UNSUPERVISED MONOCULAR DEPTH ESTIMATION WITH CLEAR BOUNDARIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised monocular depth estimation has made great progress after deep learning is involved. Training with binocular stereo images is considered as a good option as the data can be easily obtained. However, the depth or disparity prediction results show poor performance for the object boundaries. The main reason is related to the handling of occlusion areas during the training. In this paper, we propose a novel method to overcome this issue. Exploiting disparity maps property, we generate an occlusion mask to block the back-propagation of the occlusion areas during image warping. We also design new networks with flipped stereo images to induce the networks to learn occluded boundaries. It shows that our method achieves clearer boundaries and better evaluation results on KITTI driving dataset and Virtual KITTI dataset.

1 INTRODUCTION

Monocular depth estimation becomes an active research topic as deep learning is applied in various computer vision tasks. It has many applications, from navigation through to scene understanding. A single traditional camera can be a cheaper alternative to the expensive LIDAR sensor for automotive cars if accurate estimation can be achieved. Meanwhile, single camera simplifies the design of depth estimation solution which can be adopted quite widely at a low cost.

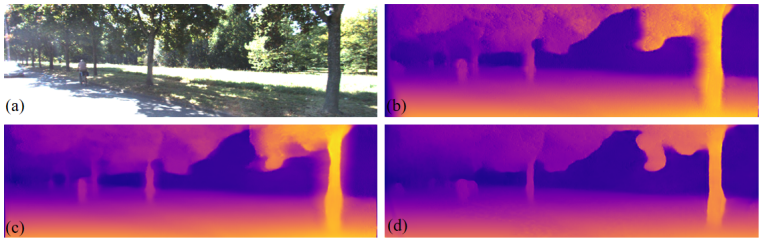

Figure 1: Blurred boundaries of predicted disparity by Godard et al. (2017) and our result. (a) input image (b) predicted left disparity (c) post-processed disparity(d) our result. To eliminate the artifacts and blurred patterns, Godard et al weighted sum the disparity of the input image and flipped disparity of the flipped input image at the cost of doubling the computation. However, there still exists blurred patterns in the post-processed disparity. Contrary to Godard et al. as shown in (b)(c), our result(d) shows clear boundaries with less artifacts and higher accuracy

One straight-forward way to train deep depth estimation models is to use ground truth depth images as the supervision signals Eigen et al. (2014). However, supervised deep learning method is eager for massive data with ground truth. Collecting large datasets with ground truth depth in varied real scenarios is challenge and expensive. Instead, training using stereo images without depth label is an alternative option. Godard et al. (2017) proposed a method to exploit the left-right consistency of stereo images to tackle the monocular depth estimation, which achieved quite promising results. However, the depth predicted by their method has blurred boundaries. The issue is mainly due to the occlusions during the image warping. Though it can be alleviated in some extent with proper post processing, the fundamental problem is not well addressed.

In this paper, we propose a new method to overcome the blurred boundaries when using stereo pairs to train the monocular depth model. An example is illustrated in Fig. 1. During the image warping, we generate an occlusion mask using the disparity map to block the inappropriate back-propagation gradients for occlusion areas. However, the mask only cannot guarantee clear boundaries as there is no constrain for the masked areas. Then we design new networks to fully exploit the information of stereo images. With flipped stereo pairs, the network is induced to learn clear boundaries for occlusion areas. Our method provides a solution to the fundamental learning difficulty of occluded areas introduced by image warping in depth estimation. Empirical evaluation on KITTI driving dataset (Geiger et al. (2012)) and Virtual KITTI dataset (Gaidon et al. (2016)) demonstrates the effectiveness of our approach. Moreover, we find the depth label of KITTI 2015 is usually very sparse near the object boundaries, which is not very sensitive to evaluate the clearness of boundaries.

2 RELATED WORK

Large amounts of multi-view based approaches have been proposed such as stereo matching(Scharstein & Szeliski (2002)), difference view point(Furukawa et al. (2015)) or temporal sequence(Ranftl et al. (2016)). Here we briefly review work based on single view depth estimation which is usually harder since reasoning depth from monocular colored image is an ill-posed problem.

2.1 SUPERVISED MONOCULAR DEPTH ESTIMATION

The most intuition way is treating monocular depth estimation problem as a supervised problem by taking RGB images as inputs and Lidar depth points as ground truth. Saxena et al. (2009) proposed a method known as Make3d which breaks the image into homogeneous patches. For each small homogeneous patch, Saxena et al used a Markov Random Field (MRF) to infer a set of plane parameters that capture both the 3D location and 3D orientation. However, this approach has a hard time capturing thin structures since the predictions are made locally. Eigen et al. (2014) first exploited CNN in a coarse to fine manner. Liu et al. (2016)Liu et al. (2015) proposed a network jointly explore the capacity of deep CNN and continuous conditional random field (CRF). Ladicky et al. (2014) incorporated semantic segmentations in to single-view depth estimation task since they are closely tied to the property of perspective geometry. Laina et al. (2016) proposed a residual network with up-sampling module using fully convolutional architecture. Also, reverse Huber loss was introduced. However, large amount of high-quality labelled data is needed, which is hard to require in practice.

2.2 SEMI-SUPERVISED/UNSUPERVISED BASED METHODS

To overcome the lack of high quality labelled data, several semi-supervised and fully unsupervised methods have been proposed. Flynn et al. (2016) first proposed a view synthesis based method called DeepStereo which generates new view image from nearby image. Xie et al. (2016) (Deep3D) proposed a method which generates the right image through probability distribution over all the possible disparities for each pixel. Garg et al. (2016) first proposed a warp based method by aligning reconstructed image with the ground truth left image as described in 3.1. However, their loss is not fully differentiable. Godard et al. (2017) improve this methods by introducing a novel loss. Repala & Dubey (2018) extended the network into two separate channel with 6 or 12 losses which improves the result.

Based on Garg et al, Kuznietsov et al. (2017) proposed a semi-supervised methods that exploited both the sparse Lidar points as supervision and stereo pairs as unsupervion signals. The semi-supervised method was further improved by Luo et al. (2018), they decoupled the monocular depth prediction problem into two procedure, a view synthesis procedure followed by stereo matching.

Recently, several work using only monocular temporal sequence comes out which enables more training data such as video sequence on YouTube. Zhou et al. (2017) proposed a network that predicts depth and camera pose separately. Using the predicted depth and camera pose, relative temporal image can be reconstructed by image warping with which final loss can be constructed. Mahjourian et al. (2018) performed a novel 3D loss that enforces consistency of the estimated 3D point clouds and ego-motion across consecutive frames and combined it with 2D photometric loss.

Wang et al. (2018) proposed a differentiable implementation of Direct Visual Odometry (DVO) and a novel depth normalization strategy. However, all the temporal sequence based training meet the same problem of object motion. This problem can be alleviated by including stereo pairs during training known as trinocular training(Zhan et al. (2018)Poggi et al. (2018)Godard et al. (2018)). However, all these warp based methods have the difficulty of learning occluded area which would infect the result.

3 METHODS

Godard et al achieves state of art result of unsupervised monocular depth estimation with only stereo images. Follow Godard et al, we proposed monocular depth estimation network with novel mask methods which can be trained end-to-end and without ground-truth label. Our method is superior to Godard et al in result quality with clearer boundaries, especially on dense evaluation dataset such as virtual-KITTI(Gaidon et al. (2016)).

3.1 SELF-SUPERVISED DEPTH ESTIMATION FROM STEREO PAIRS

In general, our goal is to learn a network that can predict a pixel-wise dense depth map from single colored image($d^l = d(I^l)$). However, all supervised methods have a hard time with acquiring large dense labelled data in real scenarios. Thus, several unsupervised methods have been proposed to overcome the obstacle. Among these methods, training by image reconstruction using rectified stereo pairs became more and more popular currently due to its high accuracy and easy accessibility of training data. (Garg et al. (2016)Godard et al. (2017)Repala & Dubey (2018))

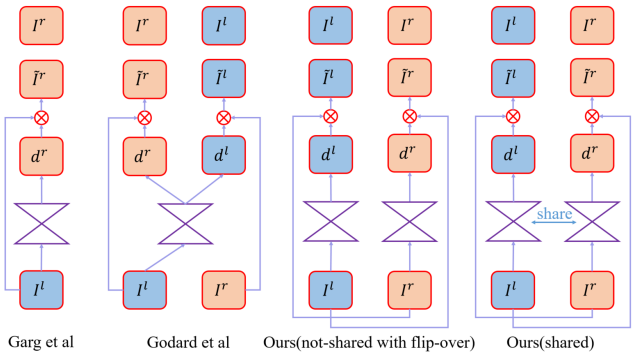

Figure 2: Network architecture. From left to Right: Naive unsupervised network proposed by Garg et al. (2016), Left-right consistency network proposed by Godard et al, Our refined network without shared parameters and our refined network with shared parameters

First proposed by Garg et al. (2016), the monodepth estimation network takes left image (I^l) as input and outputs the disparity aligned with the right image (d^r) which can be used to sample from the left image (I^l) to reconstruct the right image ($\tilde{I}^r$) during training. Thus, image reconstruction loss can be constructed between the reconstructed right image($\tilde{I}^r$) and original right image(I^r). When testing, only one colored image are required, the predicted disparity can be converted to depth simply using $depth = b * f/disparity$, where b and f are given baseline and camera focal length respectively. It is worth to mention that disparities (d^l and d^r) are a scalar per pixel as the images are rectified.

The network was further improved by Godard et al. (2017) by introducing left-right consistency loss and refined encoder-decoder network. Given the input left image, the network predicts both the left and right disparities simultaneously which enables constructing the left-right consistency. This consistency restriction leads to more accurate result and less artifacts. Also, fully differentiable backward bilinear sampling was used to reconstruct the left image which makes the model easier to optimize. With better network architecture and better loss, Godard et al achieved the state of art result of unsupervised monodepth estimation only with rectified stereo pairs, and even outperforming supervised methods. Their network architectures are shown in Fig. 2.

However, there still are unsatisfactory artifacts at the occlusion boundaries showing blurred ramps on the left side of image and of the occluders. Even with the post-processing step which weighted sums the flipped disparity of the flipped input image and the disparity of the input image, the blurred ramps are still visible near the objects, especially for those near the cameras with higher disparities as illustrated in Fig. 1.

3.2 WARPING MASK

Though common backward warping using bilinear sampler is fully differentiable which enables us to train the model end to end, some undesirable duplicates and artifacts are introduced during the warping process because of occlusions according to Lu et al. (2018). In Fig. 3, we use SYTHIA dataset (Ros et al. (2016)), a synthesized virtual driving dataset to illustrate. Though ground truth disparities are used for warping, there still exists obvious duplicates and even a huge black region on the left of the reconstructed image($\tilde{I}^l$) because of the occlusion. If those inevitable artifacts and duplicates are back propagated during the training process, unwanted high losses will be introduced forcing the network learn to blur in those regions, as the blurriness(disparity ramps) in the occluded regions will make the reconstructed image show stretched patterns(Fig 7) which are more similar to original ground truth image compared to duplicates and large black regions(Fig. 3).

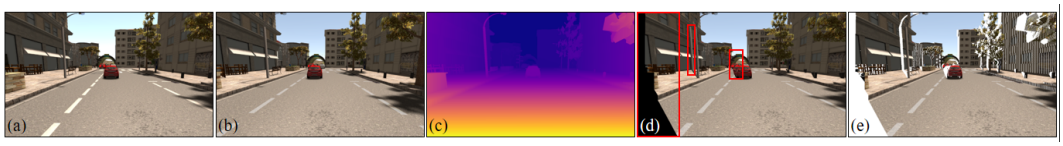

Figure 3: Problem of warping and corresponding mask. (a)left image I^l (b)right image I^r (c)disparity aligned with left image d^l (d)reconstructed left image $\tilde{I}^l$ (e)reconstructed left image after being masked. We used ground truth disparity of left image d^l(c) to warping the right image I^r(b) by bilinear sampling. However, duplicates and artifacts are obvious in reconstructed left image ($\tilde{I}^l$)(d) compared to the ground truth left image (a). We used a warping mask which is generated automatically from disparities to block the back-propagation of those artifacts. The final output is shown in (e), where the white regions are masked.

In order to block the back propagation process of those warping induced artifacts, we designed an algorithm which can mask the occlusion region automatically. In backward bilinear sampling process, a disparity map was used to sample from the source image. The intuition here is that, if any pixel in the source image has never been sampled during the bilinear sampling process, this pixel should only be visible in the source image and should be masked when reconstructing this source image later on. Thus, the mask methods takes, say, left disparity as input while generating the mask of reconstructed right image and vice versa. The pseudo code is shown in Algorithm 1 and the mask schematic diagram is shown in Fig 4.

However, the mask method alone cannot guarantee the clearness since no constrain is added in masked region. Specifically, though the masks block the back propagation of duplicates and artifacts induced by warping, they also block the further learning process of those regions. Once the disparities start to blur, hardly can we correct the network back to clearness. To solve this problem, we refined the network architecture and introduced a flip-over training scheme.

3.3 SOLVING THE PROBLEM OF BLURRINESS

Though the mask blocks the process of further learning of the blurred regions, we find that the blurred side are definite which can be exploited to reactive the learning process. For example, when the network is only trained on left images and takes corresponding right images as ground truth, the disparity ramps (blurred regions) will only appear on the left side of the occluders. So, if we randomly flipped the input images horizontally, the disparity ramps will still appear on the left side. When flipped the output disparities back for warping, the blurred regions will appear on the right side where no mask is added. Examples are shown in the last column of Fig 7. Thus, those blurred regions will make distortions when reconstructing images, and back propagate despite the

Algorithm 1: mask algorithm for occluded regions

Input: disparity of left image d^l,disparity of right image d^r,shape of input image(h,w)
Output: mask of reconstructed image $mask^r$ and $mask^r$ (Boolean array with the same height and width with the images, where Trues are masked region)

$mask^r \leftarrow Trues(h, w)$;
$mask^l \leftarrow Trues(h, w)$;
for *x,y in grid(h,w)* **do**
  if $0 \leq x - d^l(x) \leq w$ **then**
    $mask_1^r(floor(x - d^l(x)), y) = False$;
    $mask_2^r(cell(x - d^l(x)), y) = False$;
    $mask^r = mask_1^r \vee mask_2^r$
  end
  if $0 \leq x + d^r(x) \leq w$ **then**
    $mask_1^l(floor(x + d^r(x)), y) = False$;
    $mask_2^l(cell(x + d^r(x)), y) = False$;
    $mask^l = mask_1^l \vee mask_2^l$
  end
end
return $mask^l, mask^r$

mask. Also, because the flip-over scheme is performed randomly, any blurriness on the definite side will receive punishment on average which restrict the prediction to be clear. It is worth to mention that flip-over scheme will not affect masking process and we still use masks to block the back propagation of duplicates. The flip-over schematic diagram is shown in Fig. 4.

However, the predicted disparities of the right branch won't make any sense if we take the flipped left image as input and are totally mismatched with the ground truths as is shown in Fig 4. As a result, we delete the predicting branch of the right disparity and add another encoder-decoder network which takes right images as input and predicts right disparities as is shown in Fig 2. This doubled encoder-decoder network enables us to preform left-right consistency loss at the cost of doubling the training parameters. However, it won't slow down the test speed since only one branch of encoder-decoder network is used when testing. We also tried another network architecture with shared the encoder-decoder network which achieves comparable result as non-shared network while halves the training time. More details can be found in 6.1

3.4 Training Loss

We use similar training loss as Godard et al. (2017). The losses are performed on four different scale and are finally summed as the total loss. $C = \sum_{s=1}^{4} C_s$. For each scale, three different losses including appearance match loss(C_{ap}), disparity smoothness loss(C_{ds}) and LR consistency loss(C_{lr}) are performs as follows.

$$C_s = \alpha_{ap}(C_{ap}^l + C_{ap}^r) + \alpha_{ds}(C_{ds}^l + C_{ds}^r) + \alpha_{lr}(C_{lr}^l + C_{lr}^r), \tag{1}$$

Intuitively, appearance matching loss measures photometric error between the reconstructed image($\tilde{I}_{ij}^l$)) and the ground truth image(I_{ij}^l), which is defined as the weighted sum of L1 and SSIM shown as follows,

$$C_{ap}^l = \frac{1}{N}\sum_{i,j} \alpha \frac{1 - \text{SSIM}(I_{ij}^l, \tilde{I}_{ij}^l)}{2} + (1-\alpha)\left\| I_{ij}^l - \tilde{I}_{ij}^l \right\|. \tag{2}$$

Disparity smoothness loss aims to force the smoothness of predicted disparities through

$$C_{ds}^l = \frac{1}{N}\sum_{i,j} \left|\partial_x d_{ij}^l\right| e^{-\left\|\partial_x I_{ij}^l\right\|} + \left|\partial_y d_{ij}^l\right| e^{-\left\|\partial_y I_{ij}^l\right\|}. \tag{3}$$

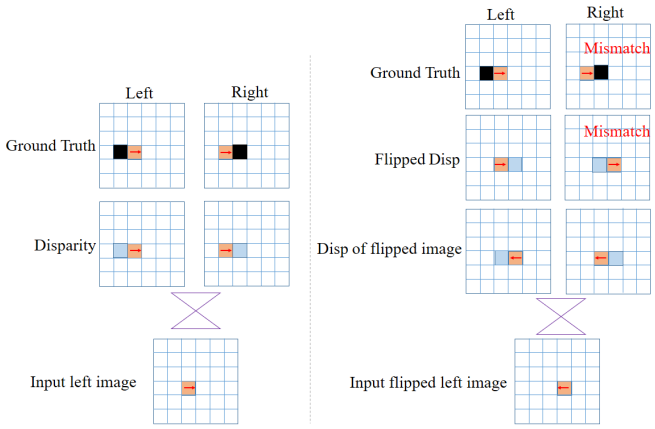

Figure 4: Schematic diagram. The orange squares refer to an object such as pedestrians or vehicles. The blue and black squares refer to blurred disparity region and masked region respectively. **Left**: Naive masking process. Warping mask(black) is overlap with blurred region(blue). As a result, masks will block the learning of those region. **Right**: Flip-over training scheme. We flipped the left images($f(I^l)$) as input and flipped the output disparities($f(d(f(I^l)))$) back to reconstruct the image($\tilde{I}^l$). Different from the diagram shown in left, the blurred region will switch to the opposite side to avoid the mask. As a result, losses will be introduced leading to clear boundaries. We can also observed that the right branch (flipped predicted right disparity of flipped left image($f(d(I^r))$) is totally mismatch with the ground truth right image, so we delete this branch and add another encoder-decoder branch as shown in Fig. 2.

Finally, L1 penalty is added to constrict the left-right consistency, which further reduce the artifacts,

$$C_{lr}^{l} = \frac{1}{N}\sum_{i,j}\left|d_{ij}^{l} - d_{ij+d_{ij}^{l}}^{r}\right|. \tag{4}$$

3.5 Implement details

Our network is based on Godard et al. (2017) and is implemented in Tensorflow (Abadi et al.). The network contains encoder network based on VGG16 or Resnet50 and decoder network with 7 up-convolutional layers, 7 merging layers and 4 disparity prediction layers. With 7 skip connections, the network can better handles features at different scales. Different from Godard et al, we modified the channel number of disparity prediction layers to predict only one disparity instead of two.

Also, we tuned the default hyper parameters α and α_{ds}. The rest are the same with Godard et al with $\alpha_{ap} = 1$, $\alpha_{lr} = 1$. The learning rate λ remains 10^{-4} for the first 30 epoch and halves every 10 epoch when trained for 50 epoch. We also tried batch norm but might lead to unstable result. Data augmentation is performed on the fly similar as Godard et al including flipping and color shifting.

4 Experiment

We trained our model on rectified stereo image pairs in KITTI and Cityscapes dataset and evaluated mainly on KITTI split(Geiger et al. (2012)), Eigen split(Eigen et al. (2014)) and virtual-KITTI datasets(Gaidon et al. (2016)). Also, we find the problem of KITTI 2015 evaluation such as sparsity and man-made defects which infects our evaluation result. When testing on dense datasets like virtual-KITTI, the superiority becomes more obvious.

4.1 KITTI split

For comparison, we use the same split as Godard et al. (2017) which 29000 out of 42382 rectified stereo image pairs with size of 1242*375 are used for training. The evaluation is performed on 200

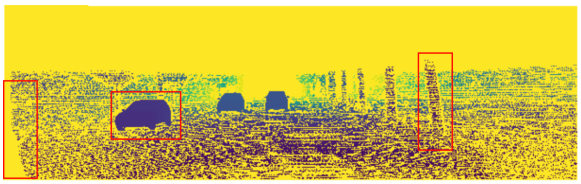

Figure 5: The sparsity problem of KITTI 2015 evaluation split. Vacancy are visible on the left of the image and near the boundaries of cars and trees, which would have an impact on the evaluation result.

high quality disparity image issued by official KITTI dataset. It is worth mentioned that though these disparity images are of better quality than projected velodyne laser point and with CAD models inserted in cars, most part of ground truth are extremely sparse, especially in terms of occluded regions. As the red boxes in the Fig. 5 shown, there is some blank space on the left side of image and occluded regions which is not involved in the evaluation. Thus, those blank space just covers the shortcomings the disparity ramps on the occluded regions. As a result, our result shows less superiority over Godard et al on KITTI stereo 2015 dataset.

As shown in Table 1, we use the metrics from Eigen et al. (2014) and *D1-all* metrics from KITTI. Our model achieves comparable result with Godard et al when both trained on KITTI dataset only for 50 epoch, and superior result when both trained longer for 100 epoch.

Method	encoder	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	*D1-all*	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard 50 epoch	vgg	K	0.124	1.388	6.125	0.217	30.272	0.841	0.936	0.975
Godard 100 epoch	vgg	K	0.124	1.378	6.146	0.216	31.522	0.837	0.936	0.975
Godard resnet pp	resnet	CS + K	0.097	0.896	5.093	0.176	23.811	0.879	0.962	0.986
Ours non-shared 50epoch	vgg	K	0.124	1.415	6.126	0.215	30.988	0.842	0.937	0.975
Ours non-shared 100epoch	vgg	K	0.121	1.228	5.977	0.209	31.10	0.846	0.941	0.977
Ours non-shared resnet pp	resnet	CS + K	**0.097**	0.960	**5.030**	**0.170**	**22.585**	**0.884**	**0.963**	0.985

Lower is better

Higher is better

Table 1: Evaluation result on KITTI 2015 stereo dataset(Geiger et al. (2012)). K means the model is only trained on KITTI dataset, while CS + K means the model is trained on Cityscapes dataset and then finetune on KITTI dataset. Also, pp means post processing step, and more details about post processing can be found in 6.3. For a fair comparison, we train the network proposed by Godard et al and our non-shared network both to 50 and 100 epoch, and find that our network has larger improvement and better performance than Godard et al. (2017) when trained longer to 100 epoch. We guess that our network requires longer time to converge. We choose different hyperparameters from Godard et al for our network. Though it is unfair to evaluation on sparse KITTI dataset, our result still outperforms that of Godard et al.

4.2 EIGEN SPLIT

Similar to Godard et al, we use the test split of 697 images proposed by Eigen et al. (2014). Each image contains 3D points captured by velodyne laser which was used to generate ground truth depth. We keep the same 22600 stereo pairs for training. According to Godard et al, all evaluation results are done under the Garg crop(Garg et al. (2016)) except for Eigen results for a fair comparison. We also present the uncropped result which our models' superiority become more obvious, because it will crop the disparity ramps on the left side for evaluation which boosting Godard et al's result. Also, ground truths are captured by velodyne Lidar thus rather sparse which would reduce our superiority over Godard et al. The evaluation result is in Table 2

4.3 VIRTUAL-KITTI

To prevent sparsity induced inaccurate result, we evaluate models on the Virtual-KITTI dataset(Geiger et al. (2012)). Virtual-KITTI dataset contains 50 monocular videos generated from five different virtual worlds in urban settings under different weather conditions with corresponding pixel-level ground truth depth. With the same resolution, scenes and camera parameters as KITTI 2015, Virtual-KITTI dataset can be implement naively when testing. Since KITTI 2015 dataset we used for training does not cover those weathers, only 2126 labbeled images without weather condi-

Method	Supervised	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. (2014) coarse	Yes	K	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen et al. (2014) fine	Yes	K	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Godard	No	K	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard resnet pp	No	CS + K	0.114	0.898	4.935	0.206	0.861	0.949	0.976
Ours	No	K	0.146	1.134	5.887	0.246	0.805	0.926	0.965
Ours non-shared resnet pp	No	CS + K	**0.113**	**0.890**	**4.894**	**0.203**	**0.866**	**0.952**	**0.978**
Garg et al L12 Aug $8\times cap50m$	No	K	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard cap 50m	No	K	0.140	0.976	4.471	0.232	0.818	0.931	0.969
Godard resnet pp cap 50m	No	CS + K	0.108	0.657	3.729	0.194	0.873	0.954	0.979
Ours non-shared cap 50m	No	K	0.1389	0.9796	4.477	0.231	0.820	0.935	0.970
Ours non-shared resnet pp cap 50m	No	CS + K	**0.107**	**0.648**	**3.703**	**0.192**	**0.878**	**0.957**	**0.980**
Godard uncropped	No	K	0.215	4.137	7.718	0.315	0.772	0.899	0.947
Godard resnet pp uncropped	No	CS + K	0.130	1.197	5.222	0.226	0.843	0.940	0.971
Ours non-shared uncropped	No	K	0.165	1.625	6.157	0.264	0.784	0.913	0.958
Ours non-shared resnet pp uncropped	No	CS + K	**0.124**	**1.071**	**5.097**	**0.216**	**0.851**	**0.945**	**0.974**

Lower is better

Higher is better

Table 2: Result on Eigen split Eigen et al. (2014). K is KITTI and CS is cityscapes dataset for training. We use the evaluation result provided in Godard et al. (2017). For a fair comparison, we use the crop the same as Garg et al. (2016) except for Eigen et al. (2014) and apply the same hyper-parameters as Godard et al on our model. Besides, we set the maximun evaluation depth to 50 meters(cap 50m) in the second row which is the same as Garg et al. (2016) , while others remain 80 meters. Also, we compared uncropped result with Godard et al, on which the superiority become more obvious.

tion are used for evaluation. The evaluation result is shown in Table 3. For visual result, please see our supplementary 6.2.

Method	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard	K	0.295	6.824	11.543	0.427	0.638	0.800	0.884
Godard pp	K	0.222	3.121	9.993	0.374	0.659	0.811	0.900
Godard resnet pp	CS + K	0.191	2.923	9.269	0.313	0.700	0.871	0.942
Ours non-shared	K	0.229	3.501	10.286	0.369	0.655	0.825	0.908
Ours non-shared pp	K	0.218	3.099	10.086	0.361	0.661	0.827	0.905
Ours non-shared resnet pp	CS + K	**0.176**	**2.216**	**8.721**	**0.283**	**0.738**	**0.893**	**0.952**
Godard crop	K	0.2165	3.1826	10.126	0.369	0.665	0.822	0.903
Godard pp crop	K	0.2093	2.9135	9.820	0.354	0.673	0.930	0.910
Godard 70 epoch crop	K	0.2149	3.2421	9.765	0.363	0.675	0.939	0.910
Godard resnet pp crop	CS + K	0.1812	2.4501	9.132	0.301	0.707	0.871	0.942
Ours non-shared crop	K	0.2123	3.129	10.050	0.353	0.678	0.836	0.916
Ours non-shared pp crop	K	0.2043	2.8846	9.839	0.341	0.678	0.841	0.921
Ours non-shared 70 epoch crop	K	0.2074	2.907	9.933	0.351	0.673	0.836	0.915
Ours non-shared resnet pp crop	CS + K	**0.1649**	**1.982**	**8.570**	**0.271**	**0.749**	**0.900**	**0.957**
Godard crop&cap 50m	K	0.198	2.277	7.119	0.336	0.707	0.820	0.919
Godard pp crop&cap 50m	K	0.193	2.132	6.294	0.321	0.714	0.858	0.927
Godard 70 epoch crop&cap 50m	K	0.199	2.348	6.992	0.334	0.713	0.861	0.924
Godard resnet pp crop&cap 50m	CS + K	0.162	1.584	6.006	0.262	0.753	0.908	0.961
Ours non-shared crop&cap 50m	K	0.193	2.165	6.945	0.315	0.715	0.867	0.934
Ours non-shared pp crop&cap 50m	K	0.186	1.981	6.768	0.304	0.721	0.871	0.939
Ours non-shared 70 epoch crop&cap 50m	K	0.188	1.991	6.792	0.314	0.718	0.869	0.932
Ours non-shared resnet pp crop&cap 50m	CS + K	**0.147**	**1.250**	**5.526**	**0.233**	**0.795**	**0.929**	**0.959**

Lower is better

Higher is better

Table 3: Evaluation result on virtual-KITTI. Once our ground truth become dense, our model out performance other models for its sharp and clear boundaries on predicted depth map. Even we crop the black edges of Godard et al (the left 10% of the whole disparity map) in the second row, our model is still superior to Godard et al. (2017) (the state of art unsupervised monocular depth estimation using left and right information only). We evaluate on 3 different methods: naive, crop the black edge, both crop the edge and set the maximun evaluation depth to 50 meters

5 CONCLUSION

In this work, we present an occlusion mask and filp-over training scheme to enable effective learning of object boundaries when using image warping. With our new network, our model achieves state of art result using only stereo images. Moreover, as warping based image reconstruction is commonly used in depth estimation problem, our method provides a solution to the fundamental difficulty of occluded areas introduced by image warping.

In the future, our method can be incorporated with more accurate network trained on trinocular data (temporal Stereo sequence) such as Zhan et al. (2018), Poggi et al. (2018) and Godard et al. (2018), which would further boost the accuracy.

REFERENCES

M Abadi, A Agarwal, P Barham, E Brevdo, Z Chen, C Citro, GS Corrado, A Davis, J Dean, M Devin, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arxiv preprint (2016). *arXiv preprint arXiv:1603.04467*.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pp. 2366–2374, 2014.

John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5515–5524, 2016.

Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

A Gaidon, Q Wang, Y Cabon, and E Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016.

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*, pp. 740–756. Springer, 2016.

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

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

Clément Godard, Oisin Mac Aodha, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *arXiv preprint arXiv:1806.01260*, 2018.

Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6647–6655, 2017.

Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 89–96, 2014.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pp. 239–248. IEEE, 2016.

Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5162–5170, 2015.

Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10): 2024–2039, 2016.

Yao Lu, Jack Valmadre, Heng Wang, Juho Kannala, Mehrtash Harandi, and Philip HS Torr. Devon: Deformable volume network for learning optical flow. *arXiv preprint arXiv:1802.07351*, 2018.

Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 155–163, 2018.

Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5667–5675, 2018.

Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. *arXiv preprint arXiv:1808.01606*, 2018.

Rene Ranftl, Vibhav Vineet, Qifeng Chen, and Vladlen Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4058–4066, 2016.

Vamshi Krishna Repala and Shiv Ram Dubey. Dual cnn models for unsupervised monocular depth estimation. *arXiv preprint arXiv:1804.06324*, 2018.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009.

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2022–2030, 2018.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pp. 842–857. Springer, 2016.

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*, pp. 340–349, 2018.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, pp. 7, 2017.

6 SUPPLEMENTARY MATERIALS

6.1 SHARED WEIGHT NETWORK

The shared weight network is similar to the non-shared weight network except for the shared weight. The architecture is shown in 2. Fortunately, the shared weight model naturally handled the problem. Because the blurriness always occur in the occluded region, say, disparity ramps (burriness) in left disparity maps will only appear on the left side of the occluders, the network cannot distinguish which side the input image belongs to and which side to blur when trained on the left and right image simultaneously with shared weights. Thus, any wrong blurriness (blurriness on the wrong side, say, right blurriness in the left disparity maps) will lead to punishment in raising reconstructed loss since no mask was adding on the other side (wrong side, say, right side). As a result, the network achieves comparable result as non-shared network while halves the training time.However, shared method is slightly inferior to non-shared network but still out performance Godard et al. The test result is shown in Fig 6.

6.2 VIRTUAL-KITTI TEST RESULT

Addition comparison results on virtual-kitti dataset are shown in Fig. 5

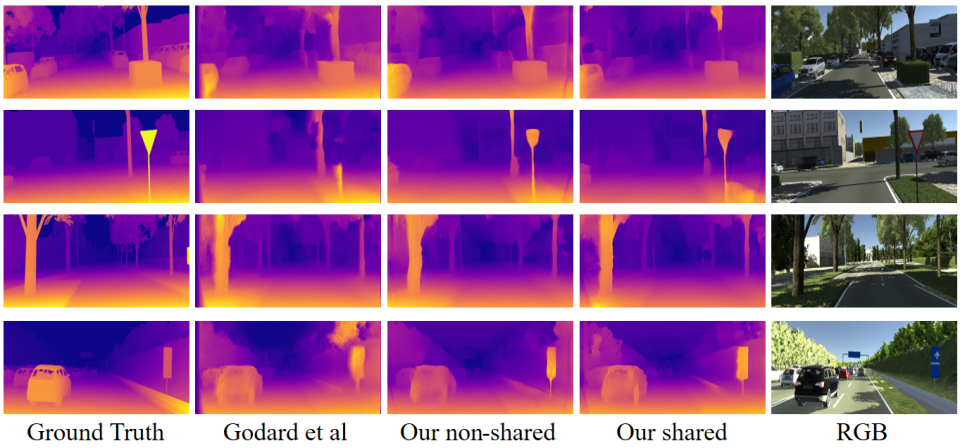

Figure 6: virtual-kitti test results

6.3 POST-PROCESSING

Godard et al weighted summed the flipped disparity of flipped image and original disparity as post processing step. To eliminate the black edge on the left side of image as shown in Fig 1(b), Godard et al used non-linear weight. Because our model do not predict black edge, so we simply average them as our post-processing.

6.4 EXAMPLE OF BLURRED RESULTS AND FLIP-OVER SCHEME

Here we use the network by Godard et al. (2017) to illustrate how flip-over scheme could solve the problem of back-propagate, so the predicted disparities are rather blurred. In Fig 7, from left to right: left disparity(d^l), reconstructed left image($\tilde{I}^l$), corresponding mask($mask^l$), ground truth(I^l), flipped disparity of flipped input image$f(d(f(I^l)))$. To minimize the reconstruction error, the network predicts blurred result forcing the network to sample from the background, which leads to more similar stretched patterns instead of high loss duplicates as shown in Fig 3(d). However, the blurred regions are aligned with the mask, which freezes the fine-tuning process of blurred regions. To solve the problem, we randomly flip the colored image as input($Input = f(I^l)$), then flipped the output back as final disparities for warping($f(d(f(I^l)))$). As shown in the last column, the blurred

regions are no longer aligned with the masks which enables further learning process. As the training process goes on, the boundaries will become clearer as shown in Fig. 1. We also use schematic diagram to illustrate in Fig. 4

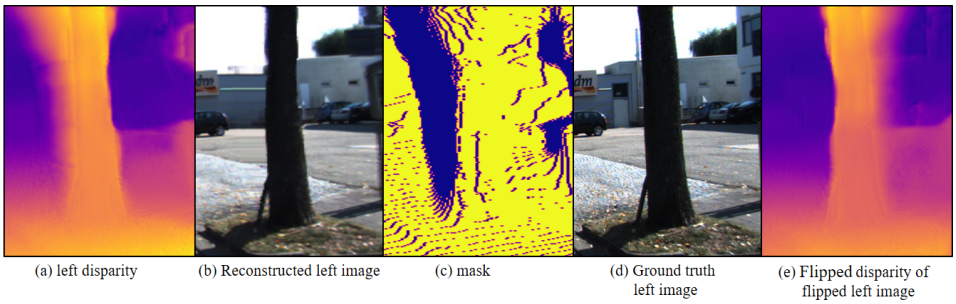

Figure 7: Example of blurred results and flip-over scheme.