This process produces the final feature. In the image-to-point attention fusion part, they utilize cross-attention to fuse two features. Finally, the fused features are input into GRU+MLP to obtain the predicted SMPL human pose.

Evaluation Metrics. We report Mean Per Joint Position Error (MPJPE), Procrustes Aligned Mean Per Joint Position Error (PA-MPJPE), Percentage of Correct Keypoints (PCK0.3), Per Vertex Error (PVE), Acceleration Error(mm/s^2) (ACCEL). The PCK0.3 is calculated as a percentage, while other indicators are in mm.

Dataset: LiDARHuman26M. In this paper, we utilize the first long-range LiDAR-based motion capture dataset, LiDARHuman26M. The dataset is collected from two independent scenes: one in a patio and the other in an open spaces within two buildings. Thirteen volunteers, including 11 males and two females, are recruited for the data collection process. Each participant was involved for a duration ranging from 15 to 30 minutes.LiDARHuman26M provides 184, 048 frames, 26, 414, 383 points, and 20 kinds of daily motions(including walking, swimming, running, phoning, bowing, etc). It consists of three modalities: synchronous LiDARpoint clouds, RGB images, and ground-truth 3D human motions from professional IMU-based mocap devices. We preprocessed the data by erasing the background and eliminating the localization error of the IMUs.

Comparison Experiments. We evaluate the proposed method, MMCA, based on the LiDARHuman26M (Li et al. 2022) dataset, which contains both RGB and LiDAR modalities. In this experiment, we follow the same evaluation as LiDARCap(Li et al. 2022). As is shown in Table 1, benefiting from the effective use of 3D spatial information, LiDAR-Based methods and LiDAR+RGB-Based methods significantly outperforms RGB-Based mothods. Due to the long collection distance of the LiDARHuman26M dataset, RGB images provide less useful information then LiDAR. When considering LiDAR input alone, MMCA demonstrates a slight improvement over LiDARCap. When combining LiDAR and RGB inputs, MMCA out-performs ImmFusion (Chen et al. 2023) and FusionPose (Cong et al. 2023). Using two modalities (RGB+LiDAR) leads to better performance than using these two modalities separately. As the number of modalities increases, the performance of MMCA improves. This emphasizes that correctly fusing the information of each modality feature makes the method robust.

Conclusion

In conclusion, we have proposed a human pose estimation method based on the multi-modal cross-attention mechanism. This approach effectively extracts and integrates features from both RGB images and point clouds, enabling the prediction of human joint information. Through our proposed method, we address issues present in traditional single-modal RGB image methods, such as spatial ambiguity and sensitivity to lighting changes, as well as the insufficient texture information in single-modal point cloud methods. Finally, we utilize the SMPL model for optimization fitting to obtain the final human body mesh model and visualize the results. The experimental errors obtained using our proposed method are consistently lower than those of any single-modal approach, providing evidence of the effectiveness of our method.We hope that the general techniques introduced here will be valuable to other researchers dedicated to improving the performance of human pose estimation models.

References

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 561–578. Springer.

Chen, A.; Wang, X.; Shi, K.; Zhu, S.; Fang, B.; Chen, Y.; Chen, J.; Huo, Y.; and Ye, Q. 2023. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2752–2758. IEEE.

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. In *AAAI*, 461–469. AAAI Press.

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.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kim, W.; Ramanagopal, M. S.; Barto, C.; Yu, M.-Y.; Rosaen, K.; Goumas, N.; Vasudevan, R.; and Johnson-Roberson, M. 2019. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4(2): 1940–1947.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5253–5263.

Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11127–11137.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2252–2261.

Li, J.; Zhang, J.; Wang, Z.; Shen, S.; Wen, C.; Ma, Y.; Xu, L.; Yu, J.; and Wang, C. 2022. Lidarcap: Long-range markerless 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20502–20512.

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

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Rempe, D.; Birdal, T.; Hertzmann, A.; Yang, J.; Sridhar, S.; and Guibas, L. J. 2021. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11488–11499.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Sun, X.; Xiao, B.; Wei, F.; Liang, S.; and Wei, Y. 2018. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, 529–545.

Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; and Bregler, C. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 648–656.

Toshev, A.; and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1653–1660.

Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; and Schmid, C. 2018. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European conference on computer vision (ECCV)*, 20–36.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Weng, Z.; Gorban, A. S.; Ji, J.; Najibi, M.; Zhou, Y.; and Anguelov, D. 2023. 3D Human Keypoints Estimation From Point Clouds in the Wild Without Human Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1158–1167.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ye, D.; Xie, Y.; Chen, W.; Zhou, Z.; and Foroosh, H. 2023. LPFormer: LiDAR Pose Estimation Transformer with Multi-Task Network. *arXiv preprint arXiv:2306.12525*.

Zanfir, A.; Marinoiu, E.; and Sminchisescu, C. 2018. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2148–2157.

Zheng, J.; Shi, X.; Gorban, A.; Mao, J.; Song, Y.; Qi, C. R.; Liu, T.; Chari, V.; Cornman, A.; Zhou, Y.; et al. 2022. Multimodal 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.