Supplementary Materials

Weisong Ren, Lijun Wang et al.

Paper ID 0405

1 Experiment Details on Monodepth2 Network

We claim that our framework can improve the depth estimation performance of existing unsupervised approaches. In our paper, we further showcase the experimental results based on the most popular Monodepth2 network. In this section, we introduce more training details of this experiment.

Detailed modification of the network. Specifically, we modify the basic single stream network design into our proposed multi-stream format, severing as the network of the teacher ensemble and the student. We use the same ResNet-18 as the original model. The modified decoder is shown in Table 1. Specifically, for the teacher ensemble, we use 4 coefficient decoders with one basis decoder. For the final student, we use only one coefficient decoder with one basis decoder.

Training details. Compared with original training strategy, we employed our proposed two-stage training scheme on the modified network introduced above. In the first stage, we train the teacher ensemble with Eq.6 in main paper. Following [3], we train the model for 20 epochs using Adam in the first stage. We use a learning rate of 1e-4 for first 15 epochs and then dropped to 1e-5 for the remainder. In the second stage, we use the proposed cost-aware distillation loss to distill the student model. The second stage is trained for 20 epochs, with a learning rate of 1e-4.

2 Visualization of Basis Maps

In our main paper, we claimed that our proposed MUSTNet can contribute to both the teacher ensemble and the student network, for the bases decoder can generate more diverse depth features. We further visualize some typical depth bases in this section. The results are shown in Figure 1. Specifically, we can observe that some depth bases (**Top-right**) have higher response on the close objects while some others bases (**Bottom-right**) focus on the learning of the far background. Additionally, some bases maintain fine-grained details of the close objects (**Top-left**) or the far background (**Bottom-right**).

3 Visualization of Depth Maps of the Teachers/Student

We further showcase the qualitative comparison of the recovered depth maps by the teacher ensemble and the student model, shown in Figure 2. We can

Basis Decoder						
layer	**k**	**s**	**ch**	**res**	**input**	**activation**
upconv5	3	1	256	32	e_conv5	ReLU
iconv5	3	1	256	16	upconv5, e_conv4	ReLU
upconv4	3	1	128	16	iconv5	ReLU
iconv4	3	1	128	8	upconv4, e_conv3	ReLU
disp4	3	1	**16**	1	iconv5	Sigmoid
upconv3	3	1	64	8	iconv4	ReLU
iconv3	3	1	64	4	upconv3, e_conv2	ReLU
disp3	3	1	**16**	1	iconv3	Sigmoid
upconv2	3	1	256	32	iconv3	ReLU
iconv2	3	1	256	16	upconv2, e_conv1	ReLU
disp2	3	1	**16**	1	iconv2	Sigmoid
upconv1	3	1	256	32	iconv3	ReLU
iconv1	3	1	256	16	upconv1	ReLU
disp1	3	1	**16**	1	iconv1	Sigmoid
Coefficient Decoder						
layer	**k**	**s**	**ch**	**res**	**input**	**activation**
c_conv1	1	1	256	32	e_conv5	ReLU
c_conv2	3	1	256	32	c_conv1	ReLU
c_conv3	3	1	256	32	c_conv2	ReLU
c_conv4	3	1	256	32	c_conv3	ReLU
c_conv5	3	1	16	32	c_conv4	ReLU
c_pool1	-	1	16	1	c_conv5	-

Table 1. Network architecture of the coefficient decoder and the basis decoder. k represents the kernel size. **ch** represents output channels. **res** represents the resolution scale. **input** represents the input features for each layer. **activation** represents the activation function. **e_conv** represents the output feature of the shared encoder.

observe that, during distilling, the teacher ensemble can select more accurate knowledge to guide the training of the student model. It reveals that, benefited from the adaptive co-teaching framework, the student model produces more accurate depth maps than the teacher ensembles.

4 Additional Visualization of the Masks

In Figure 3, we showcase the masks used in our method. Specifically, the 6th row represent the selective mask M_i^d for the distillation scheme, which is introduced in Sec 4.2 in the main paper. Additionally, as described in Sec 4.3, to address some dynamic scenes, we used additional pseudo optical flow maps generated by [4] as geometric guidance. We generate dynamic masks by computing the displacement of each pixel between the current frame and adjacent frames using a) pseudo optical flow and b) cross-view projection based on the predicted depth and pose. The generated dynamic masks are shown in 7th row.

Methods	Sup.	Resolution	Error metric↓				Accuracy metric↑		
			$AbsRel$	$SqRel$	$RMSE$	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SfMLearner	M	416×128	0.176	1.532	6.129	0.244	0.758	0.921	0.971
Vid2Depth	M	416×128	0.134	0.983	5.501	0.203	0.827	0.944	0.981
DDVO	M	416×128	0.126	0.866	4.932	0.185	0.851	0.958	0.986
EPC++	M	640×192	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Monodepth2	M	640×192	0.090	0.545	3.942	0.137	0.914	0.983	0.995
HR-Depth	M	640×192	0.082	0.484	3.776	0.127	0.925	0.986	0.996
Ours	M	640×192	**0.082**	**0.447**	**3.687**	**0.126**	**0.926**	**0.986**	**0.996**

Table 2. Quantitative results of depth estimation on KITTI improved ground truth. Best results are marked **bold**. Comparison of existing methods to our own on the KITTI 2015 [2] using the improved ground truth [5] of the eigen split. **M** represents unsupervised monocular supervision. Results are presented without any post-processing.

Methods	Sup.	Resolution	Error metric↓				Accuracy metric↑		
			$AbsRel$	$SqRel$	$RMSE$	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours	M	640×192	0.106	0.763	**4.562**	**0.182**	0.888	**0.963**	**0.983**
Ours	MF*	640×192	**0.105**	**0.753**	4.568	**0.182**	**0.890**	**0.963**	**0.983**

Table 3. The effect of the dynamic masks.

5 Qualitative Results on KITTI Improved Ground Truth

In our main paper, we evaluate our method using the evaluation method introduced by [1], which creates the ground truth depths by re-projecting LiDAR points. However, the generated ground truth maps do not handle occlusions and non-rigid parts. To solve this problem, some methods use the improved high quality ground truth depth maps introduced by [5]. These high quality images are instead reprojected using 5 consecutive LiDAR frames and uses the stereo images for better handling of occlusions. Finally, this improved ground truth depth is provided for 652 of the 697 test frames contained in the Eigen test split [1]. We evaluate our approach on the improved ground truth frames and compare to existing methods without retraining. We evaluate these methods using the same error metrics as the standard evaluation, and clip the predicted depths to 80 meters to match the Eigen evaluation. As shown in Table 2, our method still significantly outperforms all previously published methods on all metrics.

6 The Set of the Number of Depth Bases

We claim that dividing the single channel output into depth bases contributes to more accurate depth predictions. The number of the depth bases (N) is a

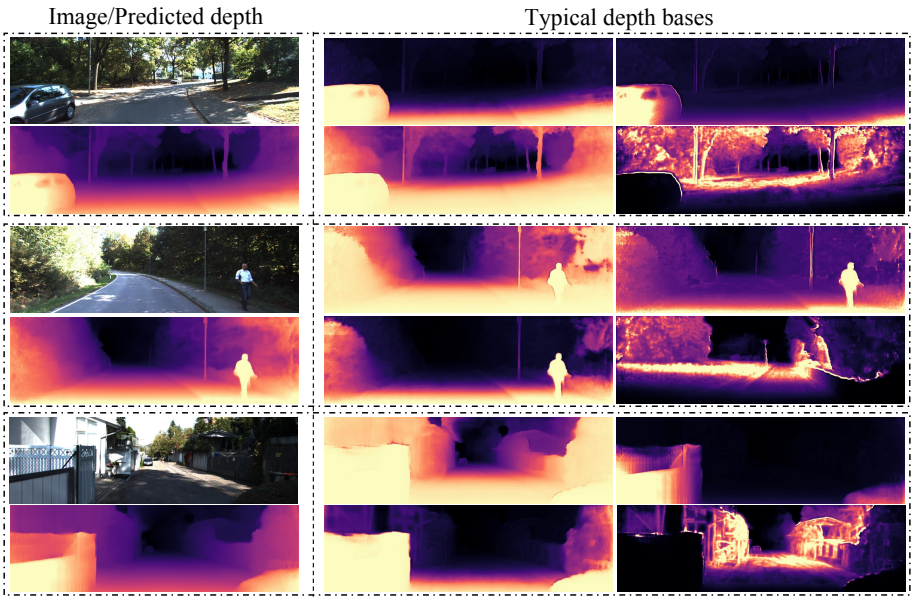

Fig. 1. Visualization of typical depth bases and the predicted depth maps.

key hyper-parameter for it influence the dimension of the depth bases and the coefficient vector. In our experiments, we have tested different set of N (4, 8, 16, 32, 64). It seems that a higher numbers of outputs be more beneficial. However, in our experiments, when N is larger than 16, the performance gain is marginal. Specifically, when N is set as 8/16/32/64, the *AbsRel* of the student model are 0.108/0.105/0.105/0.106. We conjecture that several typical depth basis are sufficient enough for diverse results. Setting a larger N will result in redundant channels. N=16 is the best set to trade-off computational complexity and depth basis diversity when training our model in the KITTI dataset.

For training the network on other datasets, we suggest to adjust this hyper-parameter according to the default setting(N=16).

7 The effect of the dynamic masks

The cost volume based masking mechanism essentially follows the static scene assumption. It fails to mask out the dynamic parts. To solve this problem, we generate dynamic masks with the help of additional optical flow information. Specifically, we compute the displacement of each pixel between the current frame and adjacent frames using a) optical flow estimation [4] and b) cross-view projection based on the predicted depth and pose. If the error of the above two methods is larger than a threshold τ_s, we mark the current pixel as dynamic and will ignore it during loss computation. The results can be seen in Table 3.

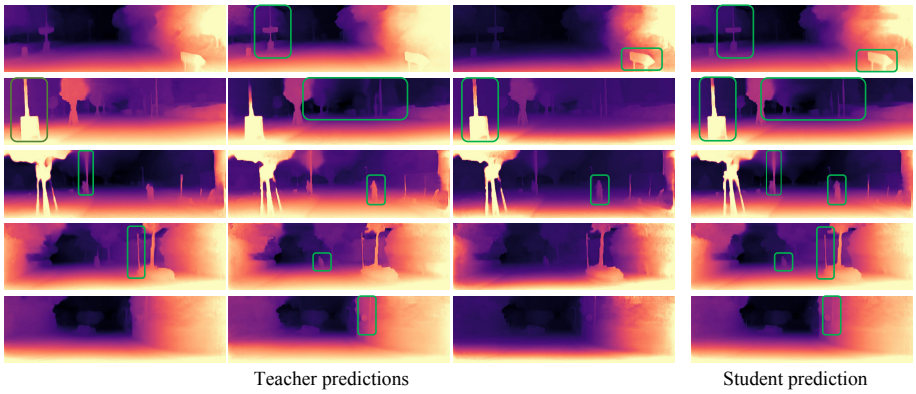

Fig. 2. Visual comparison of the depth predictions of the teacher models and the student network.

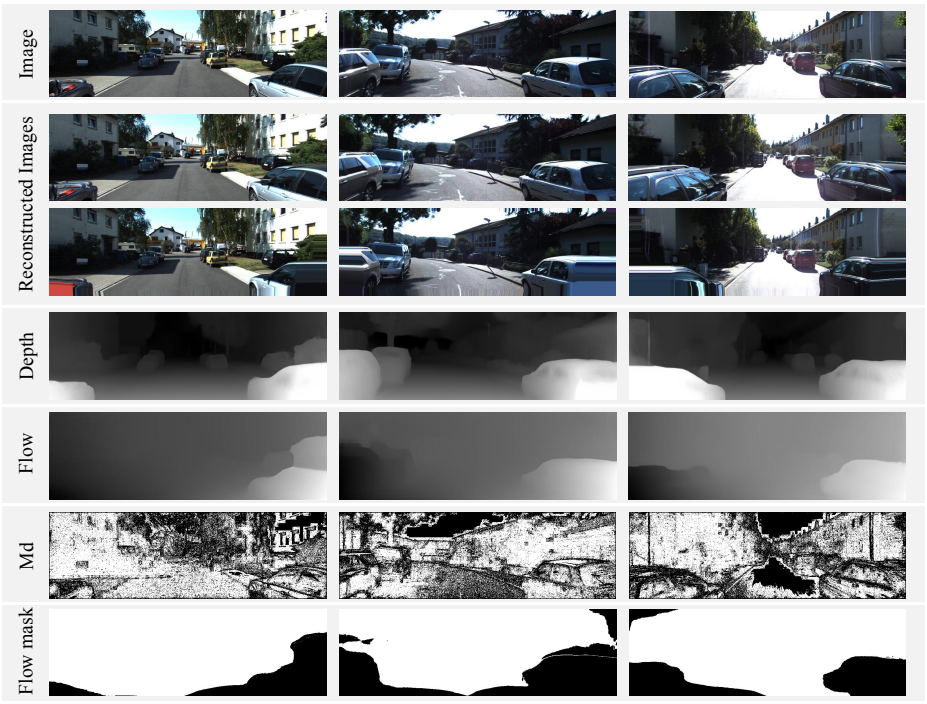

Fig. 3. Visual comparison of the our selective masks as well as the flow masks for dynamic scenes.

References

1. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
2. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
3. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838 (2019)
4. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
5. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: 2017 international conference on 3D Vision (3DV). pp. 11–20. IEEE (2017)