Adaptive Fusion of Single-View and Multi-View Depth for Autonomous Driving

Supplementary Material

Method	$\delta = 0$	$\delta = 0.01$	$\delta = 0.025$	$\delta = 0.05$	ID	R-Rel
AFNet	**0.092**	0.125	0.155	0.164	0.165	0.168
AFNet-Pose	0.093	**0.123**	**0.142**	**0.154**	**0.164**	**0.160**

Table 9. Ablation results for pose correction module on AbsRel error on DDAD [12].

6. Robustness under real-world noise poses

Because the monocular version of ORBSLAM2[25] crushes severely on some sequences, we compare it on sequences that perform moderately, and only evaluate the image sequence before it crushes. As shown in Table 10, the results of the remaining sequences in the KITTI Odometry dataset also show the robustness of our AFNet.

7. Dynamic object region mask

We claim that our adaptive fusion module can alleviate the problem that multi-view methods cannot handle dynamic objects, so we compared the performance in the region of dynamic objects on the front view camera on DDAD [12]. In order to obtain the mask of the dynamic object region, we first obtain the region mask1 through instance segmentation that is likely to have dynamic objects, such as cars, pedestrians, bicycles, etc. Then, in the time sequence, the mask1 region of the previous image and the next image are warping to the current image, and the SSIM similarity scores are calculated with the current image. The region with the similarity score less than 0.7 is taken as the final dynamic object region mask.

8. Pose Correction Module

Method. Because the prediction of multi-view branch determines the upper limit of the final accuracy of the system, we propose the pose correction module to adaptively replace the input noisy pose with the pose predicted by Posenet [17] into multi-view branch and AF module for further accuracy improvement. Specifically, we input the features $\boldsymbol{F}_{i,4}$ extracted from the feature extraction network into the decoder part of Posenet to obtain the predicted Euler angles $\boldsymbol{r}_{i,pred}$ and translation $\boldsymbol{t}_{i,pred}$ between the reference and the i-th source cameras. The Euler angles are converted to the rotation matrix $\boldsymbol{R}_{i,pred}$ for warping. Then the source images are warped according to the predicted $\boldsymbol{R}, \boldsymbol{t}$ and the input $\boldsymbol{R}, \boldsymbol{t}$ respectively as in Section 3.3 in paper, denoted as $\left\{\boldsymbol{I}^{'}_{i,pred}\right\}_{i=1}^{n-1}$ and $\left\{\boldsymbol{I}^{'}_{i,input}\right\}_{i=1}^{n-1}$. The difference is that the depth used in this warping is single-view prediction $\boldsymbol{d}_s$, since it is not associated with pose. The SSIM similarity scores between reference image $\boldsymbol{I}_0$ and warping images $\left\{\boldsymbol{I}^{'}_{i,pred}\right\}_{i=1}^{n-1}$ and $\left\{\boldsymbol{I}^{'}_{i,input}\right\}_{i=1}^{n-1}$ are calculated respectively, and the corresponding $\boldsymbol{R}, \boldsymbol{t}$ with large scores are taken as the input of multi-view branch and the adaptive fusion module.

Ablation study. To improve the depth accuracy when pose degradation is severe, we propose the pose correction module for AFNet, denoted as AFNet-Pose. As shown in Table 9, under different intensities of the pose noise δ, AFNet-Pose has a further improvement when the pose is noisy compared with AFNet, especially δ = 0.025 has a 8.4% improvement on AbsRel error.

9. Parameter comparison

In order to prove that the effectiveness of our method is not obtained by parameter stacking, we compare the performance of our method with the current single-view and multi-view fusion methods and classical single-view methods. As shown in Table 11, our AFNet with ConvNeXt [22] backbone has the highest accuracy, but the number of parameters is larger than [43], so we replace the ConvNeXt backbone with a lighter backbone MobileNetV2 [29]. It can be seen that our AFNet(MobileNetV2) has the highest accuracy and the lowest number of network parameters compared with other methods.

Sequence	00			06			07		
Pose	GT	ORB2 (ATE=0.92)	ORB1 (ATE=7.15)	GT	ORB2 (ATE=0.69)	ORB1 (ATE=13.8)	GT	ORB2 (ATE=0.48)	ORB1 (ATE=2.91)
MoRec[35]	0.054	0.063	0.388	0.063	0.078	0.373	0.053	0.058	0.416
IterMVS[34]	0.058	0.067	0.114	0.043	0.052	0.093	0.064	0.075	0.128
MaGNet[1]	0.056	0.060	0.066	0.039	0.044	0.050	0.062	0.066	0.073
AFNet	**0.052**	**0.054**	**0.058**	**0.039**	**0.041**	**0.044**	**0.055**	**0.058**	**0.063**

Table 10. Performance comparison under Ground Truth poses and SLAM system poses (ORB1 and ORB2 represents the monocular version and stereo version of ORBSLAM2[25] respectively) on KITTI [11] Odometry dataset. ATE represents the absolute trajectory error between the estimated poses and the Ground Truth poses. The reported numbers are AbsRel error.

Type	Model	DDAD [12]			KITTI [11]			parm(M) ↓
		AbsRel↓	SqRel↓	RMSE↓	AbsRel↓	SqRel↓	RMSE↓	
Single View	BTS [19]	0.169	2.81	11.85	0.059	0.245	2.756	112.8
	AdaBins [2]	0.164	2.66	11.08	0.058	0.190	2.360	78.0
Fusion Methods	MVS2D [43]	0.132	2.05	9.82	0.058	0.176	2.277	24.4
	MaGNet [1]	0.112	1.74	9.23	0.054	0.162	2.158	76.4
	AFNet(MobileNetV2)	0.091	1.45	7.28	0.040	0.124	1.751	**13.9**
	AFNet(ConvNeXt)	**0.088**	**1.41**	**7.23**	**0.039**	**0.121**	**1.743**	46.1

Table 11. Performance comparison on DDAD [12] and KITTI [11], our AFNet has higher accuracy and fewer network parameters.