Adaptive Co-Teaching for Unsupervised Monocular Depth Estimation

Weisong Ren[1], Lijun Wang[1⋆], Yongri Piao[1], Miao Zhang[1], Huchuan Lu[1,2], and Ting Liu[3]

[1] Dalian University of Technology
[2] Peng Cheng Laboratory
[3] Alibaba Group
kalilia1102@mail.dlut.edu.cn
{ljwang,yrpiao,miaozhang,lhchuan}@dlut.edu.cn
brooks.lt@alibaba-inc.com

Abstract. Unsupervised depth estimation using photometric losses suffers from local minimum and training instability. We address this issue by proposing an adaptive co-teaching framework to distill the learned knowledge from unsupervised teacher networks to a student network. We design an ensemble architecture for our teacher networks, integrating a depth basis decoder with multiple depth coefficient decoders. Depth prediction can then be formulated as a combination of the predicted depth bases weighted by coefficients. By further constraining their correlations, multiple coefficient decoders can yield a diversity of depth predictions, serving as the ensemble teachers. During the co-teaching step, our method allows different supervision sources from not only ensemble teachers but also photometric losses to constantly compete with each other, and adaptively select the optimal ones to teach the student, which effectively improves the ability of the student to jump out of the local minimum. Our method is shown to significantly benefit unsupervised depth estimation and sets new state of the art on both KITTI and Nuscenes datasets.

Keywords: Unsupervised, Monocular Depth Estimation, Knowledge Distillation, Ensemble Learning

1 Introduction

Monocular depth estimation is a fundamental research task in computer vision, with a wide application ranging from navigation [32], 3D reconstruction [10] to simultaneous localization and mapping [2]. With the rapid development of deep learning techniques, monocular depth estimation based on deep convolutional networks has achieved significant progress in recent years. Nonetheless, training

⋆ Corresponding author.

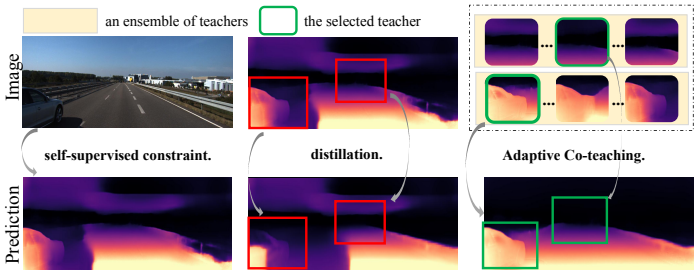

Fig. 1. Strength of our adaptive co-teaching framework. We conduct visualized comparison of our proposed adaptive co-teaching strategy (3rd column), photometric constraint based unsupervised training (1st column), and the straightforward distillation method (2nd column) on KITTI. Our scheme can transfer more accurate knowledge to the student maintaining fine-grained details.

complex deep networks entails large-scale annotated depth data which are expensive to achieve and still very limited in terms of amount and diversity. To alleviate the need of depth annotations, unsupervised approaches have recently been investigated for learning monocular depth using either stereo images [12,14] or monocular videos [15,16], which have shown promising performance.

In this paper, we focus on unsupervised depth estimation using unlabeled monocular videos. Under the assumption of static scenes and Lambertian surfaces, both depth and pose networks can be jointly learned in an unsupervised manner by minimizing the photometric reconstruction losses of view synthesis. However, the above assumption may not hold in many scenarios, leading to unstable unsupervised learning and local minimum issues in dynamic regions and non-Lambertian or low-textured surfaces.

To alleviate this issue, recent works [7,17,24] propose to exploit semantic segmentation labels as prior knowledge to facilitate training. Though impressive performance has been achieved, these methods rely on additional manual annotations, and thus fail to maintain the advantages of unsupervised learning. An alternative idea to mitigate this challenge is to leverage the distillation techniques [30,29,27,28,37]. Although distillation learning can circumvent the unstable training issue, the teacher network is trained using conventional unsupervised framework. As shown in the second column of Figure 1, it still suffers from the mentioned drawbacks, leading to low-quality pseudo labels and degraded final performance.

In light of the above issues, we propose an adaptive co-teaching framework for unsupervised depth estimation, which operates in a two-stage fashion. In the first stage, we train an ensemble of teacher networks for depth estimation in a unsupervised manner. By penalizing the correlation among teacher networks, we obtain a diversity of plausible solutions for depth estimation, providing a significantly higher chance to escape from local minimums of unsupervised learning compared to training a single teacher network. In the second stage, we transfer the learned knowledge from the teacher ensemble to the student network through

a cost-aware co-teaching loss, which allows teacher networks in the ensemble to compete with each other and is able to adaptively select the optimal supervision source to teach the student network.

As another contribution of this work, we present the MUlti-STream ensemble network (MUSTNet) to facilitate more effective teacher ensemble learning. Our MUSTNet comprises a basis decoder and multiple coefficient decoders, where the basis decoder decomposes the depth map into a set of depth bases, and each coefficient decoder predicts a weight vector to linearly combine the depth bases, giving rise to one prediction of the depth map. By integrating both the basis and coefficient decoders in a single network, the MUSTNet serves as a more compact architecture for ensemble networks. In addition, the decomposition of the depth map into depth bases permits a more elegant and convenient way to enforce model diversity within the ensemble. As shown in our experiments, by incorporating the MUSTNet into our co-teaching framework, we achieve new state-of-the-art performance on popular benchmarks.

The contributions of our approach can be summarized into three-folds:

- An adaptive co-teaching framework for unsupervised depth estimation that enjoys the strengths of knowledge distillation and ensemble learning for more accurate depth estimation.
- A novel Multi-Stream Ensemble Network which decomposes depth maps into depth bases weighted by depth coefficients, providing a compact architecture for both the teach ensemble and the student model.
- A cost-aware co-teaching loss which leverages both the ensemble teachers and the photometric constraint to adaptively distill our student network.

Our model outperforms state-of-the-art monocular unsupervised approaches on the KITTI and Nuscenes datasets.

2 Related Work

2.1 Unsupervised Monocular Depth Estimation

Monocular depth estimation is an inherent ill-posed problem. Recently, with the help of Multi-view Stereo or Structure from Motion, some works [43,1,15] propose to tackle this problem within an unsupervised learning manner replacing the need of ground truth annotations.

Unsupervised Stereo Training. [12] proposes the first approach that estimates depth maps in an unsupervised manner with the help of multi-view synthesis. This work provides a basic paradigm for unsupervised depth estimation. After that, [14] employs a view synthesis loss and a depth smoothness loss to further improve the depth estimation performance. Very recently, [38] proposes a two-stage training strategy, which firstly generates pseudo depth labels from input images and secondly refine the network with a self-training strategy. It achieves better performance than state-of-the-art unsupervised methods.

Unsupervised Monocular Training. Different from multi-view based methods, monocular video sequence based methods require additional process to obtain camera poses. [43] jointly estimates depths and relative poses between adjacent frames, and indirectly supervise the depth and pose networks by computing the image re-projecting losses. However, this training strategy highly relies on the assumption that the adjacent frames comprise rigid scenes. Consequently, non-rigid parts caused by moving objects may seriously affect the performance. To solve this problem, [1] introduces an additional network to predict per-pixel invalid masks to ignore regions violating this assumption. [15] proposes a simple yet effective auto-masking and min re-projecting method to solve the problem of moving objects and occlusion. [24] decomposes the motion to the relative camera motion and instance-wise object motion to geometrically correct the projection process. [35] presents a novel tightly-coupled approach that leverages the interdependence of depth and ego motion at training and inference time. Although these methods have achieved matured performance, it is still an open question to solve the problem introduced by the photometric loss.

2.2 Knowledge Distillation

Knowledge distillation is originally proposed by [3] and popularized by [20]. The idea has been exploited for many computer vision tasks [6,18,26] for its ability to compressing a large network to a much smaller one. Recently, some works attempt to exploit distillation for unsupervised depth estimation.

Multi-view Training. [29] propose a self-distillation strategy for unsupervised multi-view depth estimation in which a sub-network of a bidirectional teacher is self-distilled to exploit the cycle inconsistency knowledge. [38] generates pseudo depth labels from the input images and secondly refine the network with a self-training strategy. More recently, [28] also apply a self-distillation method to unsupervised multi-view depth estimation and try to generate pseudo labels based on their proposed post-processing method. However, these multi-view images based methods fail to cope with monocular videos. It is hard for these methods to generate high quality pseudo labels for the camera poses keep unknown.

Monocular Training. For monocular videos, some approaches also attempt to employ knowledge distillation to unsupervised depth estimation. [27] train a complex teacher network in a unsupervised manner and distill the knowledge to a lightweight model to compress parameters while maintain high performance. [37] inference depth from multi-frame cost volume and generate depth prior information with a monocular unsupervised approach to teach the cost volume network in those potentially problematic regions. All the above introduced methods boost unsupervised depth estimation performance. However, those approaches leverage single depth teacher for distillation, which can not contribute to the student network to jump out of local minimal. Different from these methods, our work conducts an adaptive co-teaching strategy that leverages an ensemble of diverse teachers for better distillation.

3 Problem Formulation

During training, we consider a pair of input images: source image I_s and target image I_t of size $H \times W$. Two convolutional networks are leveraged to estimate the depth map of I_t and the relative camera pose $p_{s \to t} = [R|t]$ between I_s and I_t, respectively. After that, a synthesized target image $I_{t'}$ can be generated by rendering the source image I_s with predicted depth d_t, relative pose p_t, and the given camera intrinsic K. As a consequence, the depth and relative pose can be jointly optimized by minimizing the photometric loss given by

$$\mathcal{L}_p = \alpha \left(1 - SSIM\left(I_t, I_{t'}\right)\right) + (1 - \alpha) \left\| I_t - I_{t'} \right\|_1, \tag{1}$$

where $|| * ||_1$ measures the pixel-wise similarity and $SSIM$ indicates the structural similarity to measure the discrepancy between the synthesized and the real images structure. α is the hyper-parameter used to balance these two loss terms. Following [15], we set $\alpha = 0.85$.

Furthermore, in order to mitigate spatial fluctuation, we apply the edge-aware depth smoothness loss used in [15,21] , which can be described as

$$\mathcal{L}_{\mathrm{s}}\ \left(D_i\right) = \left|\delta_x D_i\right| e^{-\left|\delta_x I_i\right|} + \left|\delta_y D_i\right| e^{-\left|\delta_y I_i\right|}. \tag{2}$$

Following [15], for each pixel we optimize the loss for the best matching source image by selecting the per-pixel minimum over the reconstruction loss.

4 Methodology

Motivation and Overview. There may exist multiple feasible solutions to the photometric constraint (1), especially at low-texture regions, leading to training ambiguity and sub-optimal convergence. In this paper, we address this issue by proposing an adaptive co-teaching framework for unsupervised depth estimation. Our philosophy is to firstly learn an ensemble of depth estimation networks in the unsupervised manner, which can deliver a variety of depth prediction for an input image. We then treat these pre-trained networks as teachers and select their optimal predictions as pseudo labels to train the final student network using our cost-aware loss. As a result, training the student network under our co-teaching framework can circumvent the unstable issue of unsupervised learning. As verified in our experiment, through knowledge distillation from ensemble teachers, the student network outperforms each individual teacher network. In the following, we first introduce a new network architecture named MUSTNet for the teacher ensemble in Sec. 4.1. We then elaborate on the proposed adaptive co-teaching framework in Sec. 4.2. Finally, Sec. 4.3 presents the implementation details of our method.

4.1 MUSTNet: MUlit-STream ensemble Network

A conventional approach to building network ensembles is to group a set of independent networks with similar architectures. The memory complexity and

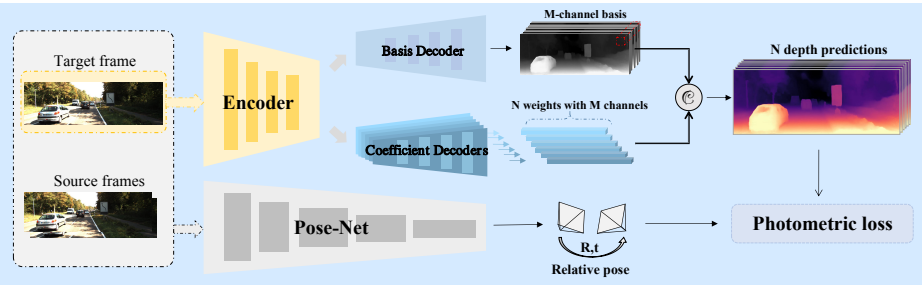

Fig. 2. Illustration of the proposed MUlti-STream ensemble Network.

computational overhead for training such an ensemble of networks will be linearly increased w.r.t. to the amount of ensemble members. Besides, the diversity of the learned network parameters can not be easily guaranteed. Inspired by the above observation, we design a more compact network structure named MUlti-Stream ensemble network (MUSTNet) for monocular depth estimation. In contrast to the above conventional approach, the proposed MUSTNet integrates a number of depth net into a single network. As shown in Figure 2, MUSTNet contains one encoder followed by one basis decoder and N coefficient decoders in parallel. Given an input image, the basis decoder produces an output of M channels with each channel as one depth basis. Meanwhile, each coefficient decoder generates a coefficient vector of M dimensions. Each coefficient can be used as weight to linearly combine the depth basis, producing an estimation of depth map. The N coefficient decoders will then make N predictions of the depth map. As a result, the MUSTNet is equivalent to an ensemble of N depth estimation networks.

Basis Decoder. The basis decoder is designed motivated by prior methods [15,27]. It receives a multi-scale feature pyramid from the encoder, and the features are then progressively combined from coarse to fine level. Different from the basic structure, we extend the output channels of the last layer convolution to M. For each channel, we employ a sigmoid activation function to generate a normalized basis, representing the disparity values.

Coefficient Decoder. The coefficient decoder receives the coarsest output from the depth encoder. It consists of three 3×3 convolution layers followed by non-linear $ReLU$ layers. An additional 3×3 convolution is add to compress feature channels to M, which keeps the same as the number of depth bases. Finally, we use a global average pooling layer to generate the coefficients.

Discussion. Our MUSTNet provides a compact structure for depth estimation, which enjoys more flexibility than conventional network ensembles from multiple aspects. As can be seen in Sec. 4.2, the diversity of the teacher ensemble is essential in our co-teaching framework. By decomposing depth estimation into depth bases and coefficient prediction, our MUSTNet permits an elegant and convenient way (introduced in Sec. 4.2) to enforce ensemble diversity. In addition, the MUSTNet structure is also scalable, i.e., it can not only serve as our teacher ensemble, but is also suitable for the student depth network by using only one coefficient decoder. By using consistent architectures for teachers and student

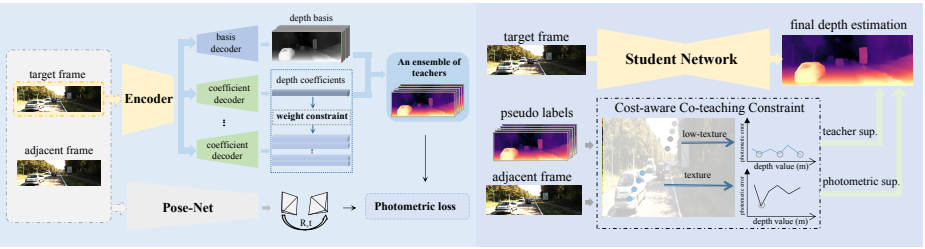

Fig. 3. Illustration of the proposed adaptive co-teaching framework. Specifically, **(a)**: In the first stage, we train the pose network (θ_{pose}) and the depth network (θ_{depth}) in an unsupervised manner. We employ an ensemble network for the depth estimation to generate a diversity of pseudo labels. **(b)**: In the second stage, we transfer the learned knowledge to a student monocular depth network leveraging the proposed cost aware co-teaching loss.

network, we expect the distillation learning to be more coherent, and thus more superior final performance.

4.2 Adaptive Co-teaching Framework

As shown in Figure 3, our adaptive co-teaching framework operates in a two-stage fashion. In the first stage, we train a teacher ensemble through unsupervised learning. By penalizing their correlation, a diversity of teachers can be achieved. The second training stage is introduced to transfer the learned knowledge to the student network through a cost-aware co-teaching loss. It allows teacher networks in the ensemble to compete with each other and can adaptively select the optimal supervision source to teach the student network for much better predictions.

Unsupervised Teacher Ensemble Learning. We adopt the MUSTNet with $N > 1$ coefficient decoders as our teacher ensemble, where each coefficient decoder correspond to an individual teacher. We design the following two constraints to pursue model diversity.

Bases diversity constraint. We prefer the components of the bases have significant different distributions, which can help the network jump out of local minimal solutions. As suggested by [25], we assume the disparity values subject to a Gaussian distribution, and force the generated depth bases to have similar variances and different means for irrelevant outputs

$$L_v = \frac{\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\mu_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\right)^2}, \tag{3}$$

where μ_n and σ_n denote the mean and variance of the n-th depth basis, respectively.

Coefficients orthogonality constraint. For each predicted depth coefficient vector w, we compute its normalized version as $\overline{w} = \frac{w}{|w|}$, and stack all the

normalized coefficients into a matrix W. The correlation of matrix W can be computed as

$$C_w = W \cdot W^T. \tag{4}$$

To penalize the relevance among coefficients, we define the coefficient orthogonality constraint as follows

$$L_w = ||C_w - E||, \tag{5}$$

where E represents the identity matrix.

The final loss function for training the teacher ensemble combines the photometric loss with the above two constraints and can be described as

$$L_{self} = L_p + \alpha L_s + \beta L_v + \gamma L_w, \tag{6}$$

where, α, β and γ are the hyper-parameters. In our work, we set $\alpha = 1e-3$, $\beta = 1e-3$, $\gamma = 1e-5$.

Student Learning via Adaptive Co-teaching. Given the pre-trained teacher ensemble, we learn a student depth network using a cost-aware co-teaching loss, which can not only identify the best teacher for distillation but also adaptively switch between distillation and unsupervised learning to select the optimal supervision sources.

Ensemble distillation loss. To aggregate multiple predictions from the teacher ensemble to synthesize satisfactory pseudo labels for distillation learning, we measure the accuracy of the predicted depth maps using the photometric reconstruction error (1). For each spatial location in the training image, we select the depth value predicted by the teacher ensemble with the minimum reconstruction error as the final pseudo label. The distillation loss is then defined as

$$\mathcal{L}_{distill} = \|d^s - d\|_1 + \sum_i \left[(\nabla_x D_i)^2 + (\nabla_y D_i)^2 \right], \tag{7}$$

where d denotes the depth map predicted by the student network, d^s denotes the synthesized pseudo label, and $D_i = \log(d_i) - \log(d_i^s)$. The first term measures the difference between the predictions and pseudo labels, while the second term enforces the smoothness of the predicted depth maps.

Adaptive switch between supervisions. In order to select the optimal supervision sources for learning the student, we propose to adaptively switch between the distillation learning (7) and unsupervised learning (1). Our basic idea is to estimate the quality of the potential solution yielded by the unsupervised learning. When unsupervised learning is likely to suffer from failure, we then switch to distillation learning. To this purpose, we introduce a cost volume [37,39] based masking mechanism. The cost volume V is constructed using the photometric reconstruction errors of K discrete depth planes. The planes are uniformly distributed over the disparity space. We set K=16 in our experiments. For each spatial location i, we compute the minimum error across all the depth planes as $e_i = \min_k(V_i[k])$. A smaller e_i mostly indicates a high-quality solution. However, this does not hold in low-texture regions, where multiple different solutions can

produce similar errors. To address this issue, we further convert the cost volume into a confidence volume P as

$$P = \text{softmax}\left(\beta \frac{1}{V}\right), \tag{8}$$

where the softmax normalization is performed for each spatial location along the K depth planes, and β is a hyper-parameter that controls the height of local peaks. β is set as 0.05 in our work. The maximum confidence for location i is further computed as $c_i = \max_k(P_i[k])$. For low-texture regions with multiple local minimum errors, their maximum confidences are relatively low (See Figure 3 (b) for an example). Therefore, unsupervised learning is only applicable to regions with low minimum errors and high maximum confidences. We define the selection mask M^u for unsupervised learning as

$$M_i^u = \begin{cases} 1, \text{if } e_i < \tau_e \text{ and } c_i > \tau_c, \\ 0, \qquad \text{otherwise}, \end{cases} \tag{9}$$

where τ_e and τ_c are two pre-defined thresholds. In our work, τ_e and τ_c are set as 0.6, 0.002, respectively. M_i^u indicates the mask value at position i. For regions with large minimum errors, we explore distillation from ensemble teachers for better supervision. However, the pseudo labels for distillation are generated based on photometric reconstruction losses, which are unreliable in low-texture regions. Therefore, the selection mask M^d for distillation learning also discards the low-texture regions and can be defined as

$$M_i^d = \begin{cases} 1, \text{if } e_i \geq \tau_e \text{ and } c_i > \tau_c, \\ 0, \qquad \text{otherwise}. \end{cases} \tag{10}$$

Finally, the cost aware co-teaching loss is formulated as a combination of the ensemble distillation loss and the unsupervised loss spatially weighted by their corresponding selection masks.

4.3 Implementation

Network architecture. For our MUSTNet, we use the ResNet-18 [19] as the encoder. Furthermore, we employ the Redesigned Skip Connection block proposed in [27] to decrease the semantic gap between different scale features and obtain sharper depth details. For the relative pose, we employ the same design as [15], which predicts 6-DoF axis-angle representation. The first three dimensions represent translation vectors and the last three represent Euler angles.

Training details. We implement our models on PyTorch and train on one Nvidia 2080Ti GPU. We use the Adam optimizer [23] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. For the first training stage of the adaptive co-teaching framework, we employ a main-assistant training strategy to pursue the ability of the teachers. We firstly train a teacher network with only one coefficient decoder, namely main teacher, for 20 epochs, with a batch size of 12. As in [15], the initial learning

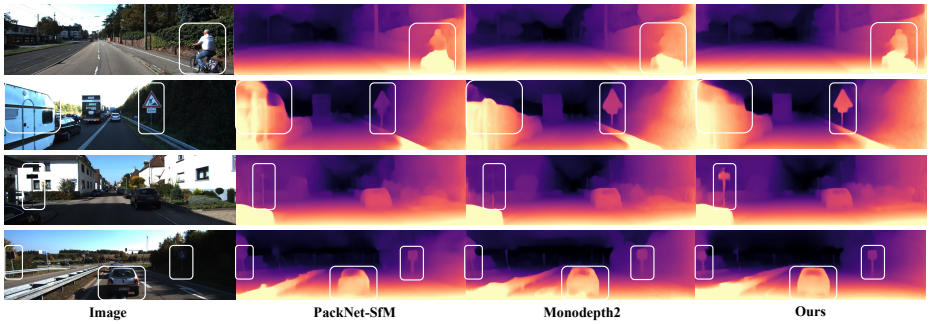

Fig. 4. Visual comparison of our method and recent works. The predicted depth maps of our method are perceptually more accurate with more details.

rate for both depth and pose models is set to 1e-4 and decays after 15 epochs by factor 10. Then, we fix the parameters of the depth and pose networks, and train additional N-1 parallel coefficient decoders, namely assistant teachers, for another 5 epoch with a learning rate of 1e-5. The second stage of the adaptive co-teaching framework is trained for 15 epochs, with a batch size of 8. The learning rate is set to 1e-4. We adopt data augmentation strategies including random color jittering and horizontal flipping to improve generalization ability.

5 Experiments

In this section, we evaluate our proposed approach on two publicly available datasets, and perform ablation studies to validate the effectiveness of our design. Source code will be released at `https://github.com/Mkalilia/MUSTNet`.

5.1 Datasets

KITTI. The KITTI benchmark [13] is the most widely used dataset for training and test monocular depth methods. We employ the training and test split of [8]. As in [15], we use the pre-processing strategy to remove static frames. In particular, we use 39810 monocular triplets for training, 4424 for validation and 697 for evaluation. We follow [15], which uses the same intrinsics for all the images and sets the camera principal point to the image center. The focal length is set as the average of that of all the samples in KITTI.

Nuscenes. Nuscenes [4] is a recently released 3D dataset for multiple vision tasks. It carries the full autonomous vehicle sensor suite: 6 cameras, 5 radars and 1 lidar, all with full 360 degree field of view. We only use this dataset for evaluation to assess the generalization capability of our approach. Based on [16], we evaluate our network on the official NuScenes-mini train-val dataset, which contains 404 front-facing images with ground-truth depth.

	Methods	Sup.	Backbone.	Error metric↓				Accuracy metric↑		
				$AbsRel$	$SqRel$	$RMSE$	RMS_{lg}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Lower	SfMLearner[43]	M	DispNet	0.198	1.836	6.565	0.275	0.718	0.901	0.960
	Struct2D[5]	M	ResNet18	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	Geo-Net[40]	MF	ResNet50	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Standard Resolution	SC-SfM[1]	M	DispNet	0.128	1.047	5.234	0.208	0.846	0.947	0.976
	MD2[15]	M	ResNet18	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	SAFE-Net[7]	MS	ResNet18	0.112	0.788	4.582	0.187	0.878	0.963	**0.983**
	HR-Depth[27]	M	ResNet18	0.109	0.792	4.632	0.185	0.884	0.962	**0.983**
	SCSI[36]	M	ResNet18	0.109	0.779	4.641	0.186	0.883	0.962	0.982
	Ours	M	ResNet18	**0.106**	**0.763**	**4.562**	**0.182**	**0.888**	**0.963**	**0.983**
	Zhou et al.[42]	M	ResNet50	0.121	0.837	4.945	0.197	0.853	0.955	0.982
	PackNet[16]	M	PackNet	0.111	0.785	4.601	0.189	0.878	0.960	0.982
	Adrian et al.[22]	M	ResNet101	0.106	0.861	4.699	0.185	0.889	0.962	0.982
Higher	Zhao et al.[41]	MF	ResNet18	0.113	**0.704**	4.581	0.184	0.871	0.961	0.984
	Lee et al.[24]	MS	ResNet18	0.112	0.777	4.772	0.191	0.872	0.959	0.982
	MD2[15]	M	ResNet18	0.115	0.882	4.701	0.190	0.879	0.961	0.982
	Fang et al.[11]	M	ResNet18	0.110	0.806	4.681	0.187	0.881	0.961	0.982
	HR-Depth[27]	M	ResNet18	0.106	0.755	4.472	0.181	0.892	**0.966**	**0.984**
	Ours	M	ResNet18	**0.104**	0.750	**4.451**	**0.180**	**0.895**	**0.966**	**0.984**
	PackNet[16]	M	PackNet	0.107	0.802	4.538	0.186	0.889	0.962	0.981
	Chang et al.[33]	M	ResNet50	0.104	0.729	4.481	0.179	0.893	0.965	0.984

Table 1. Quantitative results on KITTI dataset for distance up to 80m. M refers to methods supervised by monocular videos. **MS** refers to methods supervised by monocular videos and semantic information. **MF** refers to methods that jointly train depth and optical flow network. At test time, we scale depth with median ground-truth LiDAR information. Best results are marked **bold**.

5.2 Quantitative Evaluation

Results on KITTI Eigen split. We report the performance of our network on KITTI raw data with the evaluation metrics described in [9]. We evaluated our model at standard resolution (192×640) and high resolution (1024×320). We compare our model with state-of-the-arts. Results in Table 1 show the superiority of our approach compared with all existing ResNet-18 based unsupervised approaches [15,27,36,11]. We also outperform recent models [16,42,33,22] with much larger backbones and recent works using additional semantic information [7,24] or optical flow information [24,41]. Furthermore, we showcase the qualitative comparison in Figure 4. The depth predictions of our method are perceptually more accurate depth information with sharper depth details.

Generalization Capability. We also evaluate our approach on the recently proposed nuScenes dataset[4]. As shown in Table 2, our approach outperforms state-of-the-arts which confirms the generalization ability of our method across a large spectrum of vehicles and countries.

Methods	Sup.	Error metric↓				Accuracy metric↑		
		$AbsRel$	$SqRel$	$RMSE$	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SC-SfMLearner[43]	M	0.210	2.257	9.358	0.316	0.677	0.868	0.936
MD2[15]	M	0.199	2.236	9.316	0.311	0.697	0.869	0.936
HR-Depth[27]	M	0.196	2.191	8.894	0.308	0.702	0.869	**0.937**
Ours	M	**0.192**	**2.143**	**8.888**	**0.305**	**0.716**	**0.870**	0.936

Table 2. Quantitative results of depth estimation on nuScenes dataset at the standard resolution. Best are marked **bold**.

5.3 Ablation Studies

We conduct multiple ablative analysis on our approach, to further study the performance improvements provided by each component.

Ablation study for MUSTNet. As one of our main contributions, MUSTNet serves as the architecture of both the teacher ensemble and the student model.

The teacher ensemble: We compare three variances of the model in an unsupervised manner, shown in L1-3 in Table 3. *Baseline* employs a single stream depth decoder outputting one channel prediction which keeps the same as [15,27]. $+BD$ extends the decoder to the basis decoder with 16 output channels. The final outputs are generated by averaging all the channels. Compared with *Baseline*, both +BD and MUSTNet improve the performance. With 16 channel bases, more diverse representations of depth are embedded to guide the network out of local minimums and converge to a better solution. +BD is a special case of MUSTNet with a fixed weight $\frac{1}{N}[1, 1, ..., 1]^T$ for all frames. MUSTNet learns more appropriate representations by co-adapting bases and coefficients. Higher quality pseudo-labels can be generated leveraging the MUSTNet as the teacher model. Meanwhile, as shown in L4-5, the student model learned from the MUSTNet also beats that from the *Baseline*.

The student model: As discussed in Sec 4.1, the proposed MUSTNet is also suitable for student network by using one coefficient decoder. Compared with *baseline*, it provides an elegant way to adaptively learn feature channels and perform high quality predictions by dividing the single channel output into depth bases. As shown in L5-6 in Table 3, the results of a MUSTNet student are much better than the results of a *Baseline* student.

Ablation study for the Adaptive Co-teaching scheme. In this part we conduct 4 different training schemes on the MUSTNet to compare our adaptive co-teaching framework with the conventional unsupervised method and other distillation methods[27,31]. The results can be seen in L6-9 in Table 3.

Distill with a single teacher: L6 represents that we train only one teacher model for distillation. It is a straight forward distillation method, which is used in [27]. However, the improvement over the unsupervised baseline (L3) is marginal. The student just try to regress to pseudo labels of the teacher and can not learn more robust feature representations resulting in sub-optimal solutions.

Distill with teacher ensemble: Then we use the method provided by [31] to distill the student, shown in L7. Specifically, we train N randomly initialized

Model	P-Cons.	D-Cons.	AS.	TS.	NT.	T-Model.	PM.	*Error* ↓ *AbsRel*	*RMSE*	*Accuracy* ↑ $\delta < 1.25$
Baseline	✓			1				0.113	4.795	0.880
$\dagger + BD$	✓			1				0.112	4.808	0.881
MUSTNet	✓			1				0.109	4.656	0.883
Baseline		✓		2	1	*Baseline*	1×	0.113	4.801	0.879
Baseline		✓		2	1	*MUSTNet*	1×	0.112	4.744	0.882
MUSTNet		✓		2	1	*MUSTNet*	1×	0.110	4.718	0.883
MUSTNet		✓		2	4	*MUSTNet*	4×	0.108	4.622	0.885
MUSTNet	✓	✓		2	4	*MUSTNet*	1×	0.108	4.574	0.886
MUSTNet	✓	✓	✓	2	4	*MUSTNet*	1×	**0.106**	**4.562**	**0.888**

Table 3. Ablation study of the Adaptive Co-teaching Framework. P-Cons denotes photometric constraint. **D-Cons** denotes distillation constraint with depth pseudo labels generated by **T-Model**. **AS** denotes adaptively switch between supervisions. **TS** denotes the total training stages. **NT** denotes number of teachers. **PM** denotes the teacher has N× parameters of the student network. Results for different variants of our method are trianed at 192×640.

network with the same architecture (baseline) and obtain empirical mean as the pseudo depth. The parameters of this bootstrapped ensemble architecture is N times of our compact ensemble, while the performance gain is still marginal.

Distill with our adaptive co − teaching framework: The teacher ensemble can be easily acquired by training a MUSTNet with a basis decoder and N coefficient decoders. We distill the student model with both photometric loss and pseudo label constraint, shown in L8. Then we further add the selective masks for the final cost-aware co-teaching loss, shown in L9. Compared with other distillation methods, the performance gain brought by our framework is significant with much lighter teacher ensemble and much simpler training process.

Visualized comparison. For the sake of the pages limit, we report more visualized comparisons in the supplementary material. Specifically, i) the visualization of the depth bases. ii) the comparison of the outputs of the teacher ensemble and student model. iii) more quantitative and qualitative results.

5.4 Extension of our work

While our method significantly improves the performance of our proposed model, it also retains the advantages of being integrated into other models with two simple modifications: (i) replace the single stream depth decoder with our multi-stream decoder format, and (ii) distill a dual-stream student model leveraging our adaptive co-teaching scheme.

We conduct experiments on one famous unsupervised method [15], and the quantitative results can be seen in Table 4. More details of this experiment are presented in our supplementary materials. Our framework improve the *AbsRel* from 0.115 to 0.103 at high resolution, which further verifies our contribution.

	Methods	Sup.	TS.	Error metric↓				Accuracy metric↑		
				$AbsRel$	$SqRel$	$RMSE$	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
192*640	MD2[15]	P	1	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	Dual-stream	P	1	0.111	0.829	4.782	0.189	0.878	0.960	0.982
	Dual-stream	A	2	**0.108**	**0.795**	**4.649**	**0.185**	**0.885**	**0.962**	**0.983**
320*1024	MD2[15]	P	1	0.115	0.882	4.701	0.190	0.879	0.961	0.982
	Dual-stream	P	1	0.108	0.818	4.591	0.185	0.890	0.963	0.982
	Dual-stream	A	2	**0.103**	**0.745**	**4.451**	**0.180**	**0.894**	**0.965**	**0.983**

Table 4. Quantitative results on [13]. **P** denotes photometric constraint. **TS** denotes the total training stages. **A** represents our adaptive co-teaching scheme.

5.5 Discussion

Dynamic Mask. 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 further conduct an extensive experiment to mask out those dynamic parts with the help of additional optical flow information generated by [34]. The detailed ablation study is shown in our supplementary material.

Moving object. Despite good performance achieved by our adaptive co-teaching framework, one limitation appears on the moving objects that keep station relative to the camera. We have considered to mask out potential dynamic regions leveraging additional optical flow information. However, our framework essentially follows the static scene assumption which makes it fails to distinguish relative station parts and infinity. We support that, under the unsupervised learning pipeline, additional semantic guidance is essential to cope with this situation.

Video inference. Another limitation appears on the scale inconsistency of video inferring. For the lack of pose ground truth, the scene scale keeps unknown. We tried to train our network in a scale consistent manner leveraging the geometric loss proposed in [43]. However, the joint depth-pose learning process is affected for it enforcing a consistent scale across all images.

6 Conclusion

The core design of our work is an adaptive co-teaching framework, which aims to solve the problem of training ambiguity and sub-optimal convergence for unsupervised depth estimation. We first design a compact ensemble architecture, namely MUlti-STream ensemble network, integrating a depth basis decoder with multiple depth coefficient decoders. Meanwhile, we propose a cost-aware co-teaching loss to transfer the learned knowledge from the teacher ensemble to a student network. As verified in our experiment, training the student network under our framework can circumvent the unstable issue of unsupervised learning. Our method sets new state-of-the-art on both KITTI and Nuscenes datasets.

Acknowledgements This research is supported by National Natural Science Foundation of China (61906031, 62172070, U1903215, 6182910), and Fundamental Research Funds for Central Universities (DUT21RC(3)025).

References

1. Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in neural information processing systems **32**, 35–45 (2019)
2. Bloesch, M., Czarnowski, J., Clark, R., Leutenegger, S., Davison, A.J.: Codeslam—learning a compact, optimisable representation for dense visual slam. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2560–2568 (2018)
3. Bucila, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'06) (2006)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
5. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8001–8008 (2019)
6. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems **30** (2017)
7. Choi, J., Jung, D., Lee, D., Kim, C.: Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. In: Thirty-fourth Conference on Neural Information Processing Systems, NIPS 2020. NeurIPS (2020)
8. 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)
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014. pp. 2366–2374. Neural information processing systems foundation (2014)
10. Fan, H., Hao, S., Guibas, L.: A point set generation network for 3d object reconstruction from a single image. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
11. Fang, J., Liu, G.: Self-supervised learning of depth and ego-motion from video by alternative training and geometric constraints from 3d to 2d. arXiv preprint arXiv:2108.01980 (2021)
12. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European conference on computer vision. pp. 740–756. Springer (2016)
13. 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)
14. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
15. 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)

16. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2485–2494 (2020)
17. Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. In: International Conference on Learning Representations (2019)
18. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2827–2836 (2016)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
21. Janai, J., Guney, F., Ranjan, A., Black, M., Geiger, A.: Unsupervised learning of multi-frame optical flow with occlusions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 690–706 (2018)
22. Johnston, A., Carneiro, G.: Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 4756–4765 (2020)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Lee, S., Im, S., Lin, S., Kweon, I.S.: Learning monocular depth in dynamic scenes via instance-aware projection consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1863–1872 (2021)
25. Li, S., Wu, X., Cao, Y., Zha, H.: Generalizing to the open world: Deep visual odometry with online adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13184–13193 (2021)
26. Li, Y., Yang, J., Song, Y., Cao, L., Li, L.J.: Learning from noisy labels with distillation. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
27. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: Hr-depth: high resolution self-supervised monocular depth estimation. CoRR abs/2012.07356 (2020)
28. Peng, R., Wang, R., Lai, Y., Tang, L., Cai, Y.: Excavating the potential capacity of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15560–15569 (2021)
29. Pilzer, A., Lathuilière, S., Sebe, N., Ricci, E.: Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
30. Pilzer, A., Xu, D., Puscas, M., Ricci, E., Sebe, N.: Unsupervised adversarial depth estimation using cycled generative networks. In: 2018 International Conference on 3D Vision (3DV). pp. 587–595. IEEE (2018)
31. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3227–3237 (2020)
32. Prucksakorn, T., Jeong, S., Chong, N.Y.: A self-trainable depth perception method from eye pursuit and motion parallax. Robotics and Autonomous Systems **109**, 27–37 (2018)

33. Shu, C., Yu, K., Duan, Z., Yang, K.: Feature-metric loss for self-supervised learning of depth and egomotion. In: European Conference on Computer Vision. pp. 572–588. Springer (2020)
34. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
35. Wagstaff, B., Peretroukhin, V., Kelly, J.: Self-supervised structure-from-motion through tightly-coupled depth and egomotion networks. arXiv preprint arXiv:2106.04007 (2021)
36. Wang, L., Wang, Y., Wang, L., Zhan, Y., Wang, Y., Lu, H.: Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12727–12736 (2021)
37. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1164–1174 (2021)
38. Yang, J., Alvarez, J.M., Liu, M.: Self-supervised learning of depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7526–7534 (2021)
39. Yang, J., Mao, W., Alvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4877–4886 (2020)
40. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1983–1992 (2018)
41. Zhao, W., Liu, S., Shu, Y., Liu, Y.J.: Towards better generalization: Joint depth-pose learning without posenet. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
42. Zhou, J., Wang, Y., Qin, K., Zeng, W.: Unsupervised high-resolution depth learning from videos with dual networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6872–6881 (2019)
43. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)