FTP: A Human Pose Estimation Method Integrating Temporal and Fine-Grained Feature Fusion

Shuqiang Cai SHUQIANGCAI@STU.XMU.EDU.CN
Chennan Ma CHENNANMA@STU.XMU.EDU.CN
Xin Wang 23020211153970@STU.XMU.EDU.CN
Li Lin LINLI1210@STU.XMU.EDU.CN
Ming Yan YANMNN@STU.XMU.EDU.CN
Xinchen Lin XINCHENLIN@STU.XMU.EDU.CN
Shuqi Fan FANSHUQI@STU.XMU.EDU.CN
Siqi Shen* SIQISHEN@XMU.EDU.CN
Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, Fujian, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Human pose estimation is a significant research direction in the field of computer vision, with critical applications in human motion reconstruction and analysis. Currently proposed human pose estimation methods primarily focus on single-modality sensor information, such as RGB images and LiDAR point clouds. While these methods have achieved promising results within their respective domains, they remain limited by the inherent deficiencies of each modality, hindering their applicability across diverse real-world scenarios. With the recent introduction of numerous multi-modality human pose datasets, multi-modality approaches have begun to develop. However, existing multi-modality fusion methods mainly consider the global feature relationships between different modalities, without modeling finer-grained features or the dynamic temporal relationships between modalities. To address this issue, we propose a novel pipeline that integrates point cloud and image features, explicitly encoding fine-grained features and dynamic temporal relationships between the two modalities. Additionally, we employ a discriminator structure for semi-supervised training. Extensive experiments demonstrate that our method achieves state-of-the-art (SOTA) performance compared to previous methods.

Keywords: Human Pose Estimation and Multi-modality.

1. Introduction

Human pose estimation (HPE) has become a highly popular research field, with numerous downstream applications in augmented/virtual reality (AR/VR), robotic simulation, and human motion analysis. As research progresses, more attention is being focused on 3D Human Pose Estimation in real-world scenarios, which requires precise depth estimation and keypoint detection.

Previous HPE methods primarily used RGB images captured by monocular cameras as input. Image data provides rich texture and color information, which aids in capturing surface details of the human body. However, their sensitivity to brightness and lack of

* Corresponding author

depth information limits the effectiveness in recovering human poses in a global 3D space. Recently proposed LiDAR point cloud-based human pose estimation methods Li et al. (2022) can recover human poses at long distances and under low light conditions. Point cloud data offers precise 3D spatial information, helping network to better understand the spatial relationships of human structures and movements, especially excelling in depth estimation and volumetric perception. Nevertheless, point cloud information is highly abstract and lacks sufficient surface details, resulting in poor performance in capturing fine details of human poses. Therefore, it is essential to propose a novel method that can leverage the advantages of different modalities simultaneously.

To address the shortcomings of single-modality human pose estimation methods, researchers have proposed several multi-modality approaches. These methods utilize both the 3D spatial information from point clouds and the appearance information from images to reconstruct more accurate and reasonable human poses. ImmFusionChen et al. (2023) and FusionPoseCong et al. (2023) have developed several techniques for implementing multi-modality fusion, but they only consider temporal fusion within each modality independently, neglecting the dynamic relationships between the two modalities in both the temporal and feature dimensions. LEIRYan et al. (2024) is a pipeline capable of fusing and inferring point cloud, image, and event features simultaneously, but it also encodes the temporal information of each modality independently and does not model the deep relationships between their fine-grained features. Therefore, existing multi-modality human pose estimation algorithms still have certain shortcomings.

To overcome the limitations of previous methods, we propose a novel multi-modality human pose estimation pipeline named FTP. This method integrates dynamic features in both the temporal and fine-grained feature dimensions, effectively utilizing the visual information from images and the 3D spatial information from point clouds to make accurate predictions of human poses. By encoding features in both temporal and fine-grained dimensions, the model can better understand the relationships between different modalities, leveraging their respective strengths for precise human pose estimation. Additionally, we introduce a discriminator architecture that guides our human pose estimation network to learn priors of real human poses. We train the discriminator using existing large-scale human datasets Mahmood et al. (2019), employing generative adversarial concepts to improve the prediction quality and training effectiveness of the network. Extensive experiments demonstrate that our algorithm achieves state-of-the-art (SOTA) performance in the field of multi-modality human pose estimation using point clouds and images. We also conducted ablation studies to verify the effectiveness of each module.

In summary, our contributions are as follows:

- We designed a feature fusion block named parallel TFCA which is capable of encoding both temporal and fine-grained feature information in parallel.
- We introduce a discriminator architecture that leverages generative adversarial concepts to enhance the overall performance of the network.
- We propose FineTemporalPose (FTP), a multimodal pipeline based on **F**ine-grained and **T**emporal feature fusion for human **P**ose estimation. We conduct extensive experiments to demonstrate the superiority of our network and the necessity of its individual modules.

2. Related Work

2.1. Camera-based 3D Human Pose Estimation

Monocular camera-based human pose estimation has become a highly popular research direction. Early studies primarily focused on detecting and locating human keypoints Martinez et al. (2017); Tome et al. (2017); Cao et al. (2017); Pavlakos et al. (2017), both in 2D and 3D spaces. Subsequently, SMPL Loper et al. (2015) and SMPL-X Pavlakos et al. (2019) provided differentiable parameterized body models for the field. These models have spurred increased adoption of parameterized approaches Kanazawa et al. (2018); Kocabas et al. (2020); Bogo et al. (2016); Kolotouros et al. (2019a,b), for recovering and representing human poses. The advent of parameterized models also made it possible to create unified representations of large-scale human pose datasets. AMASS Mahmood et al. (2019), for example, is a result of integrating numerous past human datasets into a single unified format.Building on this foundation, methods like HMR Kanazawa et al. (2018) and VIBE Kocabas et al. (2020) have proposed using existing large-scale human datasets to learn human priors. These methods primarily rely on a discriminator structure, which can partially address the lack of depth information in images.

As research progresses, the focus has gradually shifted from analyzing individual frames Mehta et al. (2017); Güler et al. (2018); Sun et al. (2018); Xiao et al. (2018) to recovering human poses across entire video sequences Li et al. (2021, 2023); Kocabas et al. (2021); Wan et al. (2021); Hassan et al. (2019). This transition has brought greater attention to the temporal relationships between frames. However, performance of most methods in outdoor or large scenarios remains unstable. This instability is primarily due to the ambiguous depth estimation inherent in image data.

2.2. LiDAR-based 3D Human Pose Estimation

In recent years, LiDAR-based perception methods have experienced significant advancements. Due to LiDAR's accurate depth measurement in large-scale scenes, its application in autonomous driving and 3D perception Zhu et al. (2020); Yin et al. (2021); Han et al. (2022) has become increasingly widespread. Consequently, many researchers have begun applying LiDAR point cloud information to human pose estimation field. For instance, P4T Fan et al. (2021) and STCrowd Cong et al. (2022) use point clouds for human semantic segmentation, while LiDARCap Li et al. (2022) has proposed a point cloud-based human pose estimation pipeline capable of accurately estimating human poses in long-range scenarios.

However, LiDAR point clouds are sparse and unordered, and are susceptible to noise disturbances. Additionally, they lack the rich texture details that images provide. This is why multiple modalities are necessary for more accurately predicting human poses.

2.3. Multi-modality Fusion Approaches

With the advancement in point cloud processing, more multi-modality human pose datasets incorporating point clouds have been introduced. Datasets such as LiDARHuman26M Li et al. (2022), HSC4D Dai et al. (2022), SLOPER4D Dai et al. (2023), CIMI4D Yan et al. (2023), and RELI11D Yan et al. (2024) provide accurate human labels and include both point cloud and RGB image, each with distinct application scenarios. Based on these datasets,

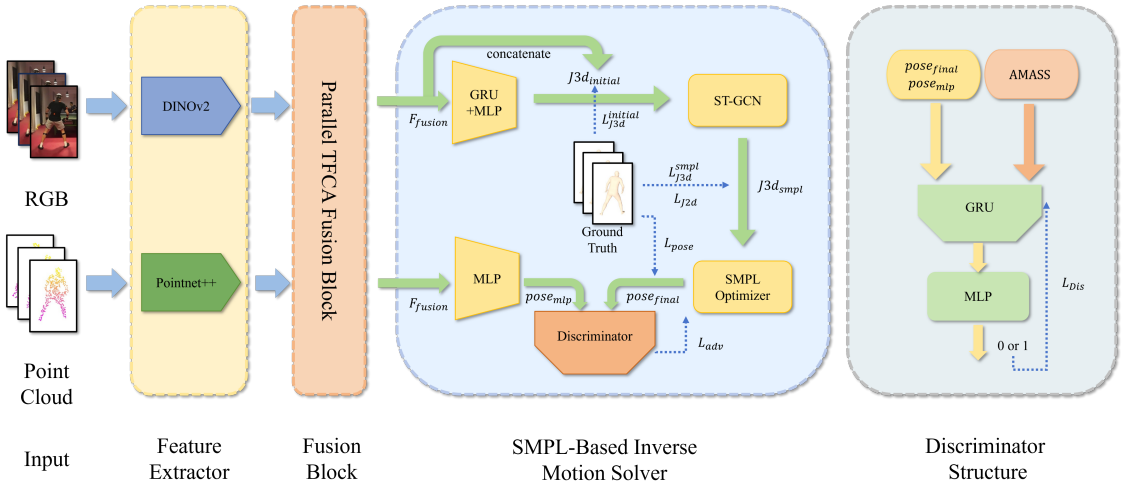


Figure 1: Overview of FTP Pipeline. FTP consists of four modules: the feature extractor, the parallel TFCA fusion block, the SMPL-based inverse motion solver and the discriminator.

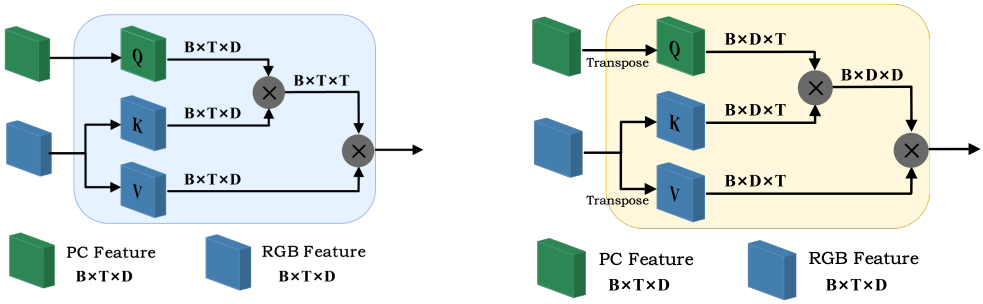
research on multi-modality human pose estimation methods has also made notable progress. Immfusion Chen et al. (2023) is a method that uses mmwave radar point clouds and camera images to recover human poses in various environments. Although mmwave radar point clouds differs from LiDAR point clouds, it remains a valuable reference. FusionPose Cong et al. (2023) proposes a method for reconstructing human poses in multi-person scenarios using point clouds and images, but it only predicts joint positions without using a parametric model, leading to reduced performance in more detailed single-person scenarios. LEIR Yan et al. (2024) simultaneously fuses RGB, point cloud, and event for human pose estimation, achieving good results. However, it does not focus on the fine-grained feature relationships between modalities, indicating potential for improvement in some scenarios.

3. Method

3.1. overview

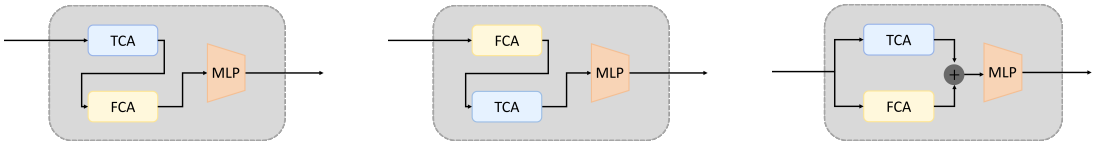
Previous multi-modality fusion methods often perform fusion solely at the feature level or just encode the temporal information of each modality independently. These approaches do not fully exploit the advantages of multi-modality data. Our goal is to integrate both the dynamic temporal information and fine-grained features of different modalities during the fusion process, thereby better combining the unique strengths of images and point clouds.

We propose FTP, a novel multi-modality baseline for human pose estimation. It's an end-to-end, weakly supervised, multi-modality fusion method. The whole pipeline is shown in the Figure 1. It employs feature extractors to process time-synchronized point clouds and images. The extracted features are then input into our designed parallel TFCA fusion block to obtain fused features. Finally, an inverse kinematics solver is used to derive the final human pose. Additionally, we employ a specially designed discriminator for generative adversarial training to accelerate the training process.



(*a*) TCA (*b*) FCA

Figure 2: Structure of TCA and FCA



(*a*) Temporal-first (*b*) Fine-grained feature-first (*c*) parallel TFCA

Figure 3: Feature Fusion Strategies

3.2. Feature Extraction

3.2.1. RGB Feature Extraction

For RGB frames, we specifically utilize a pre-trained DINOv2 network for feature extraction. The extracted image features are denoted as F_{rgb}.

3.2.2. LiDAR Feature Extraction

For point cloud frames, we first input them into a PointNet++ network to extract spatial features. These features are then processed through a GRU network to obtain spatiotemporal encoded point cloud features F_{pc}.

3.3. Fusion Process

We draw inspiration from MAED Wan et al. (2021) and propose two cross-attention mechanism-based blocks : Temporal Cross-Attention (TCA) and Feature Cross-Attention (FCA). Figure 2 shows architecture of blocks. The TCA block focuses primarily on encoding temporal relationships across different modalities, while the FCA block focuses on extracting deep relationships across the feature dimensions of different modalities.

3.3.1. Temporal Cross-Attention (TCA)

We use the point cloud features F_{pc} extracted by the feature extractor to compute the query Q_{pc}, while the image features F_{rgb} are used to compute the key K_{rgb} and value V_{rgb}. The

new features are then obtained using the following attention formula:

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

F_{tca} integrates both the spatial information of the point cloud and the appearance information of the image. Through this way we explicitly model the feature relationships between the point cloud and the image, enabling the model to establish dynamic correspondences between point cloud and image features in the temporal dimension. With Applying attention mechanism in the temporal dimension, the model can understand how different modality features change over time, which is particularly important for processing sequential data such as videos and dynamic point cloud data.

3.3.2. Feature Cross-Attention (FCA)

FCA applies the cross-attention mechanism to each feature dimension of multi-modality data, helping the model capture hidden interaction and deep dependency between different modalities. Similar to TCA, it extract Q_{pc}, K_{rgb} and V_{rgb} from the computed features. FCA then performs a matrix transpose operation on the extracted high-dimensional features, swapping the time and feature dimensions. The cross-attention mechanism is then applied to each feature dimension of the multi-modality data, helping the model capture fine-grained feature relations between point clouds and images.The attention calculation is performed using the same method as described in *Equation (1)*. FCA allows the model to concentrate on features with significant cross-modal correlations. For instance, a feature dimension in point cloud data might have a strong association with a corresponding dimension in image data. The attention architecture enable the model to identify and attend to these critical dimensional relationships.

3.3.3. Feature Fusion Strategies

We designed three different strategies to utilize the attention structures: 1) Temporal-first, 2) Fine-grained feature-first, and 3) Parallel strategy, we name it Parallel TFCA. The specific architectures are shown in the figure 3.

The temporal-first strategy focuses on the time dimension, attempting to capture the dynamic changes in data over time, which is crucial for understanding temporal progression. However, it may somewhat overlook the complex relationships in the fine-grained feature dimension. The fine-grained feature-first strategy prioritizes processing the high-dimensional features within each modality, emphasizing the complex interactions between features at specific time points. This effectively reveals detailed information at particular moments but may sacrifice an overall understanding of temporal dynamics. The parallel strategy simultaneously considers the fusion of both the time dimension and the fine-grained feature dimension, aiming to balance the integration of information from both aspects.In Section 4.4, we demonstrate through experiments that the parallel strategy achieves the best results for feature fusion.

3.4. SMPL-Based Inverse Motion Solver

After the feature fusing,the fused features F_{fusion} are subsequently processed by a designed SMPL-based inverse motion solver. Initially, we input fused features F_{fusion} into a joint solver composed of GRU and MLP to obtain the initial 3D joints $J3d_{initial}$. Next, the initial 3D joints $J3d_{initial}$ and the fused features F_{fusion} are concatenated and processed through an ST-GCN network to obtain more accurate 3D joint positions $J3d_{smpl}$. After that, an SMPL optimizer is used to derive the final pose $pose_{final}$, expressed in axis-angle format. using the parameterized model SMPL Loper et al. (2015). For each obtained value, we calculate the L2 norm loss with respect to the ground truth, denoted as $L_{J3d}^{initial}$, L_{J3d}^{smpl}, L_{pose}.Additionally,we introduce a projection lossL_{J2d}. We project $J3d_{smpl}$ onto the image plane to obtain $J2d$, and then compute the L2 norm loss between $J2d$ and the ground truth.The final loss is as follows:

$$L_{final} = L_{J3d}^{initial} + L_{J3d}^{smpl} + L_{pose} + L_{J2d} \tag{2}$$

3.5. Discriminator

Inspired by VIBE Kocabas et al. (2020) and HMR Kanazawa et al. (2018), we introduce a pose parameter discriminator into our model to enhance overall performance. Discriminators are typically used in generative adversarial networks to determine whether generated targets closely match real labels. Through the supervision of the discriminator, the model learns human pose priors, encouraging the generated pose parameters to better align with the real data distribution. This ensures that the predicted pose more accurately reflect actual human poses.

We use the large-scale human pose dataset AMASS Mahmood et al. (2019) as the real input for the pose parameter discriminator. After the feature fusion step, we add an MLP regressor to predict an initial human $pose_{mlp}$, which serves as the first pseudo-input to the discriminator. Simultaneously, the human pose $pose_{final}$ predicted by the SMPL-based inverse motion solver architecture is used as the second pseudo-input to the discriminator.We combine the distributions of both, denoted as $pose_{gen}$, then the constraint loss for the discriminator can be expressed as follows:

$$L_{Dis} = \mathbb{E}_{pose\sim A}\left[(Dis(pose)-1)^2\right] + \mathbb{E}_{pose_{gen}\sim G}\left[Dis(pose_{gen})^2\right] \tag{3}$$

Where A represents the pose distribution from the AMASS dataset, and G represents the pose distribution computed by our network, including. The loss used by the discriminator to supervise the overall network architecture is defined as follows:

$$L_{adv} = \mathbb{E}_{pose_{gen}\sim G}\left[(Dis(pose_{gen})-1)^2\right] \tag{4}$$

4. Experiment

We first introduce the datasets used in the experiments and the baselines for comparison. Subsequently, we present the quantitative results between our method and other state-of-the-art (SOTA) methods, along with visualized qualitative results. Finally, we analyze the effectiveness of each module in our method through ablation study.

Table 1: Evaluation Results on Different Datasets

Input	Method	LiDARHuman26M				RELI11D			
		ACCEL↓	PAMPJPE↓	MPJPE↓	PVE↓	ACCEL↓	PAMPJPE↓	MPJPE↓	PVE↓
PC	LiDARCap Li et al. (2022)	45.20	66.72	79.31	101.64	36.85	52.19	64.43	75.92
	P4Transformer Fan et al. (2021)	45.40	66.25	79.52	101.77	44.15	72.88	84.91	90.37
RGB	VIBE Kocabas et al. (2020)	120.49	108.19	154.61	191.55	58.33	159.38	276.54	240.39
	MAED Wan et al. (2021)	46.71	59.99	76.36	100.05	51.67	132.67	171.11	189.91
PC+RGB	ImmFusion Chen et al. (2023)	47.61	64.38	77.91	101.39	48.74	103.16	123.08	154.81
	FusionPose Cong et al. (2023)	44.52	66.70	78.17	99.36	42.29	74.51	97.58	106.31
	LEIR Yan et al. (2024)	44.51	62.94	75.09	95.96	27.07	45.72	55.36	75.92
	FTP	**37.34**	**59.92**	**72.58**	**92.91**	**18.41**	**39.86**	**48.13**	**54.96**

Table 2: Runtime analysis between different methods(unit: s).

Method	Inference time(1 epoch)	Training time(300 epoch)
VIBE	3.05	43116.32
MAED	3.51	46964.78
LiDARCap	2.37	35368.59
FTP(ours)	2.88	37727.16

4.1. Experiment Settings

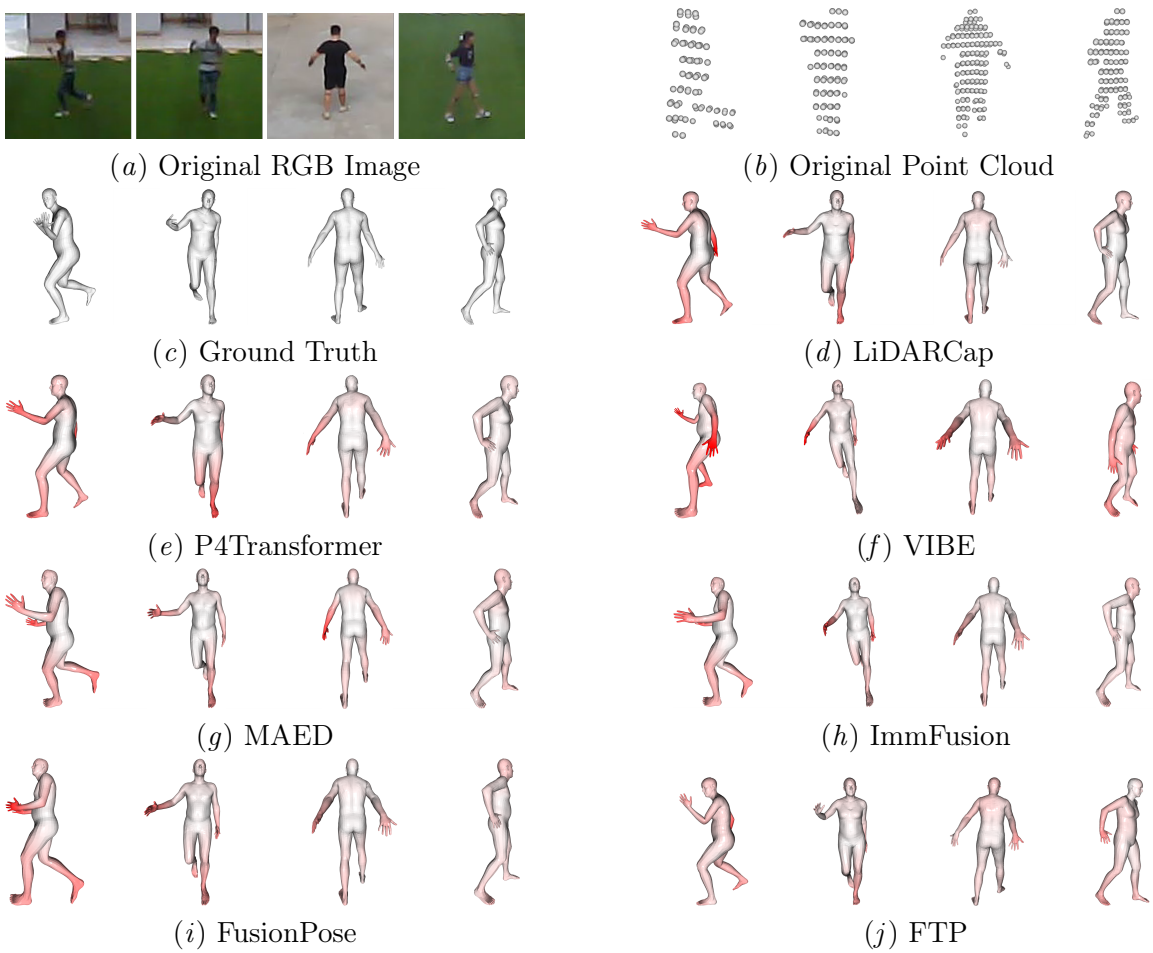
4.1.1. Baselines

We compared our method with various state-of-the-art (SOTA) methods, including camera-based, point cloud-based, and multi-modality methods. For camera-based methods, we selected VIBE Kocabas et al. (2020) and MAED Wan et al. (2021), while for point cloud-based human pose estimation, we used LidarCap Li et al. (2022) and P4Transformer Fan et al. (2021). Additionally, to demonstrate the superiority of FTP over previous multi-modality methods, we retrained and compared several recent multi-modality HPE approachs, including FusionPose Cong et al. (2023), ImmFusion Chen et al. (2023), and LEIR Yan et al. (2024). We conducted comparisons across different datasets to evaluate the performance of these methods.

4.1.2. Datasets

LidarHuman26M Li et al. (2022) contains 180k frames of point cloud and corresponding image data, including many instances of human motion at long distances (up to 28 meters). This dataset allows for the evaluation of model performance on low-resolution images and point cloud data, covering over 20 categories of daily activities. The data modalities include point clouds, video, and ground truth labels provided by IMU.

RELI11D Yan et al. (2024) is a multi-modality, high-quality human motion dataset comprising 239k frames of point clouds along with corresponding image data. It primarily focuses on rapid and complex movements that require precise positioning, covering five different activity categories: table tennis, taekwondo, boxing, fencing, and badminton. The dataset consists of four different modalities: RGB videos, Events, IMU data, and point clouds.



(a) Original RGB Image (b) Original Point Cloud

(c) Ground Truth (d) LiDARCap

(e) P4Transformer (f) VIBE

(g) MAED (h) ImmFusion

(i) FusionPose (j) FTP

Figure 4: Qualitative Results on LiDARHuman26M.(a), (b), and (c) represent the original RGB image, the point cloud, and the ground truth, respectively. (e) to (j) illustrate the results of different baselines. The redness of the human mesh indicates the magnitude of the error at that position compared to the ground truth, with redder colors indicating larger errors.

4.2. Quantitative Evaluation

The left side of Table 1 shows the quantitative metrics on the LiDARHuman26M dataset, demonstrating that our method achieved the best results on LiDARHuman26M dataset. The image-based pose estimation model MAED Wan et al. (2021) obtained better results than other single-modality methods, significantly outperforming VIBE Kocabas et al. (2020) with the same inputs. This performance difference can be partly attributed to MAED Wan et al. (2021)'s ViT-based network architecture, which highlights ViT's powerful modeling capabilities for image data. LEIR Yan et al. (2024) also achieves good results on this task, primarily due to its use of multi-layer temporal encoders and self-attention mechanisms. These techniques allow it to combine the advantages of different modalities to a certain extent, enabling the recovery of more accurate human poses.

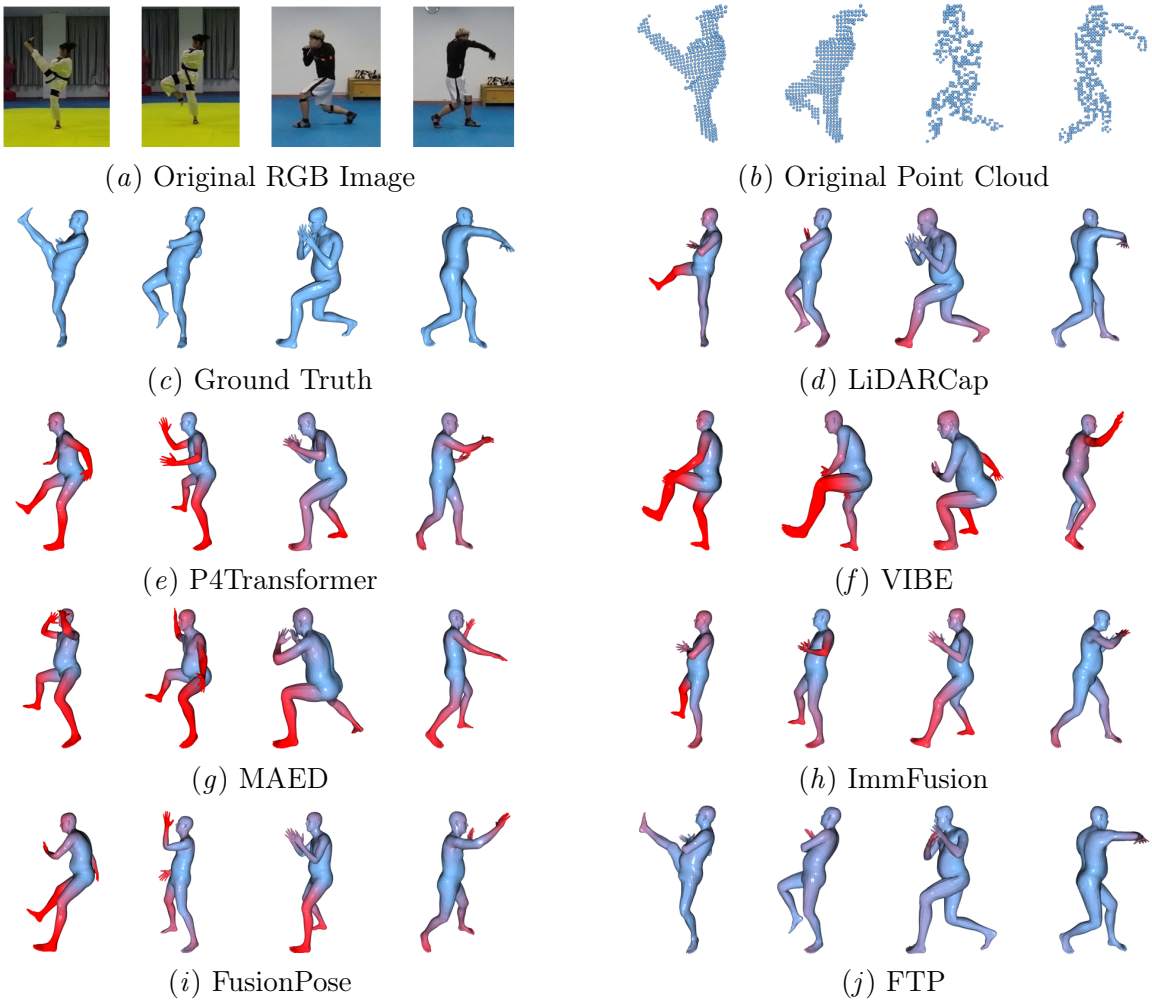
The quantitative evaluation results of all methods on the RELI11D dataset are shown in the right side of Table 1.The table clearly indicates that point cloud-based human pose estimation methods outperform image-based methods. This is primarily because, on the RELI11D dataset, which require precise positioning for fast and complex movements, the 3D spatial information provided by point clouds significantly enhances prediction accuracy. In contrast, images suffer from ambiguous results due to the lack of depth information, limiting their performance. Notably, point cloud-based methods even surpass some multimodal methods, validating our hypothesis that previous multimodal approaches failed to effectively combine the advantages of both modalities. Instead, these approaches may have been limited by the weaknesses of one modality. Our experiments demonstrate that the proposed FTP method addresses these issues and achieves best results.

To analyze the runtime complexity of the methods, we conducted an experimental evaluation of the runtime for various algorithms. The results are presented in Table 2. The results indicate that during the inference process, the cost of time between methods for one epoch is indeed in the second range, which we consider negligible. During training phase, while our discriminator architecture does introduce some additional computational overhead, it remains significantly faster than traditional image-based human pose estimation methods. This is primarily due to the complexity and computational demands of the network architectures and loss functions in those earlier methods. Although our approach involves multiple modules, the overall computational complexity is well-managed. Furthermore, while the training time of our method is marginally longer than that of point cloud-based techniques, we argue that the substantial performance improvements achieved by our method outweigh the slight increase in computational cost.

4.3. Qualitative Evaluation

The qualitative results of FTP and other pipelines on the LiDARHuman26M dataset are shown in Figure 4. The LiDARHuman26M dataset focuses on long-range human motion capture, where the resolution of both image and point cloud data is relatively low. Under these conditions, MAED Wan et al. (2021) achieved better test results compared to LiDARCap Li et al. (2022), indicating that image data can still provide richer semantic information than point cloud data, even at low resolution and from a certain distance. As shown in Figure 4, human poses in images remain recognizable despite the low resolution; in contrast, point cloud data captured at long distances often loses details of the hands and feet. Figure 4 also shows that point clouds become extremely sparse under long-range conditions, resulting in significant missing data for the arms, which leads to severe estimation errors in the arm positions by the point cloud-based methods LiDARCap and P4Transformer. In contrast, our multi-modality pose estimation method FTP effectively integrates the complementary information from both point clouds and images, achieving the best results.

Figure 5 shows the qualitative results of various methods on the RELI11D dataset. It is evident from the figure that FTP perform best qualitatively for the more challenging motions in the RELI11D dataset. This is due to FTP's dynamic fusion of key features from both point clouds and images, leveraging complementary features to achieve more accurate predictions. The results of single-modality methods reveal that rapid leg kicks and fast punches pose significant challenges for these methods. This is especially evident in point



(a) Original RGB Image (b) Original Point Cloud

(c) Ground Truth (d) LiDARCap

(e) P4Transformer (f) VIBE

(g) MAED (h) ImmFusion

(i) FusionPose (j) FTP

Figure 5: Qualitative Results on RELI11D. Each subfigure follows the same representation as in Figure 4.

cloud data, where the sparse points around the feet create difficulties for point cloud-based methods. Additionally, the lack of depth information in images causes image-based methods like VIBE Kocabas et al. (2020) to struggle with correctly estimating which leg is kicking upwards.

4.4. Ablation Study

4.4.1. Ablation Study on Fusion Strategies

We designed three different feature fusion strategies: temporal-first, fine-grained feature-first, and parallel strategy. These strategies were evaluated through ablation experiments, with the results shown in Table 3. Each strategy aims to capture and integrate key information from different dimensions to enhance the network's ability to understand dynamic changes in complex scenes.

Table 3: Fusion Strategies Ablation Study on LiDARHuman26M. T-first and F-first represent the Temporal-first and Fine-grained feature-first trategies respectively

	LiDARHuman26M			RELI11D		
Fusion Strategies	ACCEL↓	PAMPJPE↓	MPJPE↓	ACCEL↓	PAMPJPE↓	MPJPE↓
T-first	41.79	63.58	75.63	18.77	40.61	49.12
F-first	46.08	62.70	74.54	18.97	40.26	49.55
parallel TFCA	**37.34**	**59.92**	**72.58**	**18.41**	**39.86**	**48.13**

Table 4: Discriminator Ablation Study on RELI11D.

	ACCEL↓	PAMPJPE↓	MPJPE↓	PVE↓	PCK0.5↑
without *Dis*	37.81	50.97	65.07	78.43	0.97
with *Dis*	**18.41**	**39.86**	**48.13**	**54.96**	**0.98**

Experimental results show that the parallel strategy performs the best.We believe this is because it simultaneously attends to both temporal and feature details, reducing information loss during fusion and providing richer contextual information. Unlike sequential strategies, it does not prioritize one dimension over another, thus avoiding information loss. Additionally, the parallel strategy improves the model's fault tolerance to feature selection errors, as it does not need to pre-determine which dimension is more important but treats them equally and processes them concurrently. Although parallel processing may increase computational complexity, the performance gains from better information integration can offset the additional computational costs.

4.4.2. Ablation Study on Discriminator

We also conducted ablation experiments related to the discriminator on the RELI11D dataset. Table 4 presents the results. The results indicate that the discriminator architecture significantly enhances the prediction performance of our overall network. We believe this improvement is due to the discriminator learning a vast amount of human pose priors, thereby constraining the generated human poses to be closer to real human poses. Additionally, the GRU temporal encoding block within the discriminator helps constrain the human poses over time, preventing large abrupt changes between adjacent frames and thus smoothing the prediction results.

5. Conclusion

We propose FineTemporalPose (FTP), which employs an attention mechanism to model fine-grained feature relationships and dynamic temporal connections between point clouds and images, enabling neural networks to utilize diverse information more efficiently. However, our approach does not take into account the interaction between human subjects and the surrounding environment, which is a limitation to be addressed in future work. This represents a meaningful direction for further research and improvement.

References

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 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*, pages 561–578. Springer, 2016.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2752–2758. IEEE, 2023.

Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19608–19617, 2022.

Peishan Cong, Yiteng Xu, Yiming Ren, Juze Zhang, Lan Xu, Jingya Wang, Jingyi Yu, and Yuexin Ma. Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 461–469, 2023.

Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6792–6802, 2022.

Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 682–692, 2023.

Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14204–14213, 2021.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

Xiao Han, Peishan Cong, Lan Xu, Jingya Wang, Jingyi Yu, and Yuexin Ma. Licamgait: gait recognition in the wild by using lidar and camera multi-modal visual sensors. *arXiv preprint arXiv:2211.12371*, 2022.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11127–11137, 2021.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 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*, pages 2252–2261, 2019a.

Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4501–4510, 2019b.

Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20502–20512, 2022.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021.

Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12933–12942, 2023.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *Acm Transactions on Graphics*, 34(Article 248), 2015.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018.

Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2500–2509, 2017.

Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12988, 2023.

Ming Yan, Yan Zhang, Shuqiang Cai, Shuqi Fan, Xincheng Lin, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, et al. Reli11d: A comprehensive multimodal human motion dataset and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2250–2262, 2024.

Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 581–597. Springer, 2020.