Spatio-Temporally Consistent Depth Estimation for Dynamic Scenes using 3D Scene Flows

Anonymous ICME submission

***Abstract*—Dynamic depth estimation continues to be crucial but challenging mainly due to the violation of multi-view consistency raised by dynamic areas. Recent approaches have made impressive progress by *implicitly* fusing the intra-relation features, but is still limited for heterogeneous dynamic scenes. In this paper, we propose a new intra-relation feature fusion, but significantly improve the fusion quality using an *explicit* regularization from 3D scene flow cues. We first introduces a Dual Cross-Cue Fusion (D-CCF) module for depth prediction, and further build up an efficient 3D scene flow estimation as explicit 3D spatio-temporal corresponding priors to regularize the depth prediction. Finally, by jointly learning both the depth prediction and 3D scene flow estimation in a unsupervised manner, we achieve more accurate dynamic depth estimation towards spatio-temporal consistency. By extensive evaluation on challenging benchmarks (KITTI and DDAD), our approach can achieve better depth estimation results than state-of-the-art approaches in both static and dynamic areas, which especially maintains the spatio-temporal consistency for dynamic scenes.**

***Index Terms*—dynamic depth estimation, 3D scene flows, spatio-temporal consistency, unsupervised learning**

I. Introduction

In recent years, research has increasingly focused on depth estimation across various fields like autonomous driving [1] and augmented reality [2], [3]. With the success of deep learning networks [4], [5], the mainstream depth estimation approaches can be divided into single image-based [6]–[11] and multiple image-based [12]–[18] approaches, wherein the latter can achieve much better geometric consistent depth estimation relying on multi-view geometric cues. However, these approaches encounter non-negligible difficulty for dynamic scenes, mainly due to the violation of multiview consistency raised by dynamic areas [18], [19].

To overcome such a challenge, some earlier works [12], [13], [18], [19] proposed to identify the dynamic mask independently and use the mask cues to supervise dynamic depth learning. However, the quality of mask identification would significantly influence the depth estimation, which often leads to unsatisfied depth estimation, especially for unbounded dynamic area. Although subsequent works introduce extra semantic [20] or instance segmentation [21]–[23] cues to improve the dynamic depth prediction quality in self-supervised manner, those semantic cues are still unfaithful to rectify the dynamic identification.

Recent approaches [24]–[27] focus on the intra-frame feature fusion to implicitly create expressive multiple frame cues, which achieved impressive quality improvement in both static and dynamic areas. However, such implicit feature fusion is still limited to represent the complex dynamic motion in heterogeneous dynamic scenes, thus often leading to non-coherent depth prediction in both spatial and temporal field across multiple frames. Although leveraging some geometric priors like ground contacting priors [27] could alleviate such issue in the spatial domain, more effective rectification mechanism is still needed to be explored for high accurate dynamic depth estimation towards spatio-temporal consistency.

In this paper, we provide an more effective dynamic depth estimation approach with a new multiple frame feature fusion, but significantly improve the feature fusion quality using an explicit regularization from 3D scene flows. Our key observation is to explore the spatio-temporal corresponding in the 3D dynamic motion, and use it to effectively rectify the non-coherent depth prediction for heterogeneous dynamic scenes. To this end, we first introduce a new multiple frame feature fusion, which warps multiple frames each other and create a 3D volume feature to predict depth following a Dual Cross-Cue Fusion (D-CCF). Based on the depth prediction, we further lift the depth to 3D space, and perform an efficient 3D scene flows estimation across multiple frames based on superpoints [28]. Finally, we leverage the 3D scene flows to regularize the D-CCF depth prediction, and perform a joint learning of both the depth prediction and 3D scene flow estimation in an unsupervised manner, which can lead to more accurate dynamic depth prediction towards spatio-temoral consistency.

To evaluate the effectiveness of our approach, we conduct extensive experiments on the public released challenging dynamic datasets (KITTI [1] and DDAD [6]). Our approach can achieve much better depth estimation than previous approaches (like MonoRec [12], Manydepth [17], DynamicDepth [18] and MaGNet [24]), and also better accuracy than recent implicit feature fusion approaches (such as CCF [25] an AFNet [26]) in both static and dynamic regions quantitatively and qualitatively. To our best knowledge, our approach becomes a new state-of-the-art depth prediction for dynamic scenes, especially maintaining the spatio-temporal consistency across multiple frames for heterogeneous dynamic scenes.

II. Method

For a set of image sequences $\mathcal{I} = \{I_1, I_2, ..., I_N\}$ with given or estimated camera poses $\mathcal{T} = \{T_1, T_2, ..., T_N\}$, our approach aims at predicting I_t's depth map D_t by considering the intra-relation feature fusion of consecutive frames $\{I_{t-1}, I_t\}$. As shown in Fig. 1, we build up a depth prediction networks with a Dual Cross-Cue Fusion module (Sec. II-A). To achieve

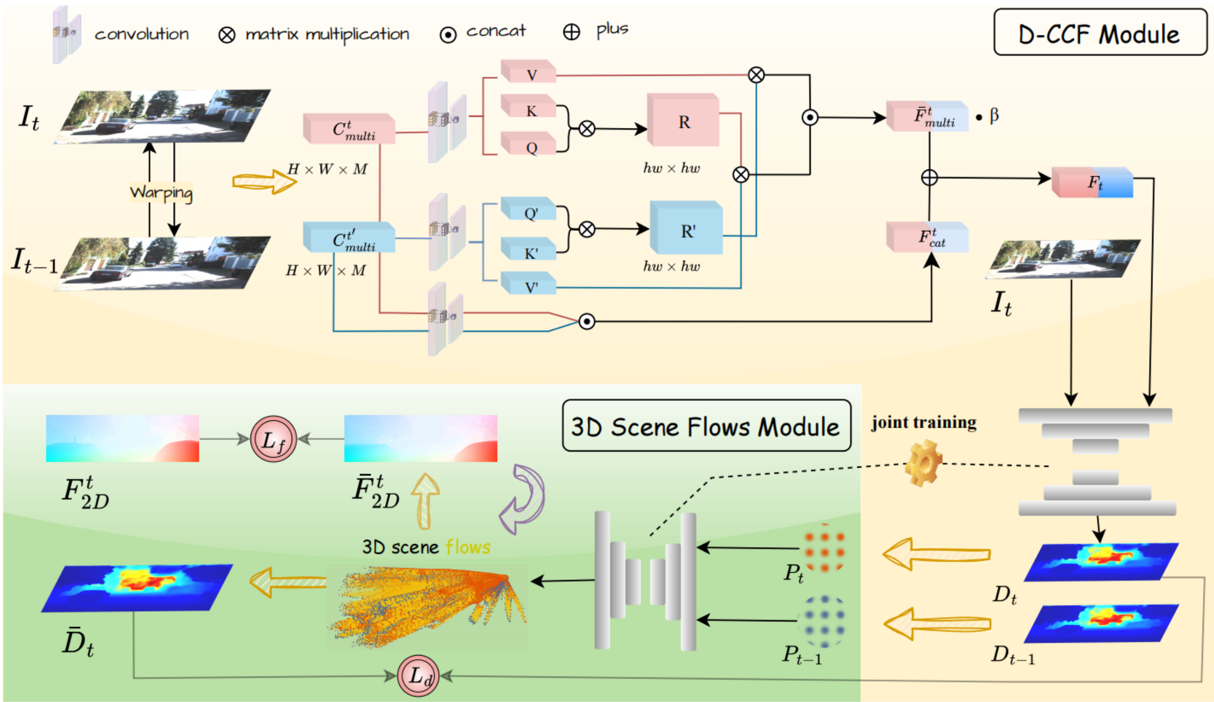

Fig. 1. The framework of the proposed approach, which mainly contains two key components: (1) a D-CCF module to predict depth map using multiple frame feature fusion, and (2) a 3D scene flow estimation module, which is used to effectively regularize the depth prediction learning.

high quality feature fusion for D-CCF, we lift the depth D_t to 3D and perform a 3D scene flow estimation (Sec. II-B) to predict the 3D motion corresponding cues between $\{I_{t-1}, T_t\}$. By jointly unsupervised learning both the depth prediction and 3D scene flow estimation (Sec. II-C), we boost up the feature fusion quality and obtain high accurate depth prediction.

A. Dual Cross-Cue Fusion

Following the paradigm of 3D cost volume [13], we propose to predict the depth map D_t by fusing the intra-relation features from consecutive frames $\{I_{t-1}, I_t\}$. Unlike previous approaches [25], [26] which need extra computation for single view depth, we directly fuse features from image domain efficiently without computing the depth prior from pre-trained models. Besides, to make the feature fusion effective, we warp the consecutive frames $\{I_{t-1}, I_t\}$ to each other in a dual manner, and further create the final feature fusion following a cross-cue attention, thus building up a Dual Cross-Cue Fusion (D-CCF) module to predict the depth map D_t.

Specifically, as shown in Fig. 1 (top), we first warp $\{I_{t-1}, I_t\}$ to each other using the camera poses $\{T_{t-1}, T_t\}$ and create multi-frame cost volumes $C^t_{multi} \in R^{H\times W\times M}$ (by warping I_{t-1} to I_t) and $C^{t'}_{multi} \in R^{H\times W\times M}$ (by warping I_t to I_{t-1}), where we uniformly sample the depth hypotheses $d \in \{d_k\}^M_{k=1}$ from the inverse depth space $[\frac{1}{d_{min}}, \frac{1}{d_{max}}]$ with M denoting the number of depth hypotheses. For each pixel (i, j) in the multi-frame cost volume $C_{multi} \in [0, 1]^{H\times W\times M}$, we compute the pixel-wise similarity between the warped image and targe image using SSIM [12], where large matching scores indicates a higher possibility to the real depth prediction.

Given multi-frame cost volumes C^t_{multi} and $C^{t'}_{multi}$, we further fuse them together to obtain the fused feature volumes F^t_{multi} and $F^{t'}_{multi}$ using a cross-cue attention $\mathcal{A}$, i.e.,

$$F^t_{multi} = \mathcal{A}(C^{t'}_{multi}, C^t_{multi})$$
$$F^{t'}_{multi} = \mathcal{A}(C^t_{multi}, C^{t'}_{multi}), \quad (1)$$

and then yield a feature volume $\bar{F}^t_{multi}$ by concatenating F^t_{multi} and $F^{t'}_{multi}$, i.e., $\bar{F}^t_{multi} = Cat(F^t_{multi}, F^{t'}_{multi})$. Besides, to retain the detailed information from the initial cost volumes, we process the cost volumes via $F^t_{cat} = Cat(Conv(C^t_{multi}), Conv(C^{t'}_{multi}))$, and create the final cross-cue features using:

$$F_t = \beta \bar{F}^t_{multi} + F^t_{cat}, \quad (2)$$

where β is a blending weight. Finally, we feed the cross-cue feature F_t to a depth network $\mathcal{D}$ along with the image information I_t to yield the final depth prediction $D_t = \mathcal{D}(I_t, F_t)$. Please refer to the supplementary materials for the details of cross-cue attention $\mathcal{A}$.

B. 3D Scene Flows Estimation

To more effectively handle dynamic regions, We leverage an iterative end-to-end framework for scene flow estimation as show in Fig. 1 (bottom) based on superpoints [28], where the superpoints are adaptively updated to enhance point-level flow prediction.

We first lift the depth map D_t to 3D, yielding point clouds, and then employ feature encoder utilized in FLOT [29] to extract features from neighboring point clouds. Subsequently, we compute the initial 3D scene flow on point clouds between the target point and source point clouds, and build up superpoint level 3D scene flow following two steps, including flow guided superpoint generation and superpoint guided flow optimization.

Flow Guided Superpoint Generation. To generate superpoints guided by flow, we first compute associations between points and superpoints. Following SPNet [28], we construct a soft association graph through adaptive learning, which calculates the association weights between each point and its K-nearest superpoint centers in the coordinate space. The computed weights are normalized, enabling each point to be assigned to its K-nearest superpoint centers. For each superpoint center, we adaptively aggregate the coordinates, flow, and feature information of its associated points, updating the superpoint center using the normalized association weights.

Superpoint Guided Flow Optimization. Our method adaptively learns flow associations at the superpoint level without relying on rigid object assumptions. Specifically, we encode the flow information of superpoints into the current iteration to guide the generation of new hidden states. Additionally, we incorporate consistency between the flow values reconstructed from superpoints generated by paired point clouds to encode confidence into the current iteration. Finally, the iterative information is input into a GRU to generate updated hidden states. A flow regressor is used to predict residual flows, which are then added to the flow from the $t-1$ iteration step to compute the flow for the current iteration. Please refer to the supplementary materials for more details.

C. Joint Learning

To improve the cross-cue feature quality in the D-CCF module, we leverage the 3D scene flows estimation to regularize the depth prediction learning by formulating a joint learning of both the depth prediction and 3D scene flow estimation simultaneously. Specifically, we formulate a loss function L by combining the depth warping loss L_d and flow warping loss L_f introduced by the 3D scene flow performed on the depth prediction as:

$$L = \lambda_d L_d + \lambda_f L_f. \tag{3}$$

Depth Warping Loss L_d. For the consecutive frames $\{I_{t-1}, I_t\}$ with their depth prediction consecutive frames $\{D_{t-1}, D_t\}$, we lift the predicted depth to 3D yeilding the 3D point clouds P_{t-1}, P_t respectively. Then we perform 3D scene flow estimation $\mathcal{F}_{t-1\to t} : P_{t-1} \to P_t$ and warp the 3D point cloud P_{t-1} yielding the warped point cloud in the current frame $\bar{P}_t = \mathcal{F}_{t-1\to t}(P_{t-1})$. By projecting the warped point cloud $\bar{P}_t$ to the current view T_t, i.e., $p_t = \pi(\bar{P}_t, T_t)$ (π is the projecting function), we can obtain a warped depth map $\bar{D}_t$ by retrieving the depth from the projected points p_t to point cloud $\bar{P}_t$. Subsequently, we formulate the depth warping loss as:

$$L_d = |D_t - \bar{D}_t|^2.$$

Flow Warping Loss L_f. Similarly, for the above projected points p_t, we calculate the 2D flow corresponding field $\bar{F}^t_{2D} \in R^{H\times W\times 2}$ with each pixel p_{ij} in $\bar{F}^t_{2D}$ representing the pixel offset δp_{ij} between the projected points p_t and the original pixel points of I_{t-1}, i.e., $\delta p_{ij} = p_{ij} - \pi(Inv(p_{ij}) \in \bar{P}_t, T_t)$, where $Inv(\cdot)$ is the 3D lifting operation which back-project the pixel p_{ij} to the warped point cloud $\bar{P}_t$. On the other hand, we calculate the 2D flow field F^t_{2D} between I_{t-1} and I_t using previous method [30], and formulate the flow warping loss L_f as:

$$L_f = |F^t_{2D} - \bar{F}^t_{2D}|.$$

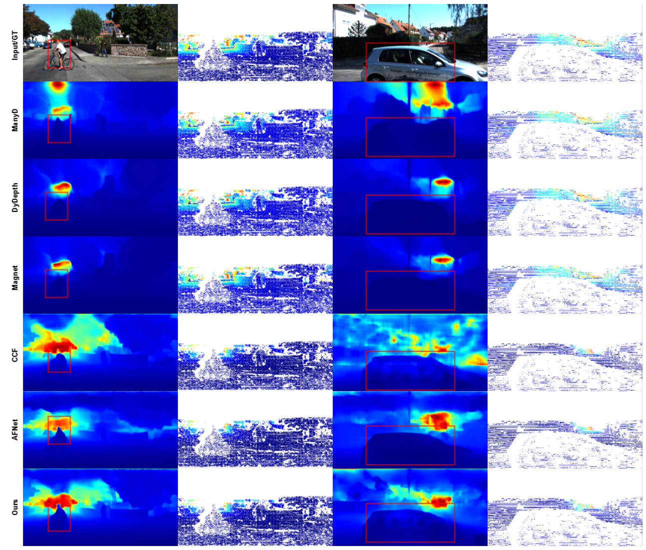

Fig. 2. Some visual comparison results evaluated on the KITTI dataset by different depth prediction approaches.

III. Experiments

To evaluate the effectiveness of our approach, we conduct experiments on two public challenging datasets (KITTI [1] and DDAD [6]) by comparing with previous state-of-the-art depth prediction approaches.

A. System Details

About the implementation. For the D-CCF module, we adopt the cross-attention network as the backbone to fuse the consecutive frame features. The full system is implemented using Pytorch framework, and we use the Adam optimizer with learning rate as $1e^{-5}$ to train all of the networks in our approach.

Parameters. For the multi-frame cost volume C_{multi}, we set the depth hypotheses configuration as $M = 32$, and set the blending weight parameter $\beta = 0.1$ to create the final cross-cue features. In the joint learning stage, we set the two blend weight parameters in the loss function L as $\lambda_d = 0.1$ and $\lambda_f = 0.3$ respectively.

Comparing Approaches. We choose two types of previous approaches for the comparison including earlier depth prediction works such as Manydepth [17], DynamicDepth [18], MonoRec [12] and MaGNet [24]), and recent implicit feature fusion approaches such as CCF [25] and AFNet [26], where

TABLE I
COMPARISON OF DIFFERENT METHODS FOR DEPTH ESTIMATION (KITTI)

Method	**Error Metric (lower is better)**				**Accuracy Metric (higher is better)**		
	AbsRel	**SqRel**	**RMSE**	**RMSE$_{log}$**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Manydepth [17]	0.071	0.343	3.184	0.108	0.945	0.991	0.998
DynamicDepth [18]	0.068	0.296	3.067	0.106	0.945	0.991	0.998
MonoRec [12]	0.050	0.290	2.266	0.082	0.972	0.991	0.996
MaGNet [24]	0.057	0.215	2.597	0.088	0.967	0.996	0.999
CCF Module [25]	0.046	0.155	2.112	0.076	0.973	0.996	0.999
AFNet [26]	0.044	0.132	**1.712**	0.069	0.980	0.997	0.999
Ours	**0.036**	**0.098**	1.721	**0.057**	**0.985**	**0.998**	**1.000**

the latter approaches are the current state-of-the-art dynamic depth prediction approaches. Besides, we conduct all of the experiments and comparison in this paper in a platform with one NVIDIA GPU V100 device.

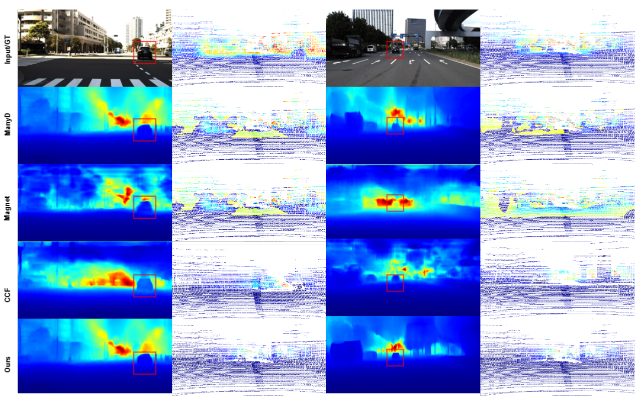

Fig. 3. Some visual comparison results evaluated on the DDAD dataset by different depth prediction approaches.

B. Evaluation on KITTI

We first conduct the comparison on the KITTI benchmark, where we train our framework on the KITTI train set and calculate the depth prediction accuracy metrics on the validation set. As like previous approaches [25], [26], we use four error metrics including AbsRel, SqRel, RMSE and RMSE$_{log}$ to evaluate the depth prediction accuracy. What's more, we also calculate the accuracy metric with different threshold including $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$ respectively.

Table I shows the quantitative comparison result between different comparing approaches. As we can see from the table, our approach can achieve mucher better accuracy metrics than all of the four earlier depth estimation approaches like Many-Depth, DynamicDepth, MonoRec and MaGNet. Comparing with the state-of-the-art depth estimation approaches like CCF and AFNet, our approach can also achieve better accuracy metrics, with only a slightly worse accuracy in RMSE (Ours 1.721) compared with AFNet (1.712), which means that our approach can consistently outperform all of the previous depth prediction approaches.

Fig. 2 show the visual comparison results for depth estimation from different approaches, where our approach can achieve more coherent depth estimation results compared with those previous approaches, especially achieving better accuracy metrics in the dynamic areas. Please refer to our supplementary materials for more visual comparison results.

C. Evaluation on DDAD

To evaluate the generalization ability across different dataset, we also perform evaluation on another challenging dynamic dataset, i.e., DDAD dataset, by comparing with those previous approaches. Similarly, as shown in Table II, our approach consistently achieves much better accuracy metrics than all of the previous approaches, which means that our approach has reliable generalization ability to achieve consistently better depth prediction than previous state-of-the-art approaches. Here since DynamicDepth [18] didn't provide the pre-trained models on DDAD, for a fair comparison we didn't compare with it. But our approach achieves much better depth prediction accuracy than it's same level approaches like Monodepth2 [6] and Manydepth [17].

Fig. 3 demonstrates some visual comparison results for depth prediction from DDAD dataset using different comparing approaches. Our approach also achieve less error metrics than previous approaches in both the static and dynamic areas. Please refer to our supplementary materials for more visual comparison results.

D. Ablation Study

We also conduct an ablation study on our approach to see how the main components take effects on the final depth prediction accuracy, including the different backbone of D-CCF module (Res-18 and Efficent-B5), the depth warping loss and the flow warping loss respectively. Specifically, we choose the KITTI dataset to conduct the experiments, and implement our framework using different backbone (Res-18 and Efficent-B5) in the D-CCF module when predicting depth information. In each backbone, we further implement different system variants by excluding the depth warping loss (denoted as 'Ours w/o L_d') and flow warping loss (denoted as 'Ours w/o L_f'), can then conduct quantitative comparison with our full system (denote as 'Ours FULL') to see how the accuracy differ.

Table III shows the quantitative comparison results for the different system variants of our approaches. As we can see in the table, different backbones (Res-18 and Efficent-B5) don't take much influence on our system, where the accuracy metrics only have a slight difference using Res-18 and Efficent-B5

TABLE II
COMPARISON OF DIFFERENT METHODS FOR DEPTH ESTIMATION (DDAD)

Method	Error Metric (lower is better)				Accuracy Metric (higher is better)		
	AbsRel	**SqRel**	**RMSE**	**RMSE$_{log}$**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [6]	0.381	8.387	21.277	0.371	0.587	-	-
Manydepth [17]	0.146	3.258	14.098	-	0.836	-	-
MonoRec [12]	0.158	3.102	**7.553**	0.227	0.854	0.931	0.961
CCF Module [25]	0.158	2.416	9.855	0.299	0.747	0.894	0.947
MaGNet [24]	0.208	2.641	10.739	0.382	0.620	0.878	0.942
Ours	**0.099**	**1.449**	8.311	**0.153**	**0.911**	**0.965**	**0.980**

TABLE III
ABLATION EXPERIMENT ON WHOLE SCENE (KITTI)

Ablation	Network	Error Metric (lower is better)				Accuracy Metric (higher is better)		
		AbsRel	**SqRel**	**RMSE**	**RMSE$_{log}$**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours w/o L_f	Res-18	0.039	0.147	1.833	0.064	0.982	0.995	0.998
Ours w/o L_d	Res-18	0.041	0.149	1.829	0.063	0.982	0.995	0.998
Ours FULL	Res-18	0.038	0.146	1.824	0.063	0.983	0.995	0.998
Ours w/o L_f	Effi-B5	0.039	0.103	1.772	0.061	0.985	0.996	1.000
Ours w/o L_d	Effi-B5	0.040	0.103	1.771	0.061	0.984	0.998	1.000
Ours FULL	Effi-B5	0.038	0.103	1.767	0.060	0.985	0.998	1.000

backbones in D-CCF module respectively. Overall, Efficent-B5 backbones will achieve better accuracy metrics than Res-18 backbone. Moreover, both 'Ours w/o L_d' and 'Ours w/o L_f' achieve consistently worse accuracy metrics than 'Ours FULL', which means that the depth warping loss L_d and flow warping loss L_f take effects to improve the depth prediction accuracy in our system. Besides, since 'Ours w/o L_d' achieves more accuracy decrease than 'Ours w/o L_f', which shows that depth warping loss L_d makes much more accuracy improvement than flow warping loss L_f during in the joint learning of our system.

Fig. 4 show several visual comprison results for depth prediction using different variants of our system, including 'Ours w/o L_f', 'Ours w/o L_d' and 'Ours FULL' respectively. As we can see in the figure, 'Ours FULL' can achieve more coherent depth prediction which shows the effectiveness of 3D scene flow module.

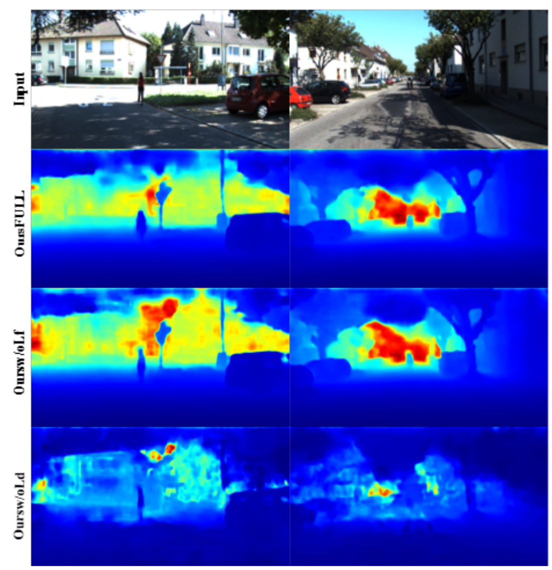

Fig. 4. The visual comparison results of different system variants evaluated on KITTI dataset.

E. Time and Memory Efficiency

We conduct a time and memory efficiency analysis on our framework. In average, our approach takes about one hour to complete the joint learning in KITTI and DDAD dataset. For the inference, our approach takes about 210ms on average to perform per-frame depth estimation, which is faster than previous approaches like CCF (with 280ms on average). What's more, per-frame depth prediction will cost about $0.2G$ GPU memory of our approach, which is also smaller than previous approaches like CCF (with 0.6G).

F. Limitation and Discussion

One main limitation of our approach comes from the 3D scene flow estimation module. Though our current solution can achieve better depth prediction than previous SOTA approaches, the quality of 3D scene flow estimation would influence the overall depth prediction improvement. One possible solution would be to leverage more effective 3D flow prediction [31], [32], which we leave for future works since the improvement of 3D scene flow is out of our main contribution in this paper. Besides, our approach also face such challenging for fast moving objects as like previous intra-relation feature fusion approaches [25], [26], which could be further improved by using more semantic cues to identify moving objects in the spatio-temporal manner.

IV. CONCLUSION

In this paper, we provide a new dynamic depth prediction approach, which leverages the explicit 3D scene flow to regularize the intra-relation feature fusion learning. By incooperating 3D scene flow to improve the feature fusion in a unsupervised manner, we show that our approach can achieve better dynamic depth prediction results towards spatio-temporal consistency. We hope that our approach could inspire more subsequent works to leveraging more effective explicit regularization to enhance the feature fusion quality for the depth prediction, towards much better spatio-temporally consistent depth prediction for heterogeneous dynamic scenes.

REFERENCES

[1] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf, "Consistent video depth estimation," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 4, pp. 71–1, 2020.

[3] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang, "Robust consistent video depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1611–1621.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2319–2328.

[6] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[7] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4009–4018.

[8] Wei Yin, Yifan Liu, and Chunhua Shen, "Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7282–7295, 2021.

[9] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen, "Towards accurate reconstruction of 3d scene shape from a single monocular image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6480–6494, 2022.

[10] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan, "Neural window fully-connected crfs for monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3916–3925.

[11] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon, "Sparse auxiliary networks for unified monocular depth prediction and completion," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2021, pp. 11078–11088.

[12] Felix Wimbauer, Nan Yang, Lukas Von Stumberg, Niclas Zeller, and Daniel Cremers, "Monorec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6112–6122.

[13] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[14] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2485–2494.

[15] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon, "Full surround monodepth from multiple cameras," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5397–5404, 2022.

[16] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou, "Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Conference on robot learning*. PMLR, 2023, pp. 539–549.

[17] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1164–1174.

[18] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li, "Disentangling object motion and occlusion for unsupervised multi-frame monocular depth," in *European Conference on Computer Vision*. Springer, 2022, pp. 228–244.

[19] Tai Wang, Jiangmiao Pang, and Dahua Lin, "Monocular 3d object detection with depth from motion," in *European Conference on Computer Vision*. Springer, 2022, pp. 386–403.

[20] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 582–600.

[21] Fabian Brickwedde, Steffen Abraham, and Rudolf Mester, "Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2780–2790.

[22] Zhe Cao, Abhishek Kar, Christian Hane, and Jitendra Malik, "Learning independent object motion from unlabelled stereoscopic videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5594–5603.

[23] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon, "Learning monocular depth in dynamic scenes via instance-aware projection consistency," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 1863–1872.

[24] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla, "Multi-view depth estimation by fusing single-view depth probability with multi-view geometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2842–2851.

[25] Rui Li, Dong Gong, Wei Yin, Hao Chen, Yu Zhu, Kaixuan Wang, Xiaozhi Chen, Jinqiu Sun, and Yanning Zhang, "Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21539–21548.

[26] Junda Cheng, Wei Yin, Kaixuan Wang, Xiaozhi Chen, Shijie Wang, and Xin Yang, "Adaptive fusion of single-view and multi-view depth for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10138–10147.

[27] Jaeho Moon, Juan Luis Gonzalez Bello, Byeongjun Kwon, and Munchurl Kim, "From-ground-to-objects: Coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10519–10529.

[28] Le Hui, Jia Yuan, Mingmei Cheng, Jin Xie, Xiaoya Zhang, and Jian Yang, "Superpoint network for point cloud oversegmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5510–5519.

[29] Gilles Puy, Alexandre Boulch, and Renaud Marlet, "Flot: Scene flow on point clouds guided by optimal transport," in *European conference on computer vision*. Springer, 2020, pp. 527–544.

[30] Zachary Teed and Jia Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[31] Chaokang Jiang, Guangming Wang, Jiuming Liu, Hesheng Wang, Zhuang Ma, Zhenqiang Liu, Zhujin Liang, Yi Shan, and Dalong Du, "3dsflabelling: Boosting 3d scene flow estimation by pseudo auto-labelling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15173–15183.

[32] Yushan Zhang, Johan Edstedt, Bastian Wandt, Per-Erik Forssén, Maria Magnusson, and Michael Felsberg, "Gmsf: Global matching scene flow," *Advances in Neural Information Processing Systems*, vol. 36, 2024.