2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

FG-Depth: Flow-Guided Unsupervised Monocular Depth Estimation

Junyu Zhu[1], Lina Liu[1*], Yong Liu[1*], Wanlong Li[2], Feng Wen[2] and Hongbo Zhang[2]

***Abstract*— The great potential of unsupervised monocular depth estimation has been demonstrated by many works due to low annotation cost and impressive accuracy comparable to supervised methods. To further improve the performance, recent works mainly focus on designing more complex network structures and exploiting extra supervised information, e.g., semantic segmentation. These methods optimize the models by exploiting the reconstructed relationship between the target and reference images in varying degrees. However, previous methods prove that this image reconstruction optimization is prone to get trapped in local minima. In this paper, our core idea is to guide the optimization with prior knowledge from pretrained Flow-Net. And we show that the bottleneck of unsupervised monocular depth estimation can be broken with our simple but effective framework named FG-Depth. In particular, we propose (i) a flow distillation loss to replace the typical photometric loss that limits the capacity of the model and (ii) a prior flow based mask to remove invalid pixels that bring the noise in training loss. Extensive experiments demonstrate the effectiveness of each component, and our approach achieves state-of-the-art results on both KITTI and NYU-Depth-v2 datasets.**

I. Introduction

Accurate depth estimation is critical for many applications in computer vision, such as robotic perception [1], [2], augmented reality [3], and 3D modeling [4]. Monocular depth estimation has become a challenging and promising field, attracting the attention of many researchers. Recently, deep learning-based monocular depth estimation [5], [6], [7], [8], [9] has been able to achieve high accuracy by narrowing the gap between predicted depth and ground truth. However, these methods are limited by the expensive annotation cost. The emergence of self-supervised approaches [10], [11], [12] addresses problems requiring depth annotations, typically trained using the photometric loss to reconstruct and warp images between target frames and source frames from monocular videos, stereo pairs, or stereo videos. The photometric loss widely used in self-supervised depth estimation is based on implicit assumptions [10] that 1) the scene is static; 2) no occlusion occurs between target frames and source frames; 3) the surface is Lambertian. However, these assumptions are so hard to be met on real data that the optimization of photometric loss is prone to be trapped in local minima [13] and the performance of the model is limited.

Regarding the above problems, Waston et al. [13] used SGM algorithm results as depth hints to guide the model to reach better minima. Zhan et al. [15] and Shu et al. [16] proposed the feature reconstruction loss to make the training loss more sensitive to low-texture regions and more robust to illumination change. [10] uses stereo pair based photometric loss to avoid the influence of dynamic objects in image warping. Other methods ignore the defects of photometric loss and try to improve depth estimation performance by introducing additional semantic segmentation constraints, such as [17]. They use semantic segmentation constraints to further the depth quality near object boundaries. However, annotating semantic segmentation in real data is expensive. Although low-cost semantic segmentation labels can be easily obtained from synthetic data, existing semantic segmentation models trained on synthetic data cannot generalize well to real data due to the domain shift [18].

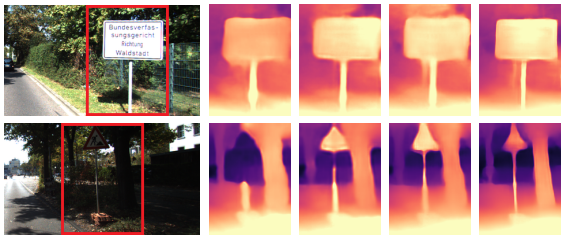

Fig. 1. **Comparison with state-of-the-art methods.** (a)MonoDepth2 [11], (b)DepthHints [13], (c)EPCDepth [14], (d)Ours.

Despite many improvements, almost all existing unsupervised methods rely heavily on photometric loss with non-negligible defects, and performance has reached a bottleneck. Therefore, we believe that it is hard to make significant progress if the training process still relies on typical photometric loss.

In order to break the bottleneck of unsupervised monocular depth estimation, inspired by [19], in this paper, we design a new loss to replace the widely used photometric loss. Similar to [19], our Depth-Net also learns monocular depth by distilling prior knowledge of an optical flow estimation network that has strong generalization and can still generalize well to real data when trained on low-cost synthetic data. But there are three main differences between our method and [19]. Besides the intuitive depth level, our proposed loss further restricts the Depth-Net from the color level. And we propose a mask to filter out out-of-range pixels at distillation time to accelerate convergence. Also, our network architectures are different from theirs.

Our framework is trained based on stereo pairs to avoid influence from moving objects. Firstly, based on the fact that the depth pseudo labels can be generated from prior flow for stereo pairs and the warping procedure in computing photometric loss is based on flow which can be synthesized

[1]Junyu Zhu, Lina Liu and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China. email:{junyuzhu, linaliu}@zju.edu.cn, yongliu@iipc.zju.edu.cn.

[2]Wanlong Li, Feng Wen and Hongbo Zhang are with Noah's Ark Lab, Huawei Technologies, Beijing, China. email:{liwanlong, wenfeng3, zhanghongbo888}@huawei.com.

*Corresponding authors: Lina Liu and Yong Liu.

from depth estimation or directly predicted by pretrained optical flow estimation network, we propose a flow distillation loss to restrict model from depth and color levels. Secondly, Depth-Net is usually only effective within a certain range, and pixels outside the estimated range also inhibit training because they always fail to match the corresponding pixels in the warping procedure. In the previous method [11], [20], there are some pixels beyond the above range that cannot be removed by the previous methods during the training. To this end, taking full advantage of the prior flow, we propose a mask to remove pixels outside the estimation range. Fig. 1 shows that our approach (Fig. 1(d)) can break the bottleneck of self-supervised monocular depth estimation compared to other methods(Fig. 1(a) to (c)).

To summarize, our main contributions are:

- We introduce a flow distillation loss for restricting the model from depth and color levels to replace the typical photometric loss.
- We propose a prior flow based mask for pixels out of the estimation range at distillation time to improve performance.
- The proposed model achieves state-of-the-art performance on the KITTI and NYU-Depth-v2 datasets.

II. Related works

A. Supervised Monocular Depth Estimation

For the task of supervised monocular depth estimation, ground truth depth labels are used to supervise the training of the model with RGB monocular images as input. The ground truth depth labels are usually captured with lidars or RGBD cameras which have the disadvantage of high cost or limited depth range and usage scenario. Most supervised methods regard monocular depth estimation as a regression task. Eigen et al. [21] is the first to employ CNN in supervised monocular depth estimation. Xu et al. [22] applied CRF to optimally combine multi-scale information derived from the inner layers of CNN to improve the performance of a CNN depth estimator. Fu et al. [5] found that a performance improvement can be achieved when the depth estimation is regarded as a classification task. With more complex and well-designed CNN architectures, [6], [7] refreshed previous records. And in recent years, thanks to the development of ViT [23], several models [24], [8], [9], [25] have been proposed to help the accuracy reach new heights.

B. Unsupervised Monocular Depth Estimation

To avoid the expensive cost of depth annotation, unsupervised methods usually use photometric loss between adjacent frames to train the model. As the earliest works in the self-supervised depth estimation field, [10] uses stereo pairs to train the Depth-Net and [12] is trained by monocular video with an extra Pose-Net to predict the relative pose between adjacent frames. [12] introduced a mask predicted by the network to reduce the influence of occlusions and moving objects. Yin et al. [26] and Wang et al. [27] excavated more geometric constraints by learning the depth and flow jointly in an unsupervised manner. Godard et al. [11] made a noticeable improvement by proposing a minimum photometric loss to handle occlusions, an auto mask to ignore pixels that violate camera motion assumptions, and a full-resolution multi-scale sampling method to make the prediction more accurate. Noting that the model can struggle during the training to find the global optimum when minimizing the photometric loss because of low-texture regions and illumination change, [13] introduced SGM algorithm results as extra constraints in training loss and [15], [16] took reconstruction on feature level into consideration. To further improve the performance, some works [28], [17], [18] brought extra semantic segmentation constraints into the training loss but extra semantic labels on real data actually increase the burden of annotation. And recently, Peng et al. [14] introduce an effective data augmentation method for stereo-based models.

C. Optical Flow Estimation

Optical flow estimation is the task of estimating per-pixel displacement between two frames. Recently, many deep learning based approaches [29], [30], [31], [32], [33] have been proposed for optical flow estimation. In supervised optical flow estimation tasks, models are usually trained on synthetic data that has dense accurate optical flow labels. After the training on the synthetic data, they can usually generalize well on the real data. As an exceptional case of optical flow estimation, stereo matching has additional constraints that the displacement is always negative along the horizontal direction and always zero along the vertical direction. Also, many stereo matching models trained on synthetic data can generalize well on real data.

III. Method

A. Method Overview

In this paper, we propose a flow distillation loss to replace the typical photometric loss (introduced in Sec. III-B), a prior flow based mask to remove pixels out of the estimation range for self-supervised depth estimation networks. Our goal is to train depth networks using stereo pairs and constrain them with our proposed flow distillation loss and prior flow based mask. The pipeline of our framework is shown in Fig. 2. In the following subsections, we first introduce the typical photometric loss and automatic mask in Sec. III-B, then describe our flow distillation loss in Sec III-C. and prior flow based mask in Sec III-D.

B. Photometric Loss and Automatic Mask

Previous stereo-based self-supervised works typically use photometric loss L_p to train the model, assuming that the surface is Lambertian and has no occlusion [12]. L_p is between target frame I_t and synthesized frame I_t^s which is interpolated on source frame I_s using predicted depth and relative camera pose, and is defined as:

$$\begin{aligned} L_p &= pe(I_t, I_t^s) \\ &= \frac{\alpha}{2}(1 - SSIM(I_t, I_t^s)) + (1-\alpha)\|I_t - I_t^s\|_1 \end{aligned}^{1} \quad (1)$$

[1]SSIM [34] is computed over a 3×3 pixel window, and $\alpha = 0.85$.

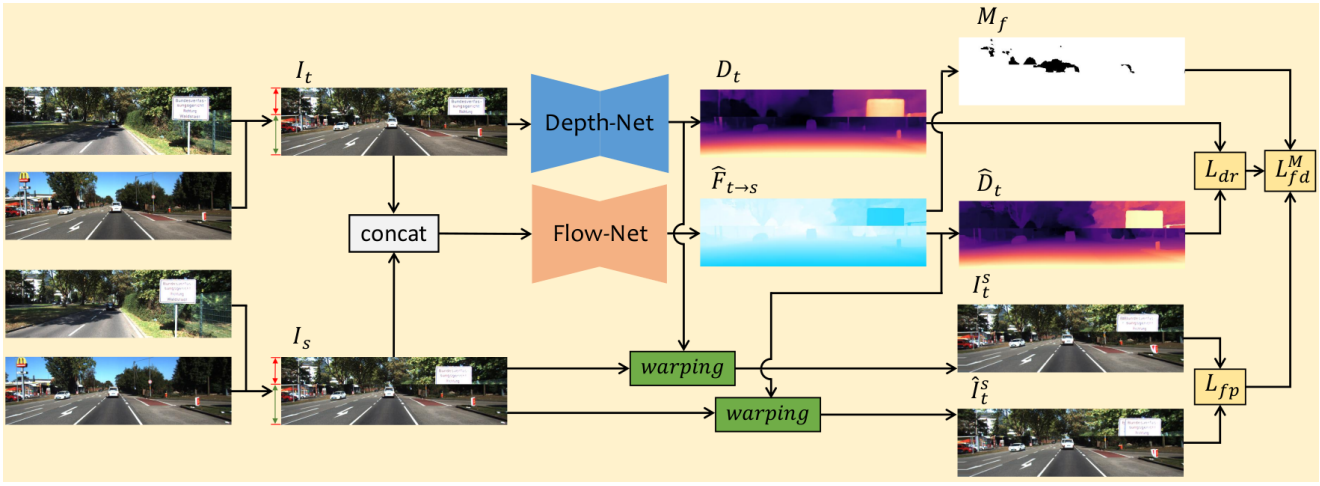

Fig. 2. **Framework illustration.** Given a real stereo pair (I_t, I_s) that is refactored from training data by data grafting [14], the Flow-Net pretrained on synthetic data infers the prior flow $\hat{F}_{t\to s}$ that can then be converted to the pesudo depth lable $\hat{D}_t$ and the mask M_f that removes those pixels out of estimation range. Multi-scale depth maps are estimated by the Depth-Net from I_t and here we only draw the maximum scale depth map D_t as an example. I_t^s and $\hat{I}_t^s$ are synthesized from I_s using D_t and $\hat{D}_t$ respectively by inverse warping. Depth regression loss L_{dr} is computed between D_t and $\hat{D}_t$. Flow-guided photometric loss L_{fp} is computed between I_t^s and $\hat{I}_t^s$. Finally, total training loss L_{fd}^M is calculated using L_{dr}, L_{fp} and M_f.

Also widely used is the automatic mask M_p for occlusion proposed in [11], which is formulated as:

$$M_p = [pe(I_t, I_t^s) < pe(I_t, I_s)]^{\,2} \tag{2}$$

It is still difficult to minimize the L_p in real data. This is because there are multiple local minima with similar magnitudes, especially in regions of low texture and not fulfilling the assumption of color consistency [15]. And the L_p is disturbed by occluded pixels. But it's hard to remove occluded pixels completely by M_p based on a simple comparison of geometry relationships. Also, removing occluded pixels means less supervision information. Furthermore, pixels out of estimation range can inhibit training but are usually neglected by previous methods. To avoid the above problems, we propose a flow distillation loss and a prior flow based mask. Our framework can achieve better performance due to an easier optimized loss function and more reliable supervision information.

C. Flow Distillation Loss

The flow distillation loss L_{fd} consists of depth regression loss L_{dr} and flow-guided photometric loss L_{fp}:

$$L_{fd} = L_{dr} + L_{fp} \tag{3}$$

The L_{dr} is given by

$$L_{dr} = log(|D_t - \hat{D}_t| + 1) \tag{4}$$

The D_t is the depth estimation and the pesudo depth lable $\hat{D}_t$ is computed from

$$\hat{D}_t = \frac{f_x b}{|\hat{F}_{t\to s}|} \tag{5}$$

where f_x is the focal length of the camera and b is the baseline of the stereo.

[2] [] here is the Iverson bracket.

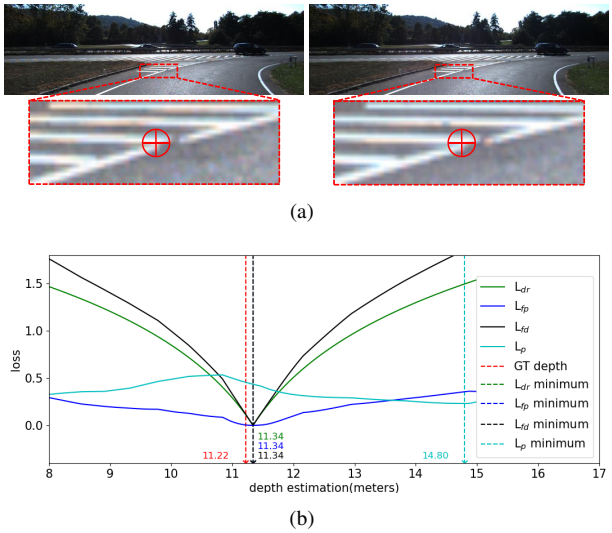

Fig. 3. **Loss visualization.** (a)A pair of matching pixels on the left image and the right image. (b)The relationship between the loss and the depth of matching points in the left subplot (a). The flow distillation loss is easier for optimization and has a more accurate global minimum when compared with photometric loss.

And the L_{fp} is adopted to express the discrepancy between the reconstructions from I_s respectively using D_t and $\hat{D}_t$:

$$\begin{cases} L_{fp} = |I_t^s - \hat{I}_t^s| \\ I_t^s = f_w(I_s, D_t) \\ \hat{I}_t^s = f_w(I_s, \hat{D}_t) \end{cases} \tag{6}$$

where f_w is the differentiable inverse warping operation.

The inspiration behind flow-guided photometric loss L_{fp} is that the warping procedure in computing photometric loss is based on the flow which can be synthesized from depth estimation or directly predicted by pretrained optical flow estimation network. Optimization of L_{fp} can be easier to reach better global minima because when L_{fp} reaches

minima, synthesized flow is closed to the predicted flow. And pretrained optical flow estimation network can predict enough accurate flow, so depth estimation can be more closed to ground truth after optimization of L_{fp}. By contrast, typical L_p is harder to be optimized because of illumination changes.

The intuitive display is shown in Fig. 3. The red pixels marked in Fig. 3(a) are the pixel pairs matched by the stereo pair. Fig. 3(b) shows the loss curve for optimizing the depth of this matching point with depth regression loss L_{dr}, flow-guided photometric loss L_{fp}, flow distillation loss L_{fd} and photometric loss L_p, respectively. There is the same global minimum for L_{dr}, L_{fp}, and L_{fd}, which is almost identical to the ground truth. And the L_{fd} curve is steeper than L_{dr} and L_{fp}, so it makes optimization of L_{fd} easier. For L_p, there are multiple minimum points, where the optimal point fails to learn the correct depth value. Therefore, the L_{fd} is easier for optimization and has a more accurate global minimum, when compared with the L_p.

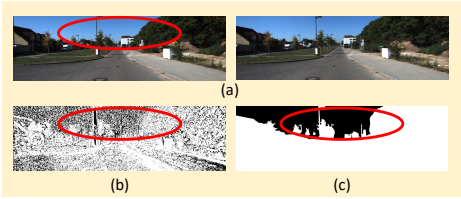

Fig. 4. **Masks visualization.** (a)Input target frame and source frame. (b)Auto-mask M_p proposed in [11]. (c)Prior flow based mask.

D. Prior Flow based Mask

We use a prior flow based mask M_f to remove those pixels out of range by checking the length of prior rigid flow. The mask value $M_f(p_i)$ of the pixel at position p_i can be formulated as:

$$\begin{cases} M_f(p_i) = \begin{cases} 1 & , \ p_i \in V \\ 0 & , \ else \end{cases} \\ V = \left\{ p_t \middle| |\hat{F}_{t\to s}(p_i)| > \dfrac{f_x b}{\delta} \right\} \end{cases} \tag{7}$$

where $\hat{F}_{t\to s}$ denotes the prior flow from target frame to source frame and δ are set to 80.

In depth estimation, out-of-range depths (greater than 80m in KITTI) drop dramatically in accuracy. In previous works, masks do not remove all those pixels out-of-range. So, noise is brought in photometric loss because those out-of-range pixels always fail to match corresponding pixels in the warping procedure. Visualization results in Fig. 4 intuitively show that compared with automatic mask M_p proposed in [11], our mask M_f can filter out out-of-range pixels more completely, making it more stable and less susceptible to noise interference.

E. Final Training Loss

We combine the flow distillation loss and prior flow based mask as:

$$L_{fd}^{M} = \frac{1}{T} \sum_{i} M_f(p_i) L_{fd}(p_i) \tag{8}$$

where T denotes the number of pixels reserved by the mask, and average over each scale.

F. Network Architecture

We implement the Flow-Net with RAFT-Stereo [35] which is based on GRU [36] and has excellent accuracy and good generalization ability. For simplicity, we directly use the official model[3] that is pretrained on Scene Flow dataset [37].

For the Depth-Net, we use the same architecture as [14] which uses ResNet18 as backbone and RSU block as the bridge between different scale features and disparity prediction blocks to output full-scale predictions. The outputs σ of the prediction blocks are further constrained between 0.1 and 80 units with $D = 1/(a\sigma + b)$.

IV. Experiments

In this section, we evaluate our proposed model on the KITTI dataset [38] to verify its state-of-art performance and we validate the generalization ability of our model on the NYU-Depth-v2 dataset [39]. Furthermore, we conduct an ablation study to demonstrate the effectiveness of our contributions.

A. Datasets

a) KITTI: KITTI dataset was captured by a driving vehicle with cameras and depth sensors around the mid-size city of Karlsruhe, in rural areas, and on highways. It is widely used for outdoor monocular depth estimation and we use the Eigen split [21] that consists of 22600 stereo image pairs for training and 697 images for testing. The train set is from 32 scenes and the test set is from other 29 scenes.

b) NYU-Depth-v2: NYU-Depth-v2 dataset was collected with a Kinect sensor in total of 582 indoor scenes. To validate the generalization ability of our model, we use the official test set that consists of 654 images with depth GTs.

B. Implementation Details

Our work is implemented in PyTorch on one Nvidia Tesla V100 GPU. For training, we use the Adam optimizer [40](β_1 = 0.9, β_2 = 0.999). The total number of epochs is set to 20 with a batch size of 12 and an input/output resolution of 192×640 unless otherwise specified. The initial learning rate is 1×10^{-4} and decays after the 10th epoch with a factor of 0.1. For evaluation, we resize the estimated depth map to the ground-truth depth resolution using bilinear interpolation.

With a 50% chance, we flip the input images horizontally, apply data grafting [14] with the same setting as [14] and add color augmentations where we perform random brightness, contrast, saturation, and hue jitter by sampling uniform distributions in ranges of [0.8,1.2], [0.8,1.2], [0.8,1.2], [0.9,1.1] respectively. The color augmentations are applied to the images that are fed to the Depth-Net rather than those fed to the Flow-Net and the loss function.

[3] https://github.com/princeton-vl/RAFT-Stereo

TABLE I

QUANTITATIVE RESULTS ON THE KITTI DATASET USING EIGEN SPLIT [21]. FOR RED METRICS, LOWER IS BETTER. AND HIGHER IS BETTER FOR BLUE METRICS. PP REPRESENTS POST-PROCESSING [10]. IN DATA COLUMN, D REFERS TO THE METHODS SUPERVISED BY THE GROUND TRUTH DEPTH, M MEANS THAT THE SUPERVISION IS FROM MONOCULAR VIDEO, S MEANS THAT THE SUPERVISION IS FROM STEREO PAIRS, $C^{\dagger}$ MEANS USING PREDICTED SEGMENTATION LABLES AND S^{*} MEANS USING EXTRA SYNTHETIC SCENEFLOW DATASET. THE BEST RESULTS ARE IN **BOLD**.

Method	PP	Data	Resolution	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	δ^1	δ^2	δ^3
DORN [5] (ResNet101)		D	385 × 513 crop	0.099	0.593	3.714	0.161	0.897	0.966	**0.986**
BTS [6] (DenseNet-161)		D	352 × 704 crop	0.091	0.555	4.033	0.174	0.904	0.967	0.984
AdaBins [8] (EfficientNet-B5)		D	352 × 704 crop	0.086	0.475	3.621	0.167	0.918	0.970	0.985
NeWCRFs [25] (swin-ViT)		D	352 × 1120 crop	**0.080**	**0.426**	**3.400**	**0.158**	**0.930**	**0.973**	**0.986**
MonoDepth2 [11]	✓	MS	192 × 640	0.104	0.786	4.687	0.194	0.876	0.958	0.980
DepthHints [13]	✓	MS	192 × 640	0.105	0.769	4.627	0.189	0.875	0.959	0.982
HR-Depth [20]		MS	192 × 640	0.107	0.785	4.612	0.185	0.887	0.962	0.982
CADepth-Net [41] (ResNet50)		MS	192 × 640	**0.102**	**0.752**	**4.502**	**0.181**	**0.894**	**0.964**	**0.983**
Guo et al. [19] w/o Fintuned (VGG-16)	✓	SS^{*}	384 × 1280	0.109	0.822	4.656	0.192	0.868	0.958	0.981
Guo et al. [19] Fintuned (VGG-16)	✓	SS^{*}	384 × 1280	0.099	0.745	4.424	0.182	0.884	0.963	0.983
MonoDepth2 [11]	✓	S	192 × 640	0.108	0.842	4.891	0.207	0.866	0.949	0.976
DepthHints [13]	✓	S	192 × 640	0.106	0.780	4.695	0.193	0.875	0.958	0.980
EPCDepth [14]	✓	S	192 × 640	0.099	0.754	4.490	0.183	0.888	0.963	0.982
Ours	✓	SS^{*}	192 × 640	**0.093**	**0.634**	**4.123**	**0.174**	**0.900**	**0.967**	**0.984**
MonoDepth2 [11]	✓	MS	320 × 1024	0.104	0.775	4.562	0.191	0.878	0.959	0.981
DepthHints [13]	✓	MS	320 × 1024	0.098	0.702	4.398	0.183	0.887	0.963	0.983
HR-Depth [20]		MS	320 × 1024	0.101	0.716	4.395	0.179	0.899	0.966	0.983
Feat-Depth [16] (ResNet50)		MS	320 × 1024	0.099	0.697	4.427	0.184	0.889	0.963	0.982
CADepth-Net [41] (ResNet50)		MS	320 × 1024	**0.096**	**0.694**	**4.264**	**0.173**	**0.908**	**0.968**	**0.984**
MonoDepth2 [11]	✓	S	320 × 1024	0.105	0.822	4.692	0.199	0.874	0.954	0.977
DepthHints [13]	✓	S	320 × 1024	0.099	0.723	4.445	0.187	0.886	0.962	0.981
EdgeDepth [17]	✓	$SC^{\dagger}$	320 × 1024	0.097	0.675	4.350	0.180	0.890	0.965	0.983
EPCDepth [14]	✓	S	320 × 1024	0.093	0.671	4.297	0.178	0.899	0.965	0.983
Ours	✓	SS^{*}	320 × 1024	**0.088**	**0.583**	**3.924**	**0.168**	**0.909**	**0.970**	**0.985**
DepthHints [13] (ResNet50)	✓	S	320 × 1024	0.096	0.710	4.393	0.185	0.890	0.962	0.983
EdgeDepth [17] (ResNet50)	✓	$SC^{\dagger}$	320 × 1024	0.091	0.646	4.244	0.177	0.898	0.966	0.983
EPCDepth [14] (ResNet50)	✓	S	320 × 1024	0.091	0.646	4.207	0.176	0.901	0.966	0.983
Ours (ResNet50)	✓	SS^{*}	320 × 1024	**0.086**	**0.575**	**3.873**	**0.166**	**0.910**	**0.971**	**0.985**

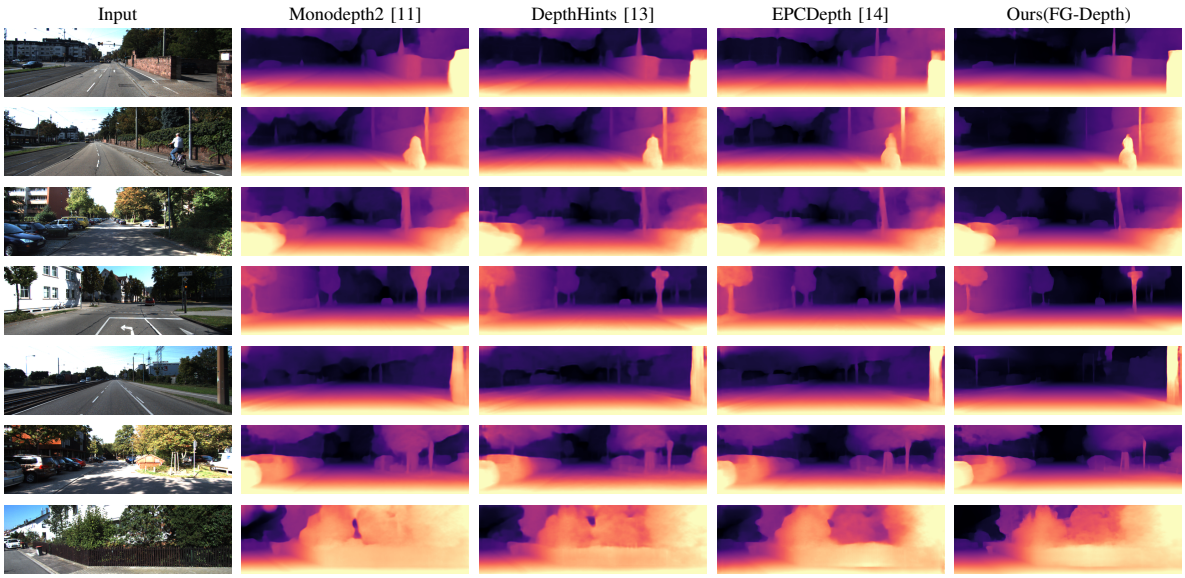

Fig. 5. **Qualitative results on the KITTI dataset using Eigen split.** Our model, FG-Depth, produces the sharpest results even in low-texture regions and on thin structures.

C. Depth Estimation Performance

Firstly, we verify the performance of our model on the KITTI dataset. For a fair comparison, we use the metrics proposed in [21] with Garg's crop [42] and a standard distance cap of 80 meters. The same as other comparative self-supervised methods, we use the same post-processing steps as theirs [10]. The quantitative results are summarized in Tab. I and the qualitative results are shown in Fig. 5.

The quantitative results show that our model, FG-Depth, comprehensively exceeds all existing unsupervised methods that are even trained with stereo video(MS). Compared with [17] which uses extra expensive semantic segmentation labels, our framework uses additional low-cost synthetic optical flow

dataset and gets better performance. Compared with [19] which also distills knowledge from optical flow network pretrianed on sceneflow dataset, FG-Depth performs better event at low resolution. Despite lack of the supervision from ground truth depth maps, the high-resolution performance of FG-Depth is even close to AdaBins [8], a recent state-of-the-art supervised method, and FG-Depth has fewer parameters meanwhile. Besides, the qualitative results show that FG-Depth can produce sharper results even in some low-texture regions and on some thin structures.

Then, we validate the performance on the NYU-Depth-v2 dataset using our model trained on the KITTI just as EPCDepth [17] did. The quantitative results in Tab. II and the qualitative results in Fig. 6 verify the strong generalization ability of our model.

TABLE II

QUANTITATIVE RESULTS ON THE NYU-DEPTH-V2 DATASET.

Method	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	δ^1	δ^2	δ^3
Monodepth2 [11]	0.362	0.718	1.312	0.384	0.477	0.758	0.898
EPCDepth [14]	0.281	0.341	0.912	0.319	0.554	0.833	0.943
Ours(FG-Depth)	**0.269**	**0.318**	**0.888**	**0.312**	**0.560**	**0.840**	**0.947**

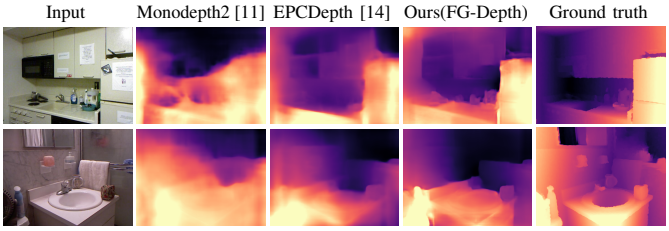

Fig. 6. **Qualitative results on the NYUV2 dataset.**

D. Ablation studies

To better understand the effect of each component of our proposed model, we perform an ablation study and present the results in Tab. III. The results show that all our components can lead to significant performance when combined together.

TABLE III

ABLATION STUDIES. $Base$ REFERS TO THE NETWORK ARCHITECTURE, L_p REFERS TO THE PHOTOMETRIC LOSS, L_{fd} REFERS TO THE FLOW DISTILLATION LOSS, M_p REFERS TO THE AUTO MASK PROPOSED IN [11] AND M_f REFERS TO OUR PRIOR FLOW BASED MASK.

Method	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	δ^1	δ^2	δ^3
$Base$+L_P	0.106	1.300	5.850	0.201	0.872	0.953	0.977
$Base$+L_P+M_p	0.104	0.919	5.176	0.202	0.873	0.953	0.976
$Base$+L_P+M_f	0.100	0.730	4.499	0.195	0.878	0.956	0.979
$Base$+L_{fd}	0.099	1.102	5.230	0.180	0.894	0.965	0.983
$Base$+L_{fd}+M_p	0.097	0.970	5.182	0.180	0.896	0.965	0.983
$Base$+L_{fd}+M_f	**0.093**	**0.634**	**4.123**	**0.174**	**0.900**	**0.967**	**0.984**

a) Flow distillation loss: Although L_p is common in previous self-supervised works, we show that it actually limits the capacity of models. Tab. III shows that in all cases, being trained with L_{fd} can outperform those with L_p.

b) Prior flow based mask: Tab. III also shows that prior flow based mask M_f significantly improves performance and its improvement is more significant than M_p proposed in [11] even though M_f dosen't remove occlusion for L_p while L_{fd} isn't disturbed by occlusion.

c) Loss function combinations: Tab. IV lists performance of different combinations of loss function. The results show that using L_{dr} can already get impressive performance and combining L_{dr} with L_{fp} can get state-of-the-art performance which is consistent with the analysis in Sec. III-C.

d) Pipeline: For a fair comparison with [19], we give results under different pipelines in Tab. V. The results on the first row and on the third row show that our networks have better performance even though at a smaller resolution. And the results in the last row show that our contributions can significantly improve the accuracy and even outperform [19] finetuned with the supervised method.

TABLE IV

ABLATION STUDIES ON LOSS FUNCTION COMBINATIONS. L_p REFERS TO THE PHOTOMETRIC LOSS, L_{dr} REFERS TO THE DEPTH REGRESSION LOSS AND L_{fp} REFERS TO THE FLOW-GUIDED PHOTOMETRIC LOSS.

Loss	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	δ^1	δ^2	δ^3
L_{dr}	0.094	0.643	4.139	0.175	0.896	0.965	**0.985**
L_{fp}	0.098	0.718	4.230	0.177	0.892	0.966	0.984
L_{dr}+L_{fp}	**0.093**	**0.634**	**4.123**	**0.174**	**0.900**	**0.967**	0.984

TABLE V

ABLATION STUDIES ON PIPLINES. $unsupFt$ AND $supFt$ RESPECTIVELY REFERS TO FITUNING THE FLOW-NET USING UNSUPERVISED AND SUPERVISED METHOD ON REAL DATA. $disp$ REFERS TO USING DISPARITY TO SUPERVISE THE DEPTH-NET FOR ALL PIXELS WITH PREDICTION OF FLOW-NET WITHOUT FITUNING AS THE PIPLINE ON THE FIRST ROW DID.

pipline	resolution	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	δ^1	δ^2	δ^3
Guo [19] w/oFt	384 × 1280	0.109	0.822	4.656	0.192	0.868	0.958	0.981
Guo [19] $unsupFt$	384 × 1280	0.099	0.745	4.424	0.182	0.884	0.963	0.983
Guo [19] $supFt$	384 × 1280	0.097	0.653	4.170	0.170	0.889	**0.967**	**0.986**
ours($disp$)	192 × 640	0.103	1.353	5.768	0.185	0.891	0.964	0.982
ours(L_{dr}+L_{fp}+M_f)	192 × 640	**0.093**	**0.634**	**4.123**	**0.174**	**0.900**	**0.967**	0.984

V. CONCLUSION

In this paper, to break the bottleneck of unsupervised monocular depth estimation, noting that optical flow estimation models have strong generalization ability and the typical photometric loss is defective, we propose a flow distillation loss and a prior flow based mask to improve the performance of the unsupervised monocular depth estimator. And the experiments demonstrate that our model, FG-Depth, can achieve state-of-the-art performance on the KITTI dataset and NYU-Depth-v2 dataset. In future work, to further improve the performance, we will explore more methods to make full use of prior optical flow and try to apply our contributions to other categories that use monocular video(M) or stereo video(MS) as input.

REFERENCES

[1] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, vol. 26, no. 2, pp. 120–144, 2009.

[2] J. Michels, A. Saxena, and A. Y. Ng, "High speed obstacle avoidance using monocular vision and reinforcement learning," in *Proceedings of the 22nd international conference on Machine learning*, pp. 593–600, 2005.

[3] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth, "Automatic scene inference for 3d object compositing," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 3, pp. 1–15, 2014.

[4] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[5] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018.

[6] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019.

[7] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using laplacian pyramid-based depth residuals," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4381–4393, 2021.

[8] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *CVPR*, 2020.

[9] D. Kim, W. Ga, P. Ahn, D. Joo, S. Chun, and J. Kim, "Global-local path networks for monocular depth estimation with vertical cutdepth," 2022.

[10] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[11] C. Godard, O. M. Aodha, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019.

[12] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.

[13] J. Watson, M. Firman, G. J.Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *ICCV*, 2019.

[14] R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *ICCV*, 2021.

[15] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *CVPR*, 2018.

[16] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in *ECCV*, 2020.

[17] S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *CVPR*, 2020.

[18] B. Cheng, I. S. Saggu, R. Shah, G. Bansal, and D. Bharadia, "S^3net: Semantic-aware self-supervised depth estimation with monocular videos and synthetic data," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX* (A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds.), vol. 12375 of *Lecture Notes in Computer Science*, pp. 52–69, Springer, 2020.

[19] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[20] X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, and Y. Yuan, "Hr-depth: High resolution self-supervised monocular depth estimation," in *AAAI*, 2021.

[21] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NeurIPS*, 2014.

[22] X. Dan, W. Wei, T. Hao, L. Hong, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," *IEEE*, 2018.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[24] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *ArXiv preprint*, 2021.

[25] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "Newcrfs: Neural window fully-connected crfs for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[26] Z. Yin and J. Shi, "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.

[27] Y. Wang, Z. Yang, P. Wang, Y. Yang, C. Luo, and W. Xu, "Joint unsupervised learning of optical flow and depth by watching stereo videos," *arXiv preprint arXiv:1810.03654*, 2018.

[28] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *European Conference on Computer Vision*, pp. 582–600, Springer, 2020.

[29] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[30] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[32] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] S. Zhao, Y. Sheng, Y. Dong, E. I.-C. Chang, and Y. Xu, "Maskflownet: Asymmetric feature matching with learnable occlusion mask," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] Z. Wang, "Image quality assessment : From error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[35] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *3DV*, 2022.

[36] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[37] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[38] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.

[39] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[41] J. Yan, H. Zhao, P. Bu, and Y. Jin, "Channel-wise attention-based network for self-supervised monocular depth estimation," in *2021 International Conference on 3D Vision (3DV)*, pp. 464–473, IEEE, 2021.

[42] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*, pp. 740–756, Springer, 2016.