2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Adaptive Visual-Aided 4D Radar Odometry Through Transformer-Based Feature Fusion

Yuanfan Zhang[1]*, Renxiang Xiao[2]*, Ziyang Hong[2], Liang Hu[2] and Jie Liu[3]

***Abstract*— Multimodal sensor fusion has been successfully utilized in many odometry and localization methods as it increases both estimate accuracy and robustness in application scenarios. To address the challenge of odometry under varying-weather conditions, we propose a novel visual 4D radar fusion based odometry in an unsupervised deep learning approach. In our method, we adopt transformer-based cascaded decoders to facilitate efficient feature extraction of images and radar point clouds. Considering that radars are weather-agnostic and information-rich cameras are susceptible to adverse weathers, we deliberately introduce an adaptive attention-based feature fusion mechanism, in which the attention shifts dynamically to adapt to changing weather conditions based on the amount of information content in image features. Through extensive comparative experiments, our method surpasses different state-of-the-art single-modal odometry estimation methods. Our code and trained model will be released publicly.**

I. INTRODUCTION

Robust Simultaneous Localization and Mapping (SLAM) systems are indispensable for autonomous vehicles to achieve long-term autonomy in real-world environments, particularly in adverse weather conditions. Due to exceptional perception performance in adverse weathers (e.g., rain, fog, snow and smoke), the millimeter-wave (mm-Wave) radar has gained recognition and become a pivotal sensor in all-weather SLAM systems. So far, quite a few radar odometry/SLAM methods have been proposed and experimented in large-scale environments [1]–[5]. While radar-based SLAM solutions excel in adverse conditions, their localization accuracy under normal weather usually lags behind visual and LiDAR-based solutions due to the inherit drawback of mm-wave radars such as sparsity in radar point clouds (PC).

Cameras which are often relatively inexpensive, provide rich visual information of the environment under adequate illumination. Compared with the mm-Wave radar, the camera can adeptly capture dense image information with color and texture details that radars lack at close range. However, a monocular camera alone cannot provide depth information as radar PC do. Conversely, mm-Wave radars excel in capturing extensive spatial geometric structures and even that of visually occluded objects, thanks to its large field of view (FOV) and superior penetration capability. Moreover, while pure visual methods struggle in detecting and handling dynamic objects in complex background, such a challenge can be effectively addressed by exploiting Doppler velocity information from mm-Wave radars.

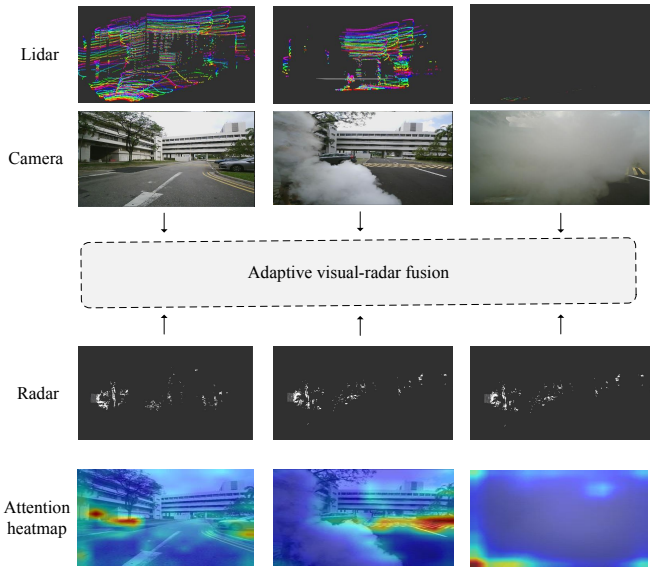

Fig. 1. The raw data from lidar and camera under smoke conditions suffers degradation of different extents, while radar remains almost unaffected. Attention heatmaps of images in the last row show locations which visual encoder pays more attention to, demonstrating that our proposed adaptive multimodal fusion mechanism of radar and visual features.

Inspired by the above observation, we aim to propose a new approach to aid 4D mm-Wave radar odometry with integration of vision. The radar provides a reliable pose prior under any weather condition, while the color and texture information provided by the camera, no matter rich or poor, is maximally utilised to further increase the accuracy of pose estimation. In such a way, the visual-aided radar odometry is not only robust to adverse weather conditions, but also comparably accurate under normal weather condition against SLAM with other types of sensors.

As demonstrated in Fig 1, we introduces a novel transformer-based adaptive multimodal fusion mechanism, enabling stable and efficient feature representation of the surrounding environment in changing weather conditions. Through this fusion, we further propose a visual-aided radar odometry that outperforms both radar-only and visual-only odometries in any weather conditions.

The main contributions of our paper are threefold:

- We propose a visual-aided radar odometry method in

This work is supported by Shenzhen Science and Technology Program (Project No. JCYJ20220818103000001).

*Equal Contribution

[1]Y. Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China.

[2]R. Xiao, Z.Hong and L. Hu are with the Department of Automation, School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen, China. For correspondence: l.hu@hit.edu.cn.

[3]J. Liu is with the Institute for AI Research, Harbin Institute of Technology, Harbin, China.

the deep unsupervised learning framework. Thanks to effective multimodal fusion of complementary sensors, the method provides robust yet accurate odometry estimate under changing weather conditions;
- We design an attention-based adaptive visual-radar fusion manner, in which reliable visual representations can be adaptively combined with radar data to constitute a more comprehensive feature expression of the surrounding environment;
- We evaluate our method on public datasets and compare it with the state-of-the-art (SOTA) open-source methods. Our method demonstrates significant improvements when visual-aiding in effective visual conditions.

II. Related Works

In this paper, our focus lies on learning-based odometry methods using both the visual camera and 4D mm-Wave radar. To begin, we will provide a succinct review of relevant literature covering radar odometry, unsupervised deep learning visual odometry, and the fusion techniques for integrating image and radar point cloud data, discussing each in turn.

A. Radar Odometry

Radar SLAM has resurged in the last decade, and evolved from the 2D mm-Wave radar to its 4D counterparts that provides additional information in the vertical direction. Most radar odometry/SLAM methods, following a similar paradigm of geometry-based visual odometry, develops novel feature extraction and matching algorithms using radar points cloud. Though efficient as its own, the geometry-based radar slam framework is hard to extend for fusing heterogeneous sensors such as cameras. On the other hand, very recently, learning based radar slam solutions have been proposed in [6], which uses 4D radar PC as input and simultaneously estimates scene flow, motion segmentation, and odometry through multiple cross-modal constraint mechanisms. Using the end-to-end learning framework, 4DRVO-Net [2] develops an efficient 4D radar–visual odometry using the feature pyramid, pose warping, and cost volume structure.

B. Unsupervised Deep Visual Odometry

Although traditional deep learning-based odometry methods [7], [8] have achieved promising results, they heavily rely on supervised training, which may not perform well in real-world scenarios. Unsupervised visual odometry has gained significant attention from researchers due to its efficiency, flexibility, and the ability to learn without manual annotation, as well as its powerful generalization capabilities. SfmLearner [9] introduces a framework that simultaneously trains deep depth estimation models and pose estimation models, and uses multi-view image reconstruction as the supervisory signal, eliminating the need for explicit supervision. The subsequent methods [10], [11] further extend this framework by incorporating additional components or techniques to improve the performance or enhance the capabilities of the model. UndeepVO [12] and DF-VO [13] train depth and pose networks on calibrated stereo videos using a photometric loss.

C. Multimodal Fusion of Images and Point Clouds

Deep learning based multimodal fusion has been widely used in various tasks of autonomous vehicles, including object detection [14], [15], semantic segmentation [16], [17], and odometry estimation [2], [3]. Currently, fusion methods for PC and images can be roughly divided into early fusion and late fusion approaches. Early fusion can be further categorized into data-driven and feature-engineered fusion.

1) Data-driven Early Fusion: Data-driven fusion involves the extraction and fusion of raw data from different sensors and mainly used for object detection and semantic segmentation tasks [18]–[20]. An early work that exemplifies this approach is F-PointNet [18], which utilizes an image detector to obtain regions of interest (ROI) and then fuses the corresponding LiDAR points within those regions. Similarly, PI-RCNN [20] identifies key points firstly, then matches and fuses the point cloud with images based on these key points. IPOD [19] filters out a significant amount of background points by exploiting image semantic segmentation information, thereby enhancing detection speed.

2) Feature-engineered Early Fusion: Feature-engineered fusion enables the interaction of cross-modality features at different levels. EPNet [21] combines LiDAR point cloud with camera images to enhance point features with semantic image features in a per-point manner, even in the absence of image annotations. Building upon EPNet, EPNet++ [22] introduces cascaded bi-directional interaction to enriches the image features and semantic information of point features. Transfuser [23] utilizes self-attention mechanism to fuse image and LiDAR representations using Transformer modules at multiple resolutions. The cross-modal fusion of images and radar PCs at the feature level not only extends global perception by fusing features of each modality in non-overlapping areas, but also enhance perception accuracy via cross-modal fusing geometric information from radar PCs and detailed color and texture from camera images. In addition, it eliminates the risk of data singularity when directly fusing data at non-overlapping areas in the data-driven early fusion approach.

3) Late Fusion: Late fusion works at the final stage of sensor fusion and is also known as decision-level fusion. The information from each modality sensor is processed through separate networks, and results from these networks are fused at the final stage. Various network structures such as Fast R-CNN and YOLO are used for object detection individually and then the detection results are then fused to obtain the final detection results in [24] and [25].

III. System Overview

Given a sequence of images $\{I_t \in \mathbb{R}^{H\times W\times 3} \mid t = 1, 2, ..., n\}$ and radar PC $\{PC_t \in \mathbb{R}^{M\times 5} \mid t = 1, 2, ..., n\}$ collected continuously, our goal is to obtain the relative pose estimation in an end-to-end manner. The overall architecture of the network is illustrated in Fig. 2, processing images and radar PC through distinct branches and utilizing fusion blocks for feature integration. A pose estimation model subsequently computes the 6-DoF transformation vectors.

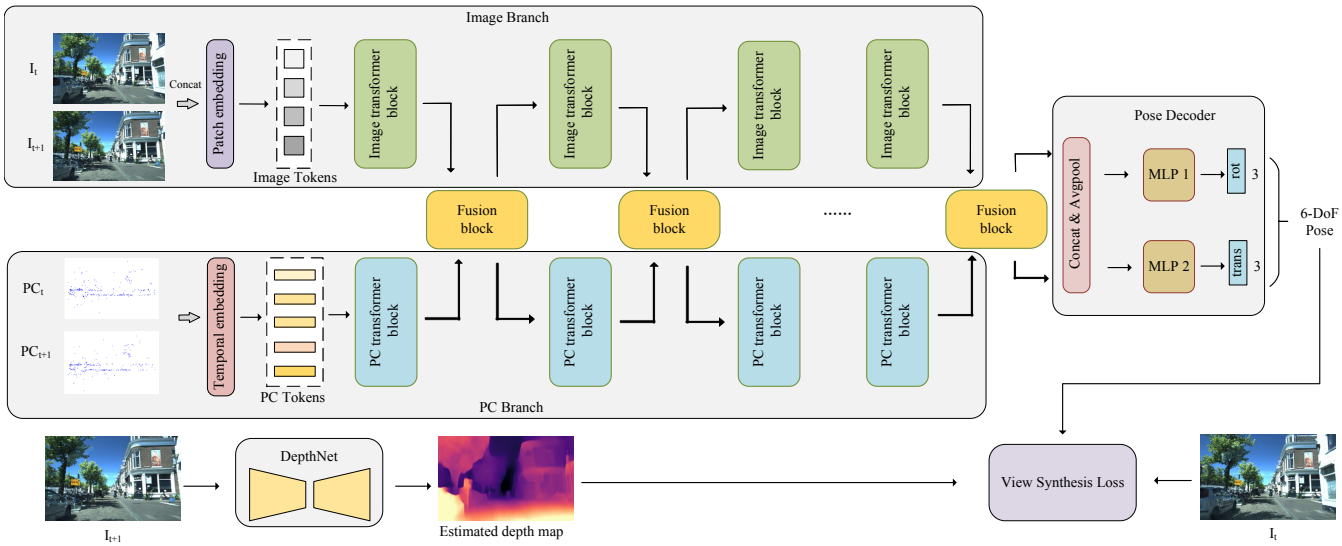

Fig. 2. The overall system architecture. Consecutive frames of images and PC are separately fed into the image branch and point cloud branch, respectively. At each stage of the encoder, an attention-based fusion block facilitates multimodal feature interaction. The pose decoder decouples computation of rotation and translation. The image view synthesis loss is used to train the model, calculated from the source view of image, estimated depth and pose.

A. Image Encoder

The image encoder takes consecutive image frames concatenated along the channel dimension, $\mathbf{I} = concat(I_t, I_{t+1})$, as input and generates a hierarchical feature representation. To learn global information from images, a number of Transformer blocks are stacked within the image encoder, each incorporating multi-head self-attention layers and feed-forward networks. The self-attention is calculated using grid-structured feature maps, and the computation within a transformer block is described as below:

$$\mathbf{B}_i(\mathbf{I}) = FFN(MSA(\mathbf{I}) + \mathbf{I}),$$

where $\mathbf{B}_i()$ represents the image transformer block, $MSA()$ represents multi-head self-attention layers, and $FFN()$ represents feed forward networks. Following the classic query-key-value self-attention mechanism, the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ are calculated via linear projections: $\mathbf{Q} = \mathbf{F}\mathbf{M}^q$, $\mathbf{K} = \mathbf{F}\mathbf{M}^k$, $\mathbf{V} = \mathbf{F}\mathbf{M}^v$, where $\mathbf{M}^q$, $\mathbf{M}^k$, and $\mathbf{M}^v$ are weight matrices, and $\mathbf{F}$ is the input feature map. Moreover, to reduce computational complexity, the dimension of $\mathbf{Q}$ and $\mathbf{K}$ is reduced by a hyper-parameter reduction rate r. The attention layer uses the dot products between the scaled $\mathbf{Q}$ and $\mathbf{K}$ to compute the attention weights and then aggregates the values for each query, which is denoted as:

$$Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_k}})\mathbf{V},$$

where D_k represents the dimensionality of the feature vector $\mathbf{K}$. Multi-heads are incorporated in the self-attention, each head independently executing the attention operation, with their outputs concatenated.

We integrate depth-wise convolution in the FFN layer to enhance spatial feature capture and positional learning, expressed as:

$$FFN(\mathbf{F}) = GELU(DWConv(MLP(\mathbf{F}))) + \mathbf{F},$$

where $DWConv()$ is the depth-wise convolution, $GELU()$ is the Gaussian error linear unit, and $MLP()$ is the multi-layer perception.

B. Point Cloud Encoder

The radar point cloud branch begins with data preprocessing where the 5-dimensional vector consisting of XYZ coordinates, radar cross-section (RCS), and velocity is transformed into 64-dimensional feature vectors by a MLP. The feature vectors are fed into the PC encoder as tokens, where temporal embeddings provide unique time codes to each cloud for differentiation, summarized as:

$$\mathbf{PC} = te(MLP(PC_t), MLP(PC_{t+1})),$$

where $\mathbf{PC}$ represents the point cloud features with time code, and $te()$ represents the temporal embedding operation. The computation within each transformer block in the point cloud branch is described as

$$\mathbf{B}_{pc}(\mathbf{PC}) = FFN(MSA(\mathbf{PC}) + \mathbf{PC}),$$

where $\mathbf{B}_{pc}$ represents the PC transformer blocks. The FFN layer is represented as

$$FFN(\mathbf{F}) = GELU(MLP(\mathbf{F})) + \mathbf{F}.$$

C. Multimodal Fusion Block

Feature alignment: Feature fusion occurs at multiple encoding layers, where fusion blocks align and fuse features from the image and point cloud branches. A fusion block mainly consists of three consecutive operation units: feature alignment projection (FAP), feature fusion using self-attention, and inverse feature alignment projection (IFAP). FAP is a learnable neural network that determines the most suitable shared feature space, where image and point features are projected as normalized features of the same dimension.

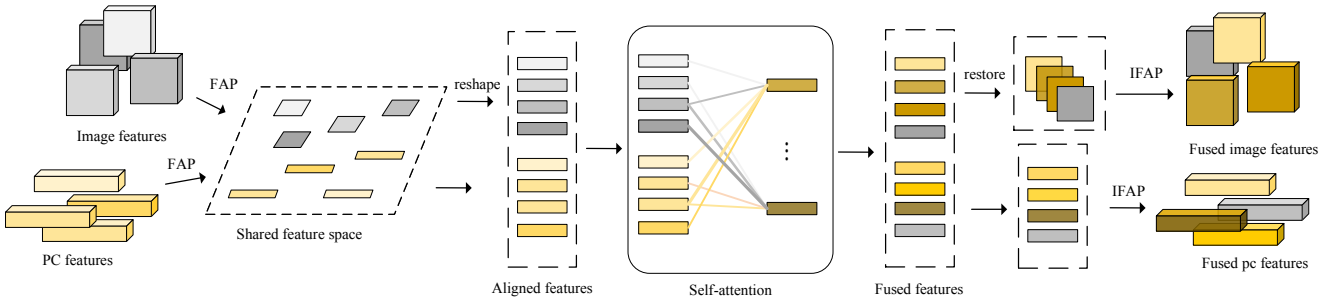

Fig. 3. Feature fusion block. Features from each modal are projected to a shared feature space for alignment. The self-attention mechanism is adopted to dynamically fuse the aligned features. The fused features are restored to original dimensions and projected to respective feature space inversely.

At the last stage, the fused features are restored to their respective feature space of images and radar PC through the IFAP.

Feature fusion: Specifically, consider a set of image features with the size of $C \times H \times W$ and point cloud features with the size of $N \times L$ as the input of the fusion blocks, where C is the number of channels and H, W is the height and width of image respectively, N is the number of radar point features and L is the length of each feature. An average pooling is applied on each channel of the image feature maps and reduces the size from $H \times W$ to $H' \times W'$, where $H'W' = L$, the same as the number of radar features. Next the image feature map is operated by FAP , reshaped to $C \times L$ and then concatenated with the point cloud features. In such a way, the size of normalized features becomes $(C+N) \times L$. The fused features outputted from self-attention layer are identical to the input normalized features. Finally, the IFAP reverses the operation of FAP. The IFAP together with image up-sampling restores fused features to original dimension of images $C \times H \times W$ and that of radar point clouds $N \times L$.

Adaptive Visual-Aiding: The adaptive visual-aiding mechanism arises from the dedicatedly designed network structures in encoders and multimodal fusion blocks. The image qualities change under varying weather conditions, leading to changes in the amounts of information content extracted by visual transformers, and hence its ratio in normalized feature, in turn the attention shifts in the fusion block. The radar branch contributes to spatial scene structure, assisting the image branch to focus on high-information areas. Concurrently, the image branch enriches the radar's scene representation by extracting color and texture features, thereby improving pose estimation accuracy.

D. Pose Decoder

We apply global average pooling to image and PC features from the last fusion block, creating a compact representation for both modalities. These 1-D vectors are then concatenated, encoding the global context of the entire FOV. This composite vector feeds into the pose decoder for final pose estimation. Considering the complexity and unit disparity between rotation (Euler angles) and translation in 6-DoF transformations, we employ two parallel MLPs to independently predict 3D rotation and translation. This approach decouples their regression in the decoder, simplifying the training process.

E. Self-supervised Training

In this self-supervised framework, pose estimation is redefined as a view-synthesis task. It incorporates a depth estimation network, leveraging depth as a intermediary variable for predicting target image appearances from alternate viewpoints. The key supervision signals for the depth and pose estimation neural network stem from the view reconstruction technique that generates scene views from new positions by predicted depth and ego-motion.

Let I_t and I_{t+1} represent consecutive frames at times t and $t+1$, with $p_t(u_t, v_t)$ and $p_{t+1}(u_{t+1}, v_{t+1})$ denoting corresponding pixels in these frames. We can obtain the I'_t reconstructed from I_{t+1} by

$$I'_t(p) = I_{t+1}(KT_{t,t+1}D_tK^{-1}p),$$

where K is the the camera intrinsics matrix, D_t is the depth value of the pixel in the tth frame, and $T_{t,t+1}$ is the pose transformation matrix between two frames, using bilinear interpolation to handle non-integer coordinates.

We define the pose estimation optimization objective as minimizing photometric reconstruction error, enabling self-supervised training through geometric constraints rather than labeled data. The photometric loss is calculated between the source image I_t and the reconstructed image I'_t:

$$L_p = \lambda_p(1 - SSIM(I_t, I'_t)) + (1 - \lambda_p) \left\| I_t - I'_t \right\|_1,$$

where λ_p is the weight parameter. And the edge-aware depth smoothness L_s is used for regularization:

$$L_s(D_t, I_t) = |\partial_x D_t| e^{-|\partial_x I_t|} + |\partial_y D_t| e^{-|\partial_y I_t|}$$

The total loss function is

$$L = L_p + \lambda L_s,$$

where λ is the weight parameter.

IV. Experimental Results

A. Datasets

We conduct experiments on two open-acess 4D radar datasets: the View of Delft (VoD) automotive dataset [26]

TABLE I

QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND OTHER ALGORITHMS ON VOD DATASET. BOLD AND UNDERLINED REPRESENT THE BEST AND SECOND-BEST RESULTS

Methods		01		02		03		04		08		14		19	
		t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
Classical-based	ICP	0.94	1.23	0.82	1.48	0.90	1.50	1.12	1.05	1.03	1.90	0.92	0.55	0.74	1.33
	NDT	1.79	1.67	0.98	1.39	0.47	0.91	1.20	0.92	0.96	1.9	1.24	0.56	1.15	0.79
	ORB-SLAM3	0.99	0.68	0.73	0.59	0.64	0.98	-	-	-	-	-	-	<u>0.14</u>	0.82
Visual-based	DeepVO	2.35	2.13	1.78	3.01	1.34	2.75	1.22	1.37	1.16	2.01	1.58	1.44	1.91	1.98
	SfmLearner	0.92	1.22	0.54	1.37	0.86	1.05	1.04	1.13	0.74	1.3	0.97	0.39	0.96	1.14
Radar-based	CMFlow	<u>0.15</u>	<u>0.09</u>	0.46	0.28	**0.12**	<u>0.20</u>	**0.07**	**0.05**	<u>0.34</u>	0.41	0.18	<u>0.13</u>	0.17	0.58
	Ours(radar only)	0.22	0.17	<u>0.26</u>	0.15	0.32	0.30	0.18	0.14	0.45	0.62	0.17	0.16	0.27	0.39
Visual-radar fusion	late fusion	0.18	0.15	0.48	<u>0.13</u>	0.36	0.27	0.22	0.19	0.4	<u>0.39</u>	<u>0.11</u>	0.14	0.26	<u>0.31</u>
	Ours	**0.11**	**0.04**	**0.14**	**0.08**	<u>0.16</u>	**0.10**	<u>0.10</u>	<u>0.09</u>	**0.10**	**0.18**	**0.04**	**0.03**	**0.12**	**0.13**

and the NTU4DRadLM dataset [27]. The VoD dataset contains 8600 frames of synchronized and calibrated 64-layer LiDAR, camera and 4D radar data collected in complex urban traffic, and provides the ego vehicle's odometry filtered combination of RTK GPS, IMU, and wheel odometry. The NTU4DRadLM is the latest open-source dataset with a benchmark of 4D radar localization algorithms, including 6 differnt sensors: 4D radar, thermal camera, IMU, 3D LiDAR, visual camera and RTK GPS. We apply our method to the NTU-cp sequence and its identical trajectories but with dual smoke events NTU-cp-smoke sequence, created by the authors of [28].

B. Implementation Details

Our proposed model is trained on an Nvidia RTX 3090 GPU and implemented on the Pytorch 1.8.0 framework. In the training process, we use an Adam optimizer and set $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and the batch size is 16. The model is trained with an initial learning rate of 0.0001 for 40 epochs and then reduced half every 5 epochs. We employ the farthest distance sampling strategy to collect the same number of radar points from each frame.

We evaluate the performance of our method using the RPE (Relative Pose Error) metric, measuring the average translational root-mean-square error (RMSE) drift (%) and average rotational RMSE drift (°/100m).

C. Performance Evaluation on VOD

Our method is compared with the following methods:

- The classical algorithms: point cloud-based (ICP [29], NDT [30]), classical visual (ORB-SLAM3 [31]);
- Learning based algorithms: deep supervised learning-based visual (DeepVO [32]), deep unsupervised learning-based visual (SfmLearner [9]), and learning-based 4D radar SOTA (CMFlow [6]) algorithms;
- Our method with radar only and our method with radar-image feature late-fusion as proposed in GRAMME [3] which uses distinct encoders for images and PC for independent modality pose estimations, and then integrates the outputs using a pose fusion network for the final pose transformation.

Sequences 01, 02, 03, 04, 08, 14, and 19 are allocated as the test set for supervised methods DeepVO and CMFlow, with the remaining sequences for training.

Fig. 4 plots the overhead view of the trajectories of our method, our method (radar only), ICP, and CMFlow on sequences 01, 02, 03, 04, 08, 14, and 19. As to quantitatively analysis, it is obvious from Table I that our method yields the best odometry estimate in most of sequences.

The performance of our method exceeds that of visual odometry, particularly in complex scenarios dense with dynamic objects such as sequence 04, 08 and 14, in which ORB-SLAM3 even can not operate completely. Our method outperforms radar-only methods , no matter the classical point cloud-based methods or the learning based CMFlow. The integration of vision improves estimate accuracy greatly even though some scenes are visually occluded. Our method outperforms its counterpart with the late-fusion strategy, which is probably attributed to more frequent information interactions between image and radar features in the feature-engineered early fusion.

D. Odometry Estimation in Suddenly Harsh Environments

1) NTU4DRadLM-cp: We utilize CP loop of the NTU4DRadLM dataset for training. To demonstrate the accuracy of our unsupervised framework, we compare our trained model against the existing SOTA algorithm, 4DRadarSLAM (without loop closure detection). The trajectory projection on the XY plane and the trajecotry error are plotted in Fig. 5 and 6 respectively, which demonstrates that our method achieves superior performance in terms of rotation and translation errors.

2) NTU4DRadLM-cp-smoke: To test the robustness of our method in extreme environments, we evaluate our method on a sequence [28] with twice smoke interference and trajectories same to NTU4DRadLM-CP sequence. We use the model trained on NTU4DRadLM-CP with comparison of the SOTA visual-LiDAR-IMU fusion algorithm R2LIVE [33], As shown in Fig. 7, during smoke, LiDAR only detects

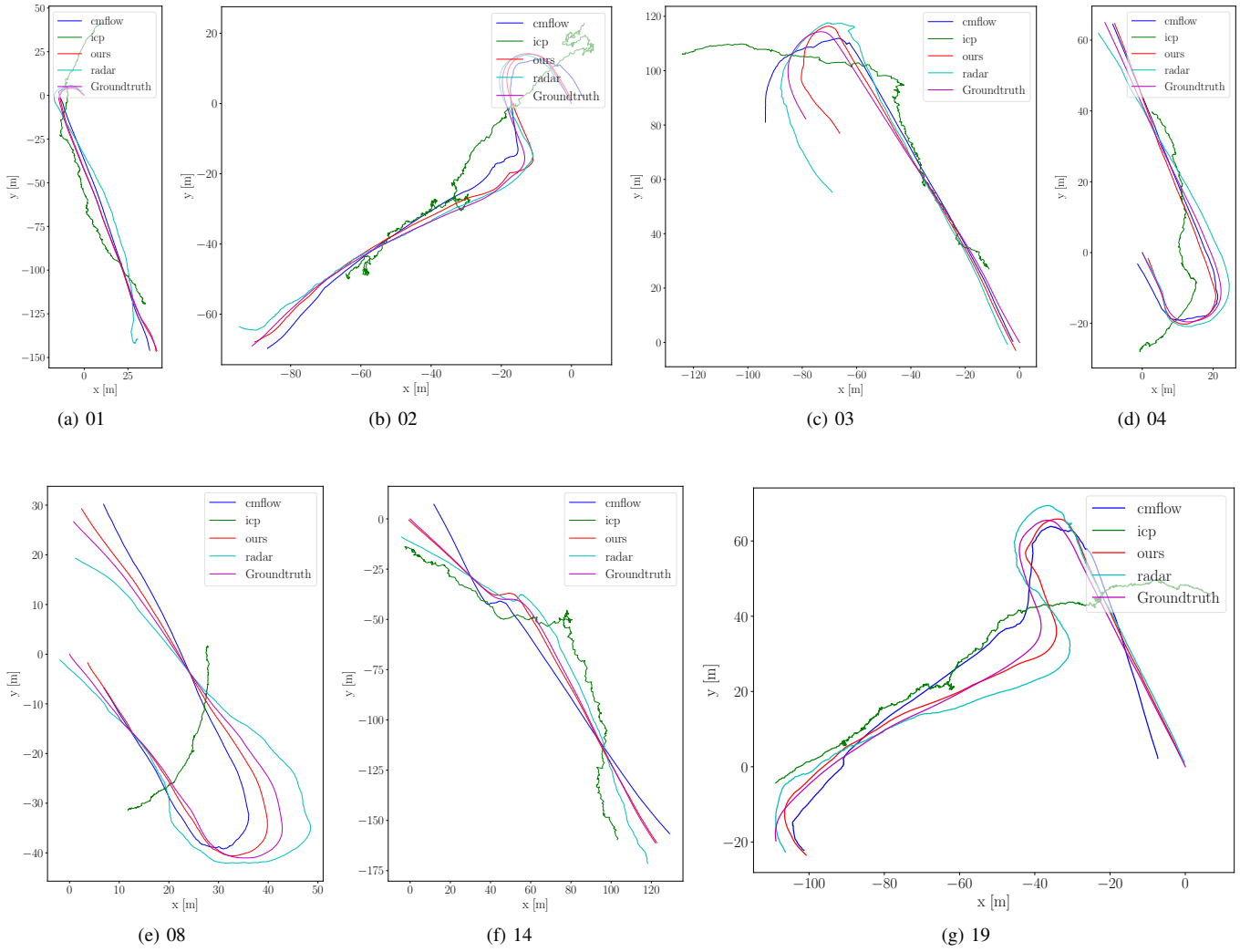

Fig. 4. The trajectories visualization of our method, our method (radar only), ICP and CMFlow on sequences 01, 02, 03, 04, 08, 14 and 19. Trajectories estimated by our method are closer to the ground truth trajectories

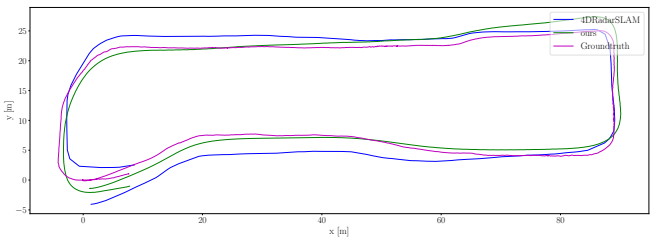

Fig. 5. The trajectories visualization of our method and 4DRadarSLAM.

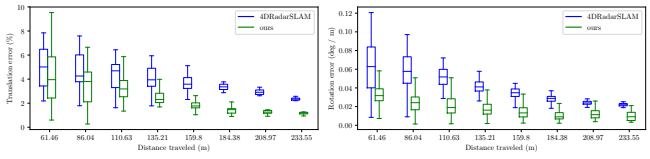

Fig. 6. The box plots of trajectory errors between our method and 4DRadarSLAM.

a minimal number of PC and the camera also fails under smoke, leading to failures in pose estimation. In contrast, our method remains functional, not only when smoke emergence and dissipation but also in smoke areas where visual inputs are highly unreliable. This highlights the robustness and accuracy of our multimodal visual-radar odometry under suddenly-changing weather conditions.

V. CONCLUSIONS

In this work, we introduce an unsupervised learning based visual-aided radar odometry method that adaptively integrates reliable visual patch with radar features for a richer environmental representation. The effectiveness of our method is validated through experiments on the public datasets, demonstrating a comparable performance to pure radar approaches in severe weather conditions and notable accuracy enhancement in visual-clear conditions due to visual enrichment.

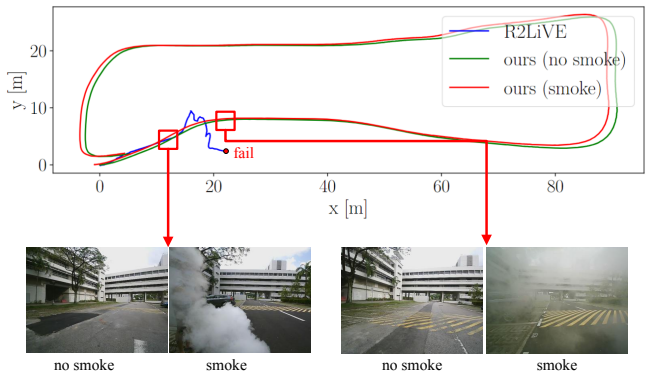

Fig. 7. Experimental results in sudden harsh environments, highlighting the superior robustness of our visual-radar fusion method.

REFERENCES

[1] J. Zhang, H. Zhuge, Z. Wu, G. Peng, M. Wen, Y. Liu, and D. Wang, "4dradarslam: A 4d imaging radar slam system for large-scale environments based on pose graph optimization," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8333–8340.

[2] G. Zhuoins, S. Lu, L. Xiong, H. Zhouins, L. Zheng, and M. Zhou, "4drvo-net: Deep 4d radar–visual odometry using multi-modal and multi-scale adaptive fusion," *IEEE Transactions on Intelligent Vehicles*, 2023.

[3] Y. Almalioglu, M. Turan, N. Trigoni, and A. Markham, "Deep learning-based robust positioning for all-weather autonomous driving," *Nature Machine Intelligence*, vol. 4, no. 9, pp. 749–760, 2022.

[4] F. Ding, Z. Pan, Y. Deng, J. Deng, and C. X. Lu, "Self-supervised scene flow estimation with 4-d automotive radar," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8233–8240, 2022.

[5] Z. Hong, Y. Petillot, and S. Wang, "Radarslam: Radar based large-scale slam in all weathers," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5164–5170.

[6] F. Ding, A. Palffy, D. M. Gavrila, and C. X. Lu, "Hidden gems: 4d radar scene flow learning using cross-modal supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9340–9349.

[7] S. Wang, R. Clark, H. Wen, and N. Trigoni, "End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 513–542, 2018.

[8] W. Wang, Y. Hu, and S. Scherer, "Tartanvo: A generalizable learning-based vo," in *Conference on Robot Learning*. PMLR, 2021, pp. 1761–1772.

[9] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[10] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[11] J.-W. Bian, H. Zhan, N. Wang, Z. Li, L. Zhang, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth learning from video," *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2548–2564, 2021.

[12] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 7286–7291.

[13] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 4203–4210.

[14] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada, "Multispectral object detection for autonomous vehicles," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 35–43.

[15] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird's eye view lidar point cloud and front view camera image for 3d object detection," in *2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018, pp. 1–6.

[16] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Lidar–camera fusion for road detection using fully convolutional neural networks," *Robotics and Autonomous Systems*, vol. 111, pp. 125–131, 2019.

[17] Z. Chen, J. Zhang, and D. Tao, "Progressive lidar adaptation for road detection," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 693–702, 2019.

[18] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[19] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Ipod: Intensive point-based object detector for point cloud," *arXiv preprint arXiv:1812.05276*, 2018.

[20] L. Xie, C. Xiang, Z. Yu, G. Xu, Z. Yang, D. Cai, and X. He, "Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 460–12 467.

[21] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 35–52.

[22] Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, and X. Bai, "Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[23] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 878–12 895, 2023.

[24] T. Kim and J. Ghosh, "Robust detection of non-motorized road users using deep learning on optical and lidar data," in *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)*. IEEE, 2016, pp. 271–276.

[25] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, and U. J. Nunes, "Multimodal vehicle detection: fusing 3d-lidar and color camera data," *Pattern Recognition Letters*, vol. 115, pp. 20–29, 2018.

[26] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+1d radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

[27] J. Zhang, H. Zhuge, Y. Liu, G. Peng, Z. Wu, H. Zhang, Q. Lyu, H. Li, C. Zhao, D. Kircali *et al.*, "Ntu4dradlm: 4d radar-centric multi-modal dataset for localization and mapping," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 4291–4296.

[28] J. Zhang, R. Xiao, H. Li, Y. Liu, X. Suo, C. Hong, Z. Lin, and D. Wang, "4drt-slam: Robust slam in smoke environments using 4d radar and thermal camera based on dense deep learnt features," 06 2023.

[29] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[30] T. Stoyanov, M. Magnusson, H. Andreasson, and A. J. Lilienthal, "Fast and accurate scan registration through minimization of the distance between compact 3d ndt representations," *The International Journal of Robotics Research*, vol. 31, no. 12, pp. 1377–1393, 2012.

[31] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[32] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 2043–2050.

[33] J. Lin, C. Zheng, W. Xu, and F. Zhang, "R2live: A robust, real-time, lidar-inertial-visual tightly-coupled state estimator and mapping," *ArXiv*, vol. abs/2102.12400, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232035489