Revealing the Reciprocal Relations between Self-Supervised Stereo and Monocular Depth Estimation

Zhi Chen [1], Xiaoqing Ye [2], Wei Yang [1*], Zhenbo Xu [1], Xiao Tan [2],
Zhikang Zou [2], Errui Ding [2], Xinming Zhang [1 *], Liusheng Huang [1]
[1] University of Science and Technology of China
[2] Department of Computer Vision Technology (VIS), Baidu Inc., China
*Corresponding Authors. E-mail: qubit@ustc.edu.cn, xinming@ustc.edu.cn

Abstract

Current self-supervised depth estimation algorithms mainly focus on either stereo or monocular only, neglecting the reciprocal relations between them. In this paper, we propose a simple yet effective framework to improve both stereo and monocular depth estimation by leveraging the underlying complementary knowledge of the two tasks. Our approach consists of three stages. In the first stage, the proposed stereo matching network termed StereoNet is trained on image pairs in a self-supervised manner. Second, we introduce an occlusion-aware distillation (OA Distillation) module, which leverages the predicted depths from StereoNet in non-occluded regions to train our monocular depth estimation network named SingleNet. At last, we design an occlusion-aware fusion module (OA Fusion), which generates more reliable depths by fusing estimated depths from StereoNet and SingleNet given the occlusion map. Furthermore, we also take the fused depths as pseudo labels to supervise StereoNet in turn, which brings StereoNet's performance to a new height. Extensive experiments on KITTI dataset demonstrate the effectiveness of our proposed framework. We achieve new SOTA performance on both stereo and monocular depth estimation tasks.

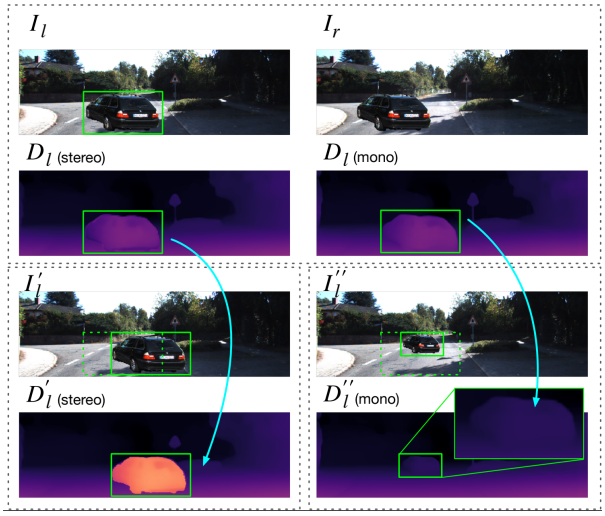

Figure 1. Characteristics of stereo and monocular models. In the upper part, we paste a car 'instance' to both left image I_l and right image I_r, respectively. $D_l(stereo)$ and $D_l(mono)$ are left disparity maps generated from the stereo and monocular models, where the brighter color means bigger disparity. In the lower left, we move the car in left image a distance to the right, and the estimated disparity of the car becomes larger, as shown in $D'_l(stereo)$. In the lower right, we shrink the car, and see the corresponding disparity becomes smaller, as shown in $D''_l(mono)$.

1. Introduction

Depth estimation from either stereo image pairs or monocular images is a fundamental problem in computer vision. It has been extensively studied due to its wide applications in robotic manipulation[37], augmented reality [35, 27] and autonomous driving [26, 43]. Current supervised depth estimation methods [4, 13], though tremendous progress has been achieved, require costly dense ground-truth data for training. Alternatively, self-supervised methods are getting increasing attention in recent years[10, 40, 3], which only requires stereo or monocular raw images.

Recent SOTA self-supervised methods mainly focus on one of monocular or stereo depth estimation problems, neglecting the reciprocal relations between them. On the one hand, stereo matching approaches aim at learning structural information by comparing the similarity of local left and right patches to obtain the optimal disparity and seeking a globally smooth disparity map. Thus for left boundaries and occlusions where only a single view can be seen, unsupervised stereo matching methods often fail to learn reliable depth. On the other hand, monocular depth estimation is an inherently ill-posed problem and it mainly relies on the ap-

pearance or semantic knowledge inside the features. Thus it is robust to occluded regions. As shown in Fig. 1, we conduct the dummy experiments to elaborate the observation.

To fully exploit the complementary knowledge of the two tasks, in this paper, we design a simple but effective framework to integrate the advantages of both stereo and monocular depth estimation networks. Generally, we train a stereo depth estimation network, named StereoNet in a self-supervised manner. Due to the invisible characteristic, estimated depth in occluded regions are not reliable. Thus an occlusion-aware distillation strategy is adopted to extract the visible estimated depths from StereoNet. Different from methods [46, 11] that adopt stereo images for novel view synthesis by a left-right depth consistency term, we propose a monocular depth estimation framework named SingleNet, under the supervision of distilled depth from StereoNet and observe a considerable improvement. The gain can be attributed to two main reasons. First, StereoNet learns more reliable depth given the stereo structural knowledge in visible regions than SingleNet under the same self-supervised training. Second, our occlusion-aware distillation strategy only adopts non-occluded depths as supervision to guide the SingleNet to learn semantic information. Furthermore, not only StereoNet can help to train SingleNet, but also SingleNet can be leveraged to improve StereoNet in turn. Even though StereoNet is generally more accurate than SingleNet, we observe that SingleNet still performs better than StereoNet on occluded pixels. Especially, along the boundary region of objects, SingleNet tends to preserve sharp edge across object borders while bleeding artifacts are obvious for StereoNet. Inspired by this observation, we propose an occlusion-aware fusion strategy, which fuses estimated depth maps from both StereoNet and SingleNet given the occlusion map. The fused depth map gives full play to its strength of the structure-based StereoNet and appearance-based SingleNet. A further hint can be conducted by adopting the fused depth as pseudo-labels for supervision to train StereoNet in turn to further improve the performance of self-supervised StereNet.

In summary, the main contributions of this work are listed below in threefold:

- We propose a simple yet effective framework to boost the performance of self-supervised stereo and monocular depth estimation by mining task-specific strengths and revealing the reciprocal relations of the two tasks.
- We put forward a novel occlusion-aware distillation strategy for training monocular depth estimation networks as well as an effective occlusion-aware fusion strategy that combines the advantages of the structure-based stereo depth estimation and the appearance-based monocular depth estimation.
- Extensive experiments on the KITTI benchmark shows that our method establishes new SOTA performances on both stereo and monocular depth estimation tasks.

2. Related Work

2.1. Stereo Depth Estimation

Stereo matching takes stereo image pairs as input and computes the depth by finding the dense pixel-wise correspondences between left and right images. For stereo depth estimation, supervised approaches [4, 13, 28] have achieved great performance with deep neural networks. GCNet [18] constructs a 3D cost volume by comparing pixel-wise features of reference and target images, then adopts soft-argmin operation to compute the best disparity. PSMNet [4] leverages a pyramid pooling module to encode cost volume, and designs a stacked hourglass 3D CNN to regress the disparity. GWCNet [13] proposes group-wise correlation to construct cost volumes, and modifies 3D hourglass refinement network to improve the performance.

Considering that dense ground-truth depth is challenging to acquire, many works [51, 40] have put great efforts into unsupervised stereo depth estimation, and exhibit considerable performance gain than traditional methods like [14, 15]. Monodepth [51] modifies the convolutional architecture of DispNet [28] to train the network without ground-truth depth as supervision. [10] borrows the architecture of effective GCNet [18] to predict the disparity map with an iterative unsupervised training framework. In UnOS [40], the authors take a lightweight network termed PWCNet [33] for stereo depth estimation by restricting the predicted optical flow to the same horizontal row.

2.2. Monocular Depth Estimation

Monocular depth estimation infers a dense depth map from the appearance feature of a single image. For monocular depth estimation, supervised works [6, 7, 20, 2] have also obtained pleasing results with learning-based methods. [6] adopts a multi-scale convolutional architecture to refine coarse depth prediction. DORN [7] converts the regression problem to quantized ordinal regression problem for higher accuracy. [44] leverages a CRF module to fuse multi-scale depth estimations. BTS [20] replaces bilinear upsampling layer with novel local planar guidance layers at multiple stages in the decoding phase. AdaBins [2] introduces an AdaBins module to divide depth range into bins where the bin widths change per image, and achieves SOTA performance on supervised monocular depth estimation.

Self-supervised approaches [8, 10, 1] learn to estimate the depth map by reducing the photometric loss between stereo image pairs, monocular video frames, or stereo video frames. [8] formulates the photometric loss between stereo pairs with an L2 loss, which results in blurry depth maps. Monodepth [10] takes a combination of SSIM [41] and L1

to measure the similarity between correspondences to improve the depth quality, and a post-processing operation is also applied, where the depth maps of original image and flipped image are averaged to obtain a more accurate depth estimation. Monodepth2 [11] introduces the per-pixel minimum reprojection loss to solve the ambiguity of photometric loss at occluded region. Methods [42, 36] utilize additional proxy labels generated from traditional Semi-Global Matching (SGM) [14, 15] as supervision to train monocular depth estimation models. Recent methods [32, 12] adopt heavier backbones to improve the quality of depth estimation at the cost of time and memory.

2.3. Distillation

Recently, the concept of knowledge distillation has been introduced to transfer the learned knowledge from a teacher model to a student model [21]. The teacher model is usually stronger and heavier, whereas the student model is more lightweight. Knowledge distillation has been successfully exploited for several computer vision tasks such as image classification [39], object detection [5], and natural language processing [17]. In this paper, we borrow the idea of knowledge distillation to transfer the learned structure-based depth knowledge from the stereo model to the monocular model with an occlusion-aware distillation strategy. For further improvement, the fused depth prediction of both stereo and monocular models are also distilled as pseudo labels to train StereoNet in turn. To the best of our knowledge, this work is the first attempt to analyse the reciprocal relations between stereo and monocular depth estimation models.

3. Preliminary

Given a pair of images I_l and I_r, stereo matching networks try to estimate a disparity map, which can be easily converted to depth map as $depth = \frac{b \cdot f}{|disparity|}$, where b is the baseline between left and right cameras and f is the camera focal length. For simplicity, we train both stereo and monocular depth estimation models to predict disparity instead of depth. D_l denotes the disparity from I_l to I_r, and D_r represents the disparity from I_r to I_l.

During the self-supervised training process of stereo matching, the generated disparity map can be applied to synthesize the corresponding view of image [11, 42]. Given learned disparity map D_l, each pixel p_l in the left image I_l is able to find its corresponding pixel $\tilde{p}_r = p_l + D_l(p_l)$ at the right image I_r. If the disparity value $D_l(p_l)$ is accurate and the pixel p_l is not occluded in the right view, the colors of $I_l(p_l)$ and $I_r(\tilde{p}_r)$ should be consistent. Based on this assumption, we are able to reconstruct the appearance of I_l by warping right image I_r according to the obtained $(p_l, \tilde{p}_r)$ pairs as follows,

$$\tilde{I}_{r\to l} = \pi(D_l, I_r), \tag{1}$$

where $\tilde{I}_{r\to l}$ denotes the reconstructed left image originated from the right image, and π is the warp operation using bilinear sampling [16].

Given the warped image $\tilde{I}_{r\to l}$, the photometric loss is employed to calculate the similarity between $\tilde{I}_{r\to l}$ and I_l. Following [48, 10], L1 and SSIM [41] are used to form our photometric loss, and the loss is computed as,

$$L_p = \gamma \frac{(1 - SSIM(I_l, \tilde{I}_{r\to l}))}{2} + (1-\gamma)|I_l - \tilde{I}_{r\to l})|, \tag{2}$$

where SSIM is computed over a 3×3 kernel and γ is set to 0.85 by default.

However, the photometric loss is unfit for texture-less or occluded regions. For pixels in the texture-less region, the photometric loss is ambiguous, thus accurate disparity cannot be guaranteed. For pixels occluded by other objects, there are no corresponding pixels available in the right image. Therefore, the edge-aware smoothness loss [11, 42] is used to alleviate these problems. The smoothness loss is as,

$$L_m = |\partial x D_l| e^{-|\partial x I_l|} + |\partial y D_l| e^{-|\partial y I_l|}, \tag{3}$$

where D_l is first mean-normalized following [38].

4. Method

We first introduce our turbine-like structure pipeline in Section 4.1. Then we elaborate on the proposed self-supervised StereoNet and our distilled monocular depth estimation network in Section 4.2 and Section 4.3, respectively. Finally, the occlusion-aware fusion module and the reutilization of fused prediction are depicted in Section 4.4.

4.1. Overall Framework

As shown in Fig. 2, the whole pipeline of our framework is composed of three main parts. In Stage 1, we design a self-supervised stereo matching network termed StereoNet training from stereo pairs. Since there is no ground truth supervision, the network is inclined to learn the correspondences between left and right patches like the traditional non-CNN approaches do. Thus we term it as structure-based learning, since the network is learned to implicitly carry out similarity comparison between patches. Given the predicted stereo disparity maps, the framework is able to compute the corresponding occlusion map. In Stage 2, rather than directly training a monocular network with photometric loss and smoothness loss as previous works [10, 11] do, we propose an occlusion-aware distillation strategy to leverage the predictions of stereo matching branch as well as the occlusion map to supervise the

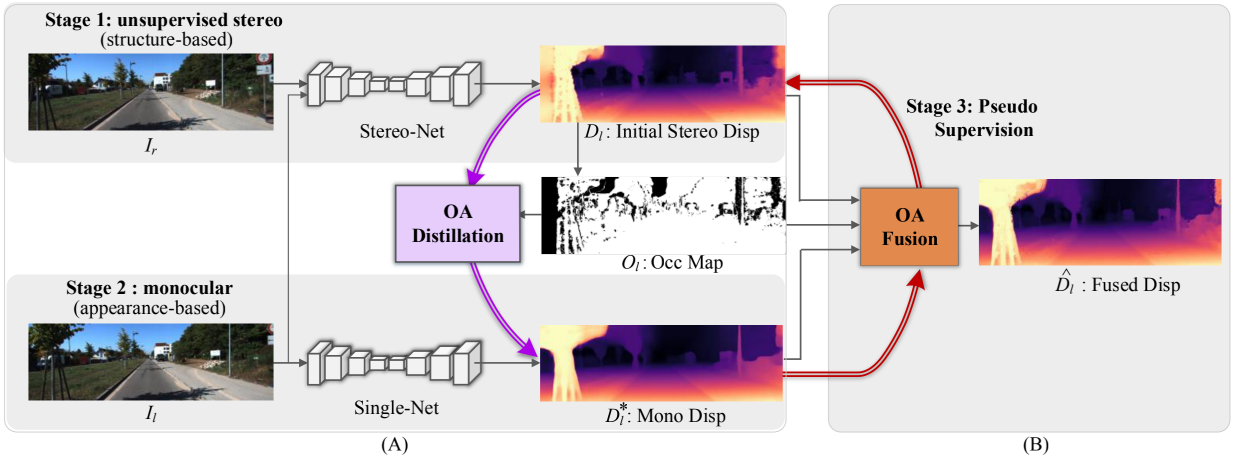

Figure 2. (A) Occlusion-aware distillation (OA Distillation). Disparity map D_l and occlusion map O_l generated from structure-based StereoNet (Stage 1) are used to guide the appearance-based SingleNet (Stage 2). (B) Occlusion-aware fusion (OA Fusion). Given disparity maps D_l and D_l^* predicted from StereoNet and SingleNet, an occlusion-aware fusion module is proposed to generate the fused disparity map $\hat{D}_l$ from D_l and D_l^*. The fused prediction can be further utilized as pseudo labels to supervise StereoNet in turn (Stage 3).

monocular depth estimation network, i.e. SingleNet. Considering that no correspondences can be found in occluded regions, predictions of StereoNet in these regions are unreliable. On the contrary, the monocular branch mainly relies on the appearance knowledge of learned features for depth perception, resulting in more consistent and smoother depth prediction. Inspired by this, the occlusion-aware fusion strategy is put forward to fuse the predictions of stereo and monocular stages given the occlusion map. Furthermore, digging deep into the mechanism of stereo and monocular depth perception, we further promote the performance of stereo matching network by introducing the fused depth map as the pseudo labels to supervise the StereoNet. By revealing the reciprocal relations of the structure-based stereo and appearance-based monocular networks, the performance of both tasks can both be boosted. Note that we do not use any labeled data during training, and in inference stage of SingleNet, only a single image is required.

4.2. Sell-supervised Stereo Branch

In Stage 1 of Fig. 2, we first train a self-supervised stereo disparity estimation model, termed StereoNet. Considering that the previous top-performed stereo matching networks, e.g. GWCNet [13], PSMNet [4], and GANet [47], usually adopt heavy 3D convolutions to trade for accuracy, we instead propose a lightweight unsupervised stereo disparity estimation framework inspired by the optic flow estimation method PWCNet [33]. It is worth mentioning that as a generic framework, multiple stereo matching or optical flow estimation networks can be adopted to instantiate the unsupervised stereo branch in our pipeline. The generalization ability will be further validated in the following experiments.

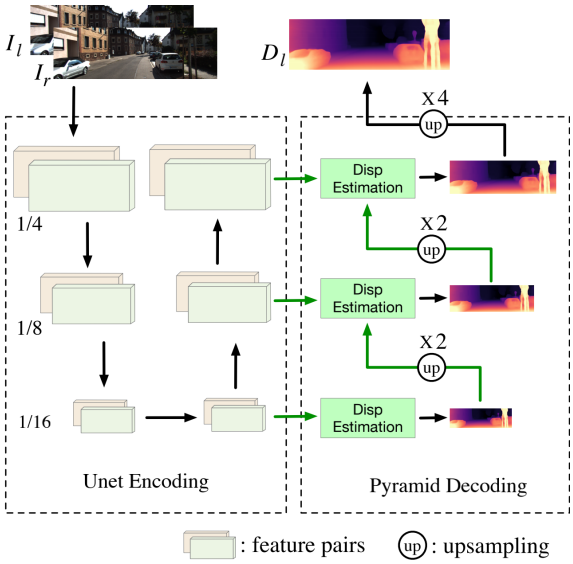

Figure 3. **StereoNet**. The proposed StereoNet is composed of two stages: Unet encoding stage to extract the feature pairs and pyramid decoding to estimate the disparity map. Disp estimation module is used to refine the disparity map with feature pairs at the corresponding layer of the same resolution.

The architecture of our proposed StereoNet is presented in Fig. 3. StereoNet takes stereo images I_l and I_r as input, and outputs the disparity map D_l. The framework of StereoNet is composed of two stages: Unet encoding stage and pyramid decoding stage. In the Unet encoding stage, we adopt a Unet model to extract hierarchical feature pairs for I_l and I_r respectively. In the pyramid decoding stage, the extracted feature map pairs are used to estimate disparity map D_l in a coarse-to-fine manner. More specifi-

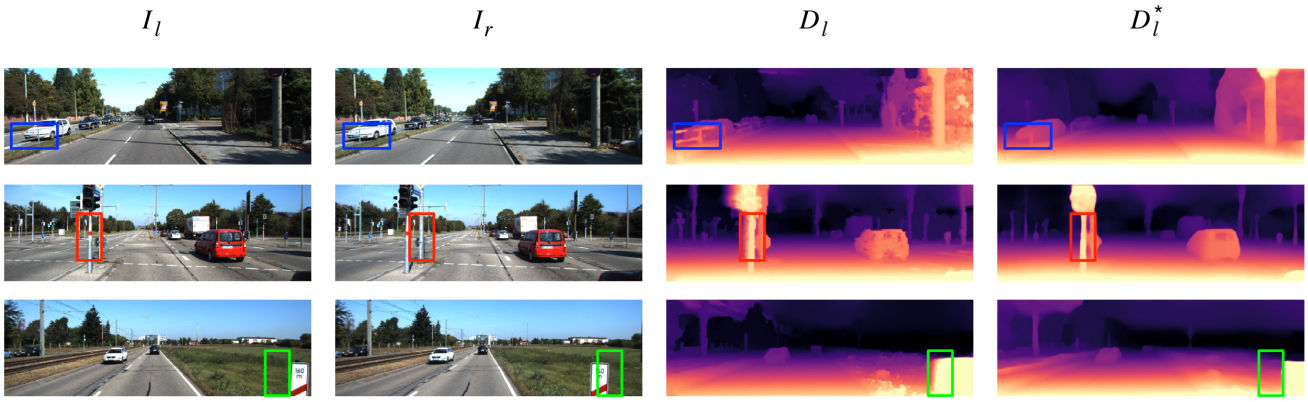

Figure 4. Sample results from StereoNet and SingleNet on KITTI datasets. I_l and I_r are left and right images. D_l and D_l^* are disparity maps estimated from StereoNet and SingleNet, respectively. Green and red squares are regions where SingleNet performs better. In Blue squares, StereoNet gives more detailed and precise disparity estimation results.

cally, in each decoding layer, the disp estimation module generates a refined disparity map based on the coarse disparity map from the preceding layer and the feature map pairs extracted by the Unet encoding module at the layer of the same resolution. The disp estimation module shares a similar architecture as that in [22]. Original PWCNet generates feature pairs with a pyramid encoding structure, where the shallower features are used to predict the higher-resolution disparity map. We believe that a deeper feature is also necessary for higher-resolution disparity estimation. Therefore, we replace pyramid encoding stage with Unet encoding structure, which brings remarkable performance improvement as shown in our experiments.

When training StereoNet, the photometric loss is only computed where pixels in the left image are not occluded, or out of the view in the right image. The occluded pixels are detected by left-right consistency checking [23, 34]. StereoNet is performed twice when computing the occlusion map. We take left and right image as the reference image respectively and compute their disparity map: D_l and D_r. If a pixel p_l is not occluded or out of the view in the right image, the disparity value $D_l(p_l)$ should be the inverse of the disparity value at the corresponding pixel $D_r(\tilde{p}_r) = D_r(p_l + D_l(p_l))$. And hence the occlusion map is detected as follows,

$$O_l = \begin{cases} 1, & |D_l + \tilde{D}_l| \geq \alpha(|D_l| + |\tilde{D}_l|) + 0.5 \\ & \text{or } (p + D_l(p)) \notin \Omega \\ 0, & others \end{cases}, \quad (4)$$

where 0.5 is used to take care of the sub-pixel accuracy for computing the occlusion map, and Ω represents the image boundary. The updated photometric loss is defined as follows,

$$\tilde{L}_p = \frac{\sum L_p \odot (1 - O_l)}{\sum (1 - O_l)}, \quad (5)$$

where $\odot$ stands for pixel-wise multiplication. And the total loss is composed of photometric loss $\tilde{L}_p$ and smoothness loss L_m.

4.3. Distilling Monocular Branch

Monocular depth estimation model predicts the disparity map D_l^* from a single image I_l. Similar to self-supervised stereo depth estimation training, traditional self-supervised monocular models are also trained by minimizing the photometric loss between reference image I_l and warped image $\tilde{I}_l$ [11, 42]. Although stereo and monocular models are both trained from stereo image pairs, the performance of monocular depth estimation is usually inferior to stereo depth estimation. In contrast to monocular methods which directly regress the disparity map from a single image I_l, stereo methods make the use of feature pairs from both images I_l and I_r, and they can typically produce a more accurate disparity map. To make our monocular depth estimation model more robust, we adopt a distillation strategy to train our monocular depth estimation model, called SingleNet.

As mentioned, a better disparity map can usually be generated from the stereo methods. It is hence preferable to explicitly exploit this disparity map to supervise monocular approaches. However, it is known that the disparity map is likely inaccurate in occluded regions by stereo based methods. We hence propose an occlusion-aware distillation strategy to train SingleNet, as shown in Fig. 2(A). Instead of using a common distillation method which takes the whole disparity map generated from stereo images as targets, we utilize only the estimated disparity values in visible regions where pixels pass the left-right consistency check. A log L1 based distillation loss is then utilized to encourage SingleNet to generate similar results as stereo based ap-

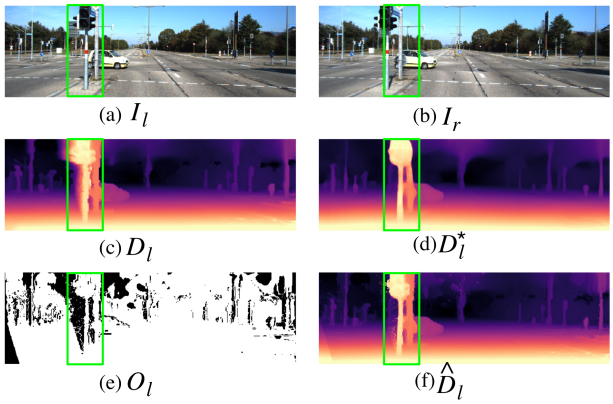

Figure 5. Example intermediate results during occlusion-aware fusion. (a,b) are left and right images, respectively. (c,d) are estimated disparity from StereoNet and SingleNet, respectively. (e) is calculated occlusion map. (f) is fused disparity map.

proaches,

$$L_d = \frac{\sum log(1 + |D_l - D_l^*|) \odot (1 - O_l)}{\sum (1 - O_l)}, \quad (6)$$

where D_l and O_l are disparity and occlusion maps generated from pre-trained StereoNet, and D_l^* is the disparity map estimated from SingleNet. Apart from the distillation loss, the edge-aware smoothness loss is also employed to train the occluded pixels. The total loss for this stage is composed of distillation loss L_d and smoothness loss L_m.

4.4. Distilling Stereo Branch

As discussed above, SingleNet is generally able to yield more preferable results in invisible regions, because SingleNet estimates the depth value (disparity value) of a pixel based on the appearance feature, which is robust compared to stereo approaches based on similarity comparison. This phenomenon is also evident in results from public benchmarks. For example, in the third row of Fig. 4, we can see that D_l gives a wrong estimation in green region of I_l, where the grass in green square is occluded by a road sign. In the second row, there are obvious bleeding artifacts along the boundary region of traffic lights, while SingleNet tends to preserve sharp disparity edge across object borders, as shown in the red square area.

To utilize the advantages of both StereoNet and SingleNet, we further propose an occlusion-aware fusion module, which fuses StereoNet and SingleNet's results to form new disparities, as shown in Fig. 2(B). Specifically, we use D_l and O_l to denote learned disparity and occlusion maps from StereoNet, and D_l^* as the disparity map from SingleNet. The fused disparity map $\hat{D}_l$ is calculated as follows,

$$\hat{D}_l = D_l \odot (1 - O_l) + D_l^* \odot O_l. \quad (7)$$

As shown in Fig. 5, the fused disparity $\hat{D}_l$ is better than both D_l and D_l^*. $\hat{D}_l$ not only preserves the details, but also ensures the sharp disparity edge. On the basis of this observation, we further take the fused disparities as supervision to train StereoNet in turn. The logistic L_1 loss is used as follows,

$$L_{ds} = log(1 + |D_l - \hat{D}_l|). \quad (8)$$

We denote the distilled StereoNet as StereoNet-D to distinguish it from StereoNet trained in Stage 1. And StereoNet-D is even better than the fused disparity used for training itself.

5. Experiments

5.1. Implementation Details

For stereo depth estimation training, we use the whole pipeline to train StereoNet. While for monocular depth estimation, only the first two stages are required for training SingleNet. At the first stage, the α used in O_l formula Eq.6 is set to 0.1. For the rest stages, it is equal to 0.01.

Our models are implemented in PyTorch [29], and trained on one Tesla V100 GPU. Our SingleNet is based on Unet architecture, where Resnet50 is used as our encoder, and the decoder is similar to [11]. For all stages, the weight for smoothness loss is all set to 0.1, and we employ the Adam [19] optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate starts from $1e^{-4}$ and is decayed by a factor of 10 after 15 epochs. We train all models for 20 epochs with a batch size of 8. During evaluation, we restore test images to the full size, and clip the estimated depth to be between 0 and 80 meters. The standard metrics described in [6] are used for comparison.

5.2. Training Datasets

The KITTI dataset [9] is the benchmark widely used for both stereo and monocular depth estimation tasks [12, 32, 42, 40]. KITTI 2015 dataset collects stereo video in 200 street scenes with sparse ground-truth depth obtained from Velodyne laser scanner. The input image resolution is 320×1024. For a fair comparison, different training splits are employed for stereo and monocular depth estimation tasks.

Stereo depth estimation. Following [40], all the raw KITTI images excluding KITTI 2015 training scenes are adopted as the training set, which consists of 29K stereo image pairs. And the 200 training image pairs of KITTI 2015 with ground-truth depth are used as the test split.

Monocular depth estimation. Following [42], we use the data split of Eigen et al. [6], which uses 22600 image pairs for training and 697 images for testing.

5.3. Evaluation

Stereo Depth Estimation. We evaluate our models on the stereo depth estimation task on KITTI 2015 training set,

Methods	Train	Test	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE$_{log}$ ↓	a1↑	a2↑	a3↑
Monodepth [10]	S	S	0.068	0.835	4.392	0.146	0.942	0.978	0.989
SsSMnet [49]	S	S	0.075	1.726	4.857	0.165	0.956	0.976	0.985
OpenWorld [50]	S	S	(0.056)	(0.692)	(3.176)	(0.125)	(0.967)	-	-
UnOS (Stereo-only) [40]	S	S	0.060	0.833	4.187	0.135	0.955	0.981	0.990
UnOS (Ego-motion) [40]	MS	S	0.052	0.593	3.488	0.121	0.964	0.985	0.992
UnOS (Full) [40]	MS	S	0.049	0.515	3.404	0.121	0.965	0.984	0.992
Ours (StereoNet)	S	S	0.052	0.558	3.733	0.123	0.961	0.984	0.992
Ours (Fusion)	S	S	0.049	**0.456**	3.478	0.112	0.964	0.987	0.994
Ours (StereoNet-D)	S	S	**0.048**	0.482	**3.393**	**0.105**	**0.969**	**0.989**	**0.994**
EPC [45]	MS	M	0.109	1.004	6.232	0.203	0.853	0.937	0.975
Ours (SingleNet)	S	M	**0.083**	**0.688**	**4.464**	**0.154**	**0.904**	**0.972**	**0.990**

Table 1. Stereo depth estimation on KITTI 2015 training set. Best results are in bold. In the Train column, S and MS refer to training on stereo pairs and stereo video, respectively. In the Test column, M and S refer to test on stereo or monocular images, respectively. Note that OpenWorld is directly trained and tested on the whole KITTI 2015 training set, thus it is not comparable with other methods.

Methods	Train	PP	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE$_{log}$ ↓	a1↑	a2↑	a3↑
Monodepth2 [11]	M		0.115	0.882	4.701	0.190	0.879	0.961	0.982
PackNet [12]	M		0.107	0.802	4.538	0.186	0.889	0.962	0.981
FeatureNet [32]	M		0.104	0.729	4.481	0.179	0.893	0.965	0.984
HR-Depth [25]	M		0.104	0.727	4.410	0.179	0.894	0.966	0.984
SuperDepth [30]	S		0.112	0.875	4.958	0.207	0.852	0.947	0.977
Monodepth2 [11]	S		0.107	0.849	4.764	0.201	0.874	0.953	0.977
Refine&Distill [31]	S		0.098	0.831	4.656	0.202	0.882	0.948	0.973
Ours (SingleNet)	S		0.095	0.697	4.435	0.186	0.891	**0.962**	**0.981**
Ours (SingleNet)	S	✓	**0.094**	**0.681**	**4.392**	**0.185**	**0.892**	**0.962**	**0.981**
MonoResMatch [36]	Ssgm	✓	0.111	0.867	4.714	0.199	0.864	0.954	0.979
DepthHints [42]	Ssgm	✓	0.096	0.710	4.393	**0.185**	0.890	**0.962**	**0.981**
EPC++ [24]	MS		0.128	0.935	5.011	0.209	0.831	0.945	0.979
Monodepth2 [11]	MS		0.106	0.806	4.630	0.193	0.876	0.958	0.980
FeatureNet [32]	MS		0.099	**0.697**	4.427	0.184	0.889	0.963	0.982
HR-Depth [25]	MS		0.101	0.716	**4.395**	0.179	**0.899**	**0.966**	**0.983**
DepthHints [42]	MSsgm	✓	**0.098**	0.702	4.398	0.183	0.887	0.963	**0.983**

Table 2. Monocular depth estimation on KITTI Eigen split. Best results are in bold. In the Train column, M, S and MS refer to training on monocular video, stereo pairs and stereo video, respectively, and sgm means additional SGM proxy labels used as supervision. PP refers to post-processing introduced by [10].

and the quantitative results are presented in Tab. 1. Our models all show great performance on KITTI 2015 training set. StereoNet is our baseline model, and is only trained in the self-supervised setting with the photometric loss and smoothness loss. From Tab. 1, we can see that StereoNet outperforms all other models trained on stereo in all metrics, and even achieves comparable performance with methods trained on stereo video. Especially, StereoNet is better than the SOTA UnOS (Stereo-only) on metrics Abs Rel (0.052 vs. 0.060) and Sq Rel (0.558 vs. 0.833), which shows the effectiveness of our proposed Unet encoding module in StereoNet.

We also give the results of our SingleNet. The performance of SingleNet trained on stereo pairs is much better than EPC [45], which is trained on stereo video. Although SingleNet's performance is inferior to some stereo depth estimation models, it can also be utilized to improve StereoNet's performance thanks to the occlusion-aware fusion module that combines the advantages of both StereoNet and SingleNet. From Tab. 1, we can see that the fusion strategy improves StereoNet from 0.052 to 0.049 on Abs Rel.

Moreover, we also distill the fused disparity to StereoNet, and the results of StereoNet-D are further improved again. Besides, the performance of StereoNet-D even surpasses SOTA UnOS (Full) trained on stereo video.

Monocular Depth Estimation. We evaluate our SingleNet on the monocular depth estimation task on KITTI Eigen split. For monocular depth estimation, we only per-

form the first two stages on Eigen training split to get a fair comparison with other monocular depth estimation methods. Tab. 2 presents all the SOTA performance trained on monocular video, stereo and stereo video, respectively. For methods [12, 11, 32, 25] trained on monocular video, the per-image median ground truth scaling [10] is used during evaluation. Our SingleNet gets the top performance among all the methods, especially in the Sq Rel metric. Moreover, SingleNet performs even better than models trained on stereo video. MonoResMatch and DepthHints leverage generated depth maps from the classical SGM [14, 15], and their performances are still lower than our SingleNet. We also test the post-processing technique introduced by [10], which further improves SingleNet's quantitative performance.

5.4. Ablation Study

Here, we conduct more experiments to show the contribution of our proposed network modules.

Unet Encoding Module. Traditional PWCNet[33] uses a pyramid encoding stage to generate image feature pairs, while our StereoNet adopts a Unet encoding stage. Tab. 3 shows the results of our StereoNet under different encoding stages. These results are evaluated after the first training stage. As shown in Tab. 3, Unet encoding can significantly improve the performance.

Encoding	Abs Rel	Sq Rel	RMSE	RMSE log	a1	a2	a3
Pyramid	0.054	0.629	30.822	0.127	0.959	0.983	0.991
Unet	**0.052**	**0.558**	**30.733**	**0.123**	**0.961**	**0.984**	**0.992**

Table 3. Contribution of Unet encoding in StereoNet.

Occlusion-Aware Distillation. To show the effectiveness of our proposed occlusion-aware distillation strategy, we also present the results under self-supervised training. For self-supervised training, we only perform the first stage by replacing StereoNet with SingleNet. Considering some algorithms [42, 31] using SGM as supervision, we also conduct experiments with SGM-based occlusion-aware distillation strategy. These results are all shown in Tab. 4, and trained on Eigen split. As can be seen, our method achieves the best performance.

We also present the impact of different αs used in the occlusion-aware distillation strategy. Different from usual distillation, we only take estimated disparity values in non-occluded region to supervise SingleNet. When computing occlusion map, α is used to control which pixels are seen as occluded. The results of SingleNet under different αs are presented in Tab. 5. All the results are trained on stereo split. We can see that the best results are obtained when $\alpha = 0.01$.

Train	Abs Rel	Sq Rel	RMSE	RMSE log	a1	a2	a3
self.	0.102	0.817	4.678	0.196	0.881	0.957	0.979
sgm.	0.100	0.834	4.576	**0.186**	0.889	**0.962**	**0.981**
stereo.	**0.095**	**0.697**	**4.435**	**0.186**	**0.891**	**0.962**	**0.981**

Table 4. Comparison of different supervision types for SingleNet. Self. means self-supervised training. Stereo. and sgm. represent distillation from stereo and SGM, respectively.

α	Abs Rel	Sq Rel	RMSE	RMSE log	a1	a2	a3
1	0.086	0.724	4.577	0.161	0.898	0.967	0.986
0.1	0.086	0.708	4.509	0.159	0.898	0.968	0.987
0.01	**0.083**	**0.688**	**4.464**	**0.154**	**0.904**	**0.972**	**0.990**

Table 5. Comparison of different αs in occlusion-aware distillation on SingleNet. These results are trained on stereo split.

Occlusion-Aware Fusion. We also evaluate the occlusion-aware fusion strategy under different occlusion maps controlled by α. Experiments under different αs are all performed on stereo split, and are presented in Tab. 6. We can see that best results are also obtained when α is equal to 0.01.

α	Abs Rel	Sq Rel	RMSE	RMSE log	a1	a2	a3
1	0.052	0.559	3.737	0.123	0.961	0.984	0.992
0.1	0.050	0.484	3.531	0.115	**0.964**	**0.987**	**0.994**
0.01	**0.049**	**0.456**	**3.478**	**0.112**	**0.964**	**0.987**	**0.994**

Table 6. Comparison of different α in occlusion-aware fusion.

6. Conclusion

In this paper, we proposed a simple yet effective framework to improve both stereo and monocular models in an unsupervised collaborative fashion. The introduced occlusion-aware distillation module leverages the predicted depths from stereo pairs by StereoNet to improve our monocular depth estimation network, named SingleNet. We also designed an occlusion-aware fusion module, which fused estimated depths from StereoNet and SingleNet on the basis with calculated occlusion map. And the fused depths were then taken as pseudo labels to supervise StereoNet in turn, which brought further performance improvement. SOTA performances on both stereo and monocular tasks are obtained on the KITTI benchmark.

Acknowledgement

This work was supported by the Anhui Initiative in Quantum Information Technologies (No. AHY150300).

References

[1] Juan Luis Gonzalez Bello and Min-Soeng Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *arXiv: Computer Vision and Pattern Recognition*, 2020.

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. *arXiv preprint arXiv:2011.14141*, 2020.

[3] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and I. Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *ArXiv*, abs/1908.10553, 2019.

[4] Jia-Ren Chang and Y. Chen. Pyramid stereo matching network. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 742–751, 2017.

[6] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[8] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[12] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020.

[13] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3268–3277, 2019.

[14] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005.

[15] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

[17] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

[18] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[21] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017.

[22] L. Liu, Jiangning Zhang, Ruifei He, Y. Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6488–6497, 2020.

[23] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8770–8777, 2019.

[24] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2624–2641, 2019.

[25] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. *arXiv preprint arXiv:2012.07356*, 2020.

[26] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2069–2078, 2019.

[27] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey.

IEEE transactions on visualization and computer graphics, 22(12):2633–2651, 2015.

[28] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas 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*, pages 4040–4048, 2016.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[30] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9250–9256. IEEE, 2019.

[31] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019.

[32] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020.

[33] Deqing Sun, X. Yang, Ming-Yu Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[34] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010.

[35] Fulin Tang, Yihong Wu, Xiaohui Hou, and Haibin Ling. 3d mapping and 6d pose computation for real time augmented reality on cylindrical objects. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2887–2899, 2019.

[36] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.

[37] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018.

[38] Chaoyang Wang, J. M. Buenaposada, Rui Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.

[39] Chong Wang, Xipeng Lan, and Yangang Zhang. Model distillation with knowledge transfer from face classification to alignment and verification. *arXiv preprint arXiv:1709.02929*, 2017.

[40] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8071–8081, 2019.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[42] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2162–2171, 2019.

[43] Di Wu, Zhaoyong Zhuang, Canqun Xiang, Wenbin Zou, and Xia Li. 6d-vnet: End-to-end 6-dof vehicle pose estimation from monocular rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[44] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5354–5362, 2017.

[45] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[46] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018.

[47] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019.

[48] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

[49] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017.

[50] Yiran Zhong, Hongdong Li, and Yuchao Dai. Open-world stereo video matching with deep rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–116, 2018.

[51] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1567–1575, 2017.