SINGLE IMAGE DEPTH ESTIMATION BASED ON SPECTRAL CONSISTENCY AND PREDICTED VIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Single image depth estimation is a critical issue for robot vision, augmented reality, and many other applications when an image sequence is not available. Self-supervised single image depth estimation models target at predicting accurate disparity map just from one single image without ground truth supervision or stereo image pair during real applications. Compared with direct single image depth estimation, single image stereo algorithm can generate the depth from different camera perspectives. In this paper, we propose a novel architecture to infer accurate disparity by leveraging both spectral-consistency based learning model and view-prediction based stereo reconstruction algorithm. Direct spectral-consistency based method can avoid false positive matching in smooth regions. Single image stereo can preserve more distinct boundaries from another camera perspective. By learning confidence map and designing a fusion strategy, the two disparities from two approaches are able to be effectively fused to produce the refined disparity. Extensive experiments indicate that our method exploits both advantages of spectral consistency and view prediction, especially in constraining boundaries and correcting wrong predicting regions.

1 INTRODUCTION

Depth prediction and scene reconstruction are required in robot navigation, augmented reality, autonomous driving, and many other scenarios. Depth maps can be obtained directly from active sensors such as LiDAR or Kinect. However, due to their high price and limited valid range, stereo reconstruction is utilized as a major approach. Stereo matching algorithms, such as Semi-Global Matching (SGM) or Semi-Global Block Matching (SGBM), infer the disparity from two or more images. However, stereo matching usually results in wrong correspondences on regions with low-texture and repeated pattern. Also, binocular and baseline setup are required.

Recent works show outstanding performance of deep neural networks in depth prediction from largely collected training datasets based on supervised frameworks (Eigen et al. (2014) Mayer & et al. (2016)) or unsupervised schemes (Garg et al. (2016) Godard et al. (2017) Poggi & et al. (2018) Godard & et al. (2019)) from a single image without any priors. However, direct single image depth estimation usually generates confusing and blurred output in object boundaries (with and without occlusions) due to missing constraints from a different perspective like stereo or SfM reconstruction.

In this paper, we improve upon these methods above with a novel combined training objective and enhanced network architecture that significantly increases the quality of our final results. We propose two separate depth estimation algorithms, spectral consistency based single image depth estimation and single image stereo matching based on predicted views. To benefit from both the depth estimation approaches, an efficient fusion strategy is proposed with the help of a confidence map to fuse spectral-consistency based single image depth estimation and predicted-view based stereo algorithm in a self-supervised approach. First, we learn an end-to-end deep neural network to predict disparity map from a single image by applying spectral consistency constraints. Then we exploit it to predict a view from the given image and apply the single image stereo algorithm between the original view and predicted view images to extract more information and produce clear boundary, which can be constrained from a different camera perspective. Following the aforementioned depth prediction networks, a fusion network is introduced to select the highly confident depth between the two depth maps and correct low-confidence regions. The disparity maps developed from the two newly designed single image depth estimation algorithms along with the help of the fusion strategy

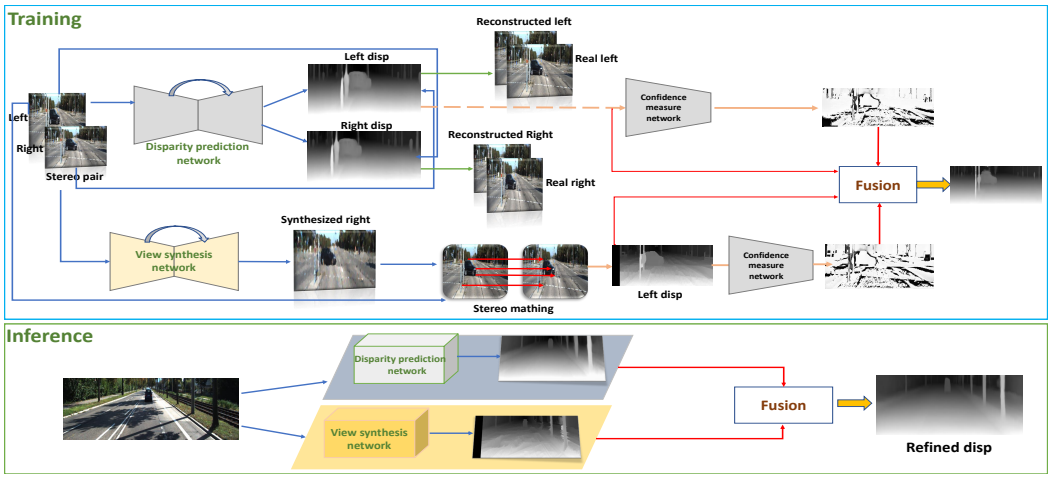

Figure 1: **Overview of the proposed pipeline.** For training, both spectral consistency based disparity estimation and view prediction take stereo image pairs as input. In inference, a single image is used to estimate the depth. Then the fusion module takes both of the disparities to produce a refined output which prevent blurred object boundary and wrong-matching regions.

lead to a highly accurate depth estimation. The entire pipeline is shown in Fig. 1, which composes both training (using stereo images) and testing stages (using just a single image).

The key contributions are as follows: 1) We improve a self-supervised deep neural network to predict disparity from single image based on the spectral consistency for image reconstruction; 2) We propose a single image stereo reconstruction approach based on neighboring-view prediction; 3) A novel strategy is developed to fuse the spectral consistency constrained estimation and single image stereo reconstruction to achieve a highly confident depth.

2 Related Work

Stereo matching algorithm: Most conventional stereo matching algorithms follow matching cost computation, cost aggregation, disparity computation and refinement as the basic steps between left and right image pairs. Semi-Global Matching (SGM) or Semi-Global Block Matching (SGBM) based methods leverage both local and global features, and perform fast approximation from all directions. In contrast to the hand-crafted matching cost metric, CNNs are explored to match between patches. Bontar & et al. (2016) investigated a series of CNN architectures for binary classification of pairwise matching and applied it for disparity estimation. Mayer & et al. (2016) proposed to train an end-to-end network DispNet on a large synthetic dataset to infer disparity as well as optical flow. As main contribution to this work, 1-D correlation along the disparity line is applied to approximate the cost volume. Following this work, GCNet by Kendall et al. (2017) proposed to deploy 3-D convolutions on 4-D volume to mimic the matching costs and obtain the best disparity over the volume. PSMNet designed by Chang & et al. (2018) applies a multi-scale pyramid matching to explore global context information as well as local cues to improve disparity.

Depth prediction from single image: Early approaches for monocular depth estimation utilize hand-craft features to build statistic model for depth prediction Saxena et al. (2007) Bipin & et al. (2015). Most recent frameworks focus on deep neural networks for predicting depth from single image (Eigen et al. (2014) Liu et al. (2015) Fu & et al. (2018) Lee & et al. (2019)). Eigen et al. (2014) generated depth maps by deploying networks capable of detecting global and textured features using the AlexNet structure Krizhevsky & et al. (2012). Following this work Liu et al. (2015) applied the continuity of the depth values and treated depth estimation as a continuous conditional random field (CRF) learning problem. Cao & et al. (2017) took this concept further by formulating depth estimation as a pixel-wise classification task, using conditional random field (CRF) as a post-processing scheme. DORN presented by Fu & et al. (2018) proposed a regression based network for monocular depth estimation. Lee & et al. (2019) later adopted the same reconstruction strategy and ordinal loss function, extending it for estimating relative depths at various scales. Targeting at solving the supervision from ground truth, Garg et al. (2016) proposed differentiable inverse warping to learn matching from stereo image pairs . Godard et al. (2017) built on top of this work and introduced novel left-and-right pixel disparity consistency loss to improve the performance. Instead

of exploiting geometrical cues from stereo pairs, (Zhou et al. (2017) Yin & Shi (2018) Mahjourian et al. (2018) Zou & et al. (2018) Ranjan & et al. (2019)) achieve success in attempts to explore monocular depth estimation methods by combining ego-motion from unlabeled video sequences.

Disparity enhancement from fusion: Multiple approaches attempt to refine disparities from different cues such as diverse information (i.e. temporal and spatial) Zhan & et al. (2018) Mun & et al. (2015), different tasks (i.e. depth and semantic segmentation) Ramirez & et al. (2018) to improve accuracy of the initial map. Mun & et al. (2015) applied motion prediction to the traditional iterative stereo matching method to compute moving region, Zhan & et al. (2018) utilized stereo video sequences to extract both spatial (left-right pairs) and temporal (forward-backward) information to further improve the performance and scale. Ramirez & et al. (2018) and Ochs et al. (2019) both proposed to train a CNN architecture to optimize depth prediction by jointly learning semantic segmentation.

3 Unsupervised Simultaneous Single Image Depth Estimation

In this section, we first describe the overall structure of how we build the entire pipeline by combing the single image depth estimation based on spectral consistency and from view prediction. Second, we introduce each component of the proposed method, including the intuition and designed unsupervised loss functions. Our framework involves three main components, single image depth estimation based on spectral consistency cues, single image stereo reconstruction based on the predicted view, and depth map fusion.

3.1 Spectral Consistency Based Depth Estimation

Our spectral consistency based depth estimation scheme transfers single image depth estimation task as an image reconstruction issue. For each image (I^l or I^r) from a stereo pair, our model targets at constructing a deep neural network to predict its corresponding disparity map ($\widetilde{d^l}$ or $\widetilde{d^r}$). Then the predicted disparity map (e.g., left image disparity $\widetilde{d^l}$) is used to reconstruct the right image $\widetilde{I^r}$ by warping with the biniliear interpolation operation. L1 loss is combined with Structural Similarity Index Metric (SSIM) term (Godard et al. (2017)) to constrain the reconstructed image to be spectrally consistent with the original image and build the photometric error function. The loss is defined as:

$$L_{me} = 0.85 * \frac{1}{N} \sum_{xy} \frac{(1 - SSIM(I_{xy}, \tilde{I}_{xy}))}{2} + 0.15 * (||I_{xy} - \tilde{I}_{xy}||_1) \tag{1}$$

where I_{xy} refers to the image pixel at x th row and y th column in the original input images, and $\tilde{I_{xy}}$ represents the reconstructed image pixel. To enforce the disparity maps to be smooth and avoid substantial gradient change on flat regions. Edge-aware smoothness loss term is applied.

$$L_{sm} = \frac{1}{N} \sum_{xy} (|\partial x d_{xy}| e^{-||\partial x I_{xy}||_1} + |\partial y d_{xy}| e^{-||\partial y I_{xy}||_1}) \tag{2}$$

where ∂x and ∂y are disparity gradient operators in horizontal and vertical direction respectively. d_{xy} is the predicted disparity value in the corresponding x-y coordinate.

In order to maintain the coherence between the predicted left and right disparity, we extend the left-to-right consistency constraint Godard et al. (2017) with the reverse Huber (berHu) penalty term Zwald & Lambert-Lacroix (2012). The improved disparity consistency loss now becomes:

$$L_{cs} = \begin{cases} |d_{xy}^{l(r)} - d_{xy+d_{xy}^{l(r)}}^{r(l)}| & |d_{xy}^{l(r)} - d_{xy+d_{xy}^{l(r)}}^{r(l)}| \leq c, \\ \frac{(d_{xy}^{l(r)} - d_{xy+d_{xy}^{l(r)}}^{r(l)})^2 + c^2}{2c} & |d_{xy}^{l(r)} - d_{xy+d_{xy}^{l(r)}}^{r(l)}| > c. \end{cases} \tag{3}$$

where constant $c = \frac{1}{5} max(|d_{xy}^{l(r)} - d_{xy+d_{xy}^{l(r)}}^{r(l)}|)$. The berHu loss in L1 norm is in the range of $[-c, c]$ and L2 norm is out of this range, thus empirically demonstrating a good balance between thees two.

To solve the border artifact in Garg et al. (2016) and Godard et al. (2017) as a result of zero padding in the border regions, we apply the closest pixel to replace those regions that are out of the boundary. The accumulative constraints from monocular depth prediction based on spectral consistency cue L_{mono} is comprised by L_{me}, L_{sm} and L_{cs}, and the corresponding weights are 1.0, 0.1 and 0.8 respectively from our experiments.

3.2 SINGLE IMAGE STEREO BY PREDICTED VIEW

Spectral consistency based single image depth estimation can generate satisfatory depth output for smooth regions. However, due to the little constraints from a different perspective and smoothness, sharp object boundaries are usually not preserved. To deal with this issue, we also propose a single image stereo algorithm that can enhance the depth estimation accuracy of spectral consistency single image depth estimation approach. To realize the single image stereo reconstruction method, a binocular view image is predicted from a GAN-based view synthesis method, which is trained from stereo pairs. Once the view from another perspective is predicted, stereo matching algorithm SGBM will be employed to estimate the scene depth. As binocular views capture the same scene from different perspectives, stereo image prediction problem can be simplified as an image generation task from different domain representations. In this problem, the source domain represents images from left camera view and target domain samples are the images from right view. Then our objective is to learn the transformation and inverse transformation between these two domains. Assuming the input modality is S, and target representation is T, we aim to learn both conversions of $h_{S->T}$ and $h_{T->S}$. Extending the basic idea from Cycle-GAN network (Zhu & et al. (2017)), we apply multi-scale generators and discriminators to extract both local and holistic features from S and T. The mapping function from input views S to target views T is expressed as:

$$L_{GAN}(G_{s->t}, D_t, S, T) = E_{s\in p(s)}[log(1 - D_t(G_{s->t}))] + E_{t\in p(t)}[log(D_t)] \tag{4}$$

where $G_{s->t}$ represents the generator to create images from the source domain to be similar as the images in the target domain, and and D_t is discriminator to identify the real images and generated images from $G_{s->t}$. By combing the image transformation from source to target views and target views back to source views, the total adversarial losses L_G equals to $L_{GAN}(G_{s->t}, D_t, S, T) + L_{GAN}(G_{t->s}, D_s, T, S)$.

To decrease the instability and uncertainty of the mapping function between source views and target views, a cycle-consistency loss (Yi & et al. (2017) Kim & et al. (2017)) is utilized here to build the forward cycle consistency and backward consistency, i.e. $S \overset{G_{s->t}}{\rightarrow} \widetilde{T} \overset{G_{t->s}}{\rightarrow} \widetilde{S} \simeq S$ and $T \overset{G_{t->s}}{\rightarrow} \widetilde{S} \overset{G_{s->t}}{\rightarrow} \widetilde{T} \simeq T$. Thus the loss can be given as:

$$L_{cyc} = E_{s\in p(s)}[\|G_{t->s}(G_{s->t}(s)) - s\|_1] + E_{t\in p(t)}[\|G_{s->t}(G_{t->s}(t)) - t\|_1] \tag{5}$$

where E is the expectation of loss values of all the training samples. s and t separately represent the left-view modality and right-view modality. And L1 norm is applied to compute the distance between source images and translated source images $G_{t->s}(G_{s->t}(s))$ and correspondingly original target domain images and translated target images $G_{s->t}(G_{t->s}(t))$. Though cycle consistency loss was originally introduced for unpaired image data, we found it also has outstanding performance on the paired datasets. The full optimization for synthesis network turns into $L_{stereo} = L_G + \lambda L_{cyc}$ where λ is set to 10 and progressively decreased after half of the training process.

After predicting a different view image to compose a stereo image pair, we calculate the disparity map by extending SGBM semi-global matching algorithm with ELAS semi-local matching approach. Taking a aggregation cost from all neighbor directions into account, a weighted least squares filter is applied to refine the initial matching result by SGBM. Further enhancing the matching performance by ELAS semi-local matching approach, a Bayesian inference strategy is utilized to improve the disparity based on image similarity score.

3.3 DEPTH FUSION BASED ON CONFIDENCE MAP

Spectral consistency constrained singe image depth estimation can estimate the smooth and flat region accurately, but suffers from blur boundary and depth discontinuities issue. Single image stereo reconstruction can lead to the boundary and textured regions matching more effectively due to another camera perspective view constraints. However, it is less capable in resolving the matching confusion of surface without textures or with repeated textures. To tackle with these problems, we explore confidence maps to improve the refined disparity. Specifically, confidence maps are trained separately in a self-supervised fashion without using ground truth depth labels. Given one disparity map as input, we can assign labels to each pixel value belonging to the range of [0, 1] (0 is not confident at all and 1 is totally confident) on a confidence map C based on conventional confidence

(a) (b) (c) (d) (e)

Figure 2: Confidence maps from the corresponding predicted disparities for two approaches respectively. (a) Input color image. (b) Predicted disparity from spectral consistency constrained method. (c) Predicted disparity from single image stereo reconstruction. (d) Confidence maps for (b). (e) Confidence map for (c).

measures provided in Tosi & et al. (2017). For predicted-view based single image stereo method, we adopt the Winner Margin (WMN), Average Peak Ratio (APKR), Left-Right Consistency (LRC), Matching Score (MC) and Distance to Left Border (DLB) as the metrics. For spectral consistency based disparity prediction method, only Median Deviation of diaprtiy (MED) , Disparity Agreement (DA) and Variance of the Disparity Values (VAR) are applied. Then the final confidence score S_c for a pixel is the sum of all confidence scores for different measure metrics C_k, as $S_c = \sum_k C_k$. The confidence maps from the corresponding disparities of two branches are given as Fig. 2.

The final refined disparity $d_{refined}$ can be selected and expressed as a combination of the initial disparity maps $\widetilde{d_1}$ and $\widetilde{d_2}$ from two branches :

$$d_{refined} = \begin{cases} \widetilde{d_1(x,y)}, & \text{if } S_{c1}(x,y) < S_{c2}(x,y) \\ \widetilde{d_2(x,y)}, & \text{if } S_{c1}(x,y) \geq S_{c2}(x,y) \end{cases} \tag{6}$$

We output the actual value for each pixel in confidence score map $S_{c1}(x,y)$ and $S_{c2}(x,y)$. With our fusion strategy, the refined disparity map is able to leverage both benefits of spectral consistency disparity estimation network and view-prediction based single image stereo technique, achieving a better performance in visual and quantitative evaluation which will be discussed in Sec. 4.

3.4 End-to-end training of the whole pipeline

These two networks can be combined for joint training once being trained to obtain the ability of geometric reasoning for the task of view synthesis and stereo matching separately.

4 Experimental Results

In this section, we will describe the dataset and experiment setup. Then the proposed method for disparity prediction and refinement is evaluated on KITTI 2015 (Menze & et al. (2015)) and Cityscapes (Cordts et al. (2016)) dataset compared with other recent state-of-the-art approaches. An ablation analysis is provided in Sec. 4.3 to prove the effectiveness of each component,

4.1 Datasets

KITTI stereo dataset contains 61 scenes (42382 stereo images) with an embedded LiDAR calibrated together with the left color camera to build sparse ground truth depth map. The 697 images covering 29 scenes are served as the test split and the remaining 22600 images from the rest 32 scenes is used for training. We conduct our experiment quantitatively and visually to show the performance of the proposed method compared with other recent approaches. To compare with other works in a consistent manner, we only evaluate on a cropped region proposed by Eigen et al. (2014). We provide our result using both the cap of 0-80m (following Yang et al. (2019)) and 1-50m (following Garg et al. (2016)). This requires to discard the pixels on which the depth is outside the proposed range. Cityscapes dataset is a large-scale dataset collected in over 50 cities in Germany which containing 19 semantic classes and 22973 stereo image pairs. DrivingStereo dataset (Yang et al. (2019)) was originally used for large-scale stereo matching in urban scenarios. In our experiment, we cropped the bottom part of the images to exclude the car hood and resize it to meet our network setting as additional examples to show the ability of our method when generalizing to a new scene without further training on it.

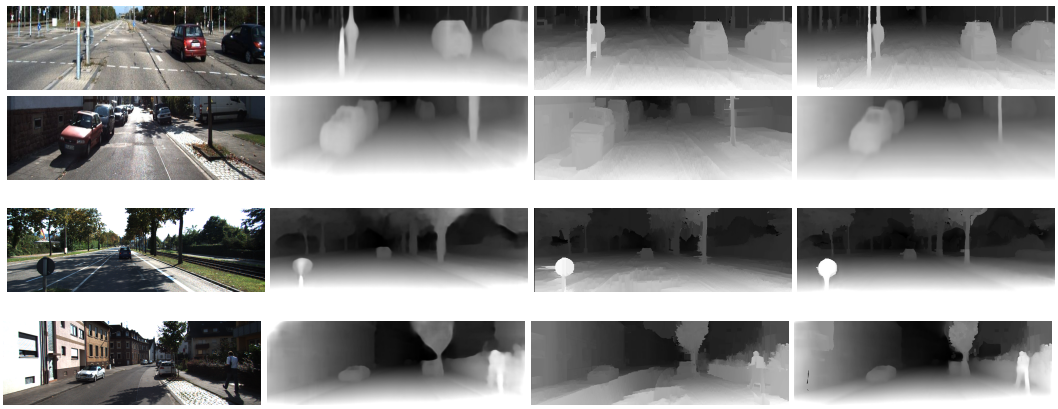

Figure 3: Depth estimation samples from our refinement compared with depths from our two depth estimation separately without fusion. Left to right: Input color image; Output from spectral consistency based estimation only; Output from single image stereo network only; Our full output.

Methods	Input type	Fusion	Error
Stereo only	Stereo	No	5.09%
Stereo gt only	Stereo	No	3.32%
Mono only	Mono	No	7.68%
Fusion by Ferrera & et al. (2019)	Stereo	Yes	3.03%
Our fusion strategy	Mono	Yes	3.01%

Table 1: Comparison of our different modules with and without the proposed fusion strategy.

4.2 Implementation

The architecture in our work is implemented with PyTorch framework. For the spectral consistency based method, images are resized to 512 $\times$ 256 before feeding into the ResNet-50 based network. Adam optimizer with β_1 =0.9 and β_2 = 0.999 together with a batch size of 8 and learning rate of 1e-4 are applied during training. Same data augmentation strategy as in Godard et al. (2017) is employed to augment data and train the network more robust. During the inference time, the predicted disparity map are post-processed to eliminate the effect of dis-occlusions (Godard et al. (2017) Poggi & et al. (2018)). Then the output are up-sampled to fit the original image size for later fusion and evaluation.

For the predicted view-based single image stereo network, We train it from scratch with also Adam where β_1 =0.9 and β_2 = 0.999 and $\epsilon = 10^{-8}$. The encoder starts with three convolutional layers, followed by LeakyRelu layers. The filter size starts from 7×7 and gradually decrease to 3×3 to dig into more details in small feature maps. The decoder consists of 4 convolutional layers with Relu activation, excepting tanh activation for the last layer. Batch normalization, up-sampling, dropout, and skip-connection layerS are also applied in the decoder. The initial learning rate is set to be 1e-4, and we linearly decrease the rate to zero over the totally 100 epochs. During the inference phase, only one test image is required to generate the refined output disparity from our pipeline. And standard metrics for monocular depth estimation as Abs rel, Sq rel, RMSE, RMSE log and three accuracy metrics under different threshold 1.25, 1.25^2 and 1.25^3 are applied for evaluation.

4.3 Performance Analysis

To compare with the performance of each component of our proposed method and explore the effectiveness of the proposed fusion pipeline, We first show visual comparison on given examples with and without our refinement in Fig. 3. It can be observed that refined outputs leverage both benefits from the two frameworks, especially in overcoming the blur boundary in disparity prediction network and miss-matching regions in single image stereo matching. It can be observed that in the regions with object occlusions such as traffic lights, trees and billboards, our full pipeline is able to deal with the object occlusions and separate them better. Furthermore, we evaluate it on KITTI 2015 stereo evaluation dataset containing 200 images for testing. We report the result of the percentage of miss-classified pixels (over 3 pixels) relative to the input in Table 1. We first show the error percentage of each of our separate prediction methods without fusion. Then we demonstrate that the refined result largely outperforms initial disparities from either of these two frameworks alone (41.6% enhancement over single image stereo matching and 60.8% over disparity prediction from single image). Last, by comparing with Ferrera & et al. (2019) which also adopts fusion idea, our

Methods	Training type	Error				Accuracy		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen Coarse Eigen et al. (2014)	Supervised	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen Fine Eigen et al. (2014)	Supervised	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Deep CRF Liu et al. (2016)	Supervised	0.202	1.614	6.523	0.275	0.678	0.895	0.965
SfMLearner Zhou et al. (2017)	Unsupervised	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Vid2depth Mahjourian et al. (2018)	Unsupervised	0.163	1.240	6.220	0.250	0.762	0.916	0.968
DFNet Zou & et al. (2018)	Unsupervised	0.150	1.124	5.507	0.223	0.806	0.933	0.973
GeoNet-ReNet Yin & Shi (2018)	Unsupervised	0.155	1.296	5.857	0.233	0.793	0.931	0.973
CC Ranjan & et al. (2019)	Unsupervised	0.140	1.070	5.326	0.217	0.826	0.941	0.975
Garg et.al. Garg et al. (2016)	Unsupervised	0.152	1.226	5.849	0.246	0.784	0.921	0.967
MonoDepth Godard et al. (2017)	Unsupervised	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Unsup-depthGAN Pilzer & et al. (2018)	Unsupervised	0.152	1.388	6.016	0.247	0.789	0.918	0.965
3-Net Poggi & et al. (2018)	Unsupervised	0.142	1.207	5.702	0.240	0.809	0.928	0.967
EveryPixel++ Luo & et al. (2019)	Unsupervised	0.141	1.224	5.548	0.218	0.811	0.934	0.972
Monodepth2 Godard & et al. (2019)	Unsupervised	0.130	1.144	5.485	0.232	0.831	0.932	0.968
Ours w/o stereo	Unsupervised	0.139	1.102	5.483	0.224	0.818	0.932	0.972
Our full wo pp	Unsupervised	0.125	1.181	5.142	0.218	0.838	0.941	0.978
Our full	Unsupervised	**0.120**	**1.061**	**5.138**	**0.216**	**0.840**	**0.942**	**0.978**

Table 2: Quantitative comparison with other recent methods. All methods just use KITTI training split for a fair comparison. We compare with both supervised and unsupervised methods taking single images (Eigen et al. (2014) Liu et al. (2016) Garg et al. (2016) Godard et al. (2017) Godard et al. (2017) Poggi & et al. (2018) Luo & et al. (2019) Godard & et al. (2019)) or monocular video (Zhou et al. (2017) Mahjourian et al. (2018) Zou & et al. (2018) Ranjan & et al. (2019)) or image pair Pilzer & et al. (2018) as input.

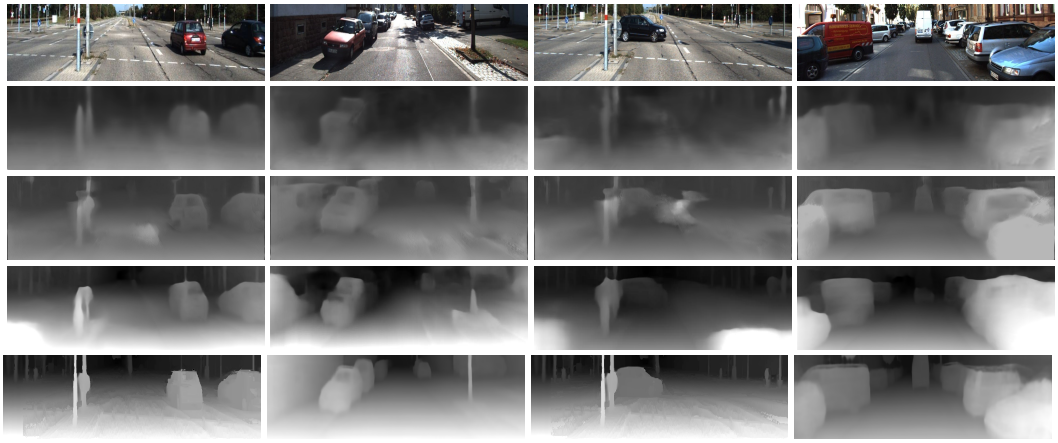

Figure 4: Visual performance of the proposed method compared with other recent methods (Godard & et al. (2019) Poggi & et al. (2018) Yin & Shi (2018)). First row: Input image; Second row: Disparity from GeoNet (Yin & Shi (2018)). Third row: Disparity from 3-Net (Poggi & et al. (2018)); Fourth row: Disparity from Monodepth2 (Godard & et al. (2019)); Last row: Our pipeline output.

method is still able to maintain better performance (3.03% v.s. 3.01%). Note that Ferrera & et al. (2019) must input stereo images during inference and our framework only requires one image.

To go insight into the proposed method, we investigate the performance of our pipeline in relation to the state-of-the-art methods. Referred to Table 2, we conduct the experiments in 1 to 80 meters' range and compare the result with other supervised methods (Eigen et al. (2014) Liu et al. (2016)) and self-supervised methods (Garg et al. (2016) Zhou et al. (2017) Godard et al. (2017) Yin & Shi (2018) Pilzer & et al. (2018) Mahjourian et al. (2018) Poggi & et al. (2018) Zou & et al. (2018) Luo & et al. (2019) Ranjan & et al. (2019) Godard & et al. (2019)). Our proposed method performs the best evaluation in both loss and accuracy metrics. In particular, we observe that our method with the refined process is able to achieve an obvious improvement in the Sq Rel metric. With and without combing predicted-view based single image stereo algorithm makes a difference in the prediction results, which further proves that the intuition of this work. Relevant performance can be better perceived from the visual comparison in Fig. 4. The results from Yin & Shi (2018) and Poggi & et al. (2018) have many artifacts and blur regions in the frame due to the wrong predictions. Also, it can be noticed that there is also an overall scale problem in their predictions. For Poggi & et al. (2018), it appears some large discontinuities, especially on the ground. Monodepth2 (Godard & et al. (2019)) achieve the best visual performance among the three comparisons, but it still shows weak ability to preserve object boundaries and deal with occlusions compared with our pipeline.

It can be observed from Fig. 4 that our full pipeline generate more clear object boundaries and efficiently prevent from the miss-matching issue on flat regions.

Method	Mean SSIM	Mean PSNR	MAD
Zhou et al. (2016)	0.59	15.00	0.25
Sun et al. (2018)	0.68	18.49	0.15
Ours	0.74	19.98	0.13

Table 3: Quantitative comparison with other neighbor view synthesis methods. Higher mean SSIM/PSNR and lower MAD means better performance in synthesising.

Table. 3 shows mean SSIM, mean PSNR and Mean Absolute Difference (MAD) metrics for each method across the test set. We use these three metrics to measure if one method is averagely better than another. For both PSNR and SSIM, one method is always better than the others with the higher score, and for MAD, one method is better if with the smaller values than others. From Table 3, our method achieves best numerical score in both SSIM (0.74) and PSNR (19.98) than Sun et al. (2018) and Zhou et al. (2016). With respect to MAD, our method achieves a 13.3% and 48.0% decrease in comparison with Sun et al. (2018) and Zhou et al. (2016) respectively, demonstrating the effectiveness of our predicted images used for stereo reconstruction.

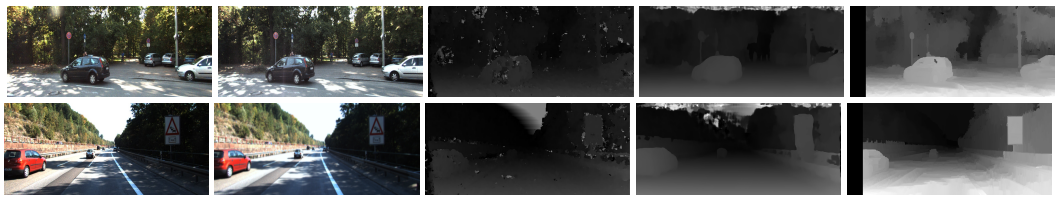

Figure 5: Comparison with other learning-based stereo matching methods from our synthesized image pairs. Left to right: Left input image; Synthesized right image; Disparity output from Luo et al. (2016); Disparity output from Chang & et al. (2018); Our single image stereo output.

Additional experiments analyze the effectiveness of our single image stereo matching approach based on predicted view. Fig. 5 shows visual examples of the performance of our single image stereo matching algorithm in comparison with the recent learning-based disparity estimation method from stereo images. In particular, from the comparison with Luo et al. (2016), we can notice that our proposed method is able to prevent many wrong matching on flat regions. Compared with the most recent state-of-the-art stereo matching method Chang & et al. (2018), which uses ground truth labels for training, our method still can achieve comparable performance, especially in the sky.

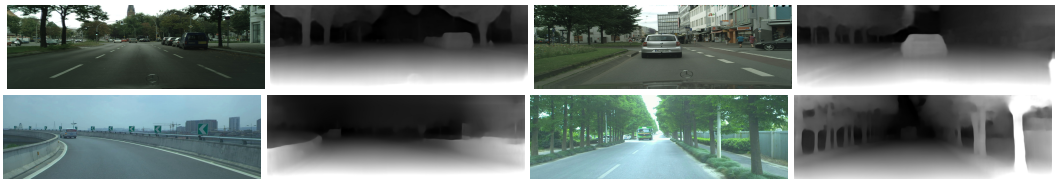

Figure 6: Qualitative results of our pipeline on Cityscapes and DrivingStereo stereo dataset without training or fine-tuning. Upper two images are from Cityscapes, the rest are from DrivingStereo.

Figure 7: Examples of 3D reconstruction from the original input color image and the estimated disparity map from our full pipeline.

Finally, we illustrate further examples in Fig. 6 and Fig. 7. With the model only trained on KITTI split of Eigen, we tested directly on the DrivingStereo and Cityscapes dataset though with differences in camera parameters, weather and city. Our method benefits from both deep neural network depth prediction and physical-informed stereo reconstruction with just one image input, which can overcome the issues from both depth estimation strategies.

5 CONCLUSION

In this paper, we investigated a single image depth estimation framework that is comprised of two newly proposed depth estimation approaches, spectral constrained single image depth estimation algorithm and single image stereo reconstruction based on predicted view. We also propose a novel pipeline for dense depth fusion from only one single image as input. Experiments show that the proposed approach leads to a more precise depth estimation performance compared with either only applied monocular disparity prediction network or single image stereo algorithm, and is able to generate comparable results compared with the state-of-the-art approaches. Moreover, our proposed full pipeline is able to infer a high-quality disparity with more clear object boundaries and less miss-predictions in flat regions compared with most recent approaches. Different with existing methods, our pipeline is able to predict accurate disparity map only from one single image as input.

REFERENCES

Kumar Bipin and et al. Autonomous navigation of generic monocular quadcopter in natural environment. In *ICRA*, 2015.

Jure Bontar and et al. Stereo matching by training a convolutional neural network to compare image patches. *The journal of machine learning research*, 2016.

Yuanzhouhan Cao and et al. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *CSVT*, 2017.

Jia-Ren Chang and et al. Pyramid stereo matching network. In *CVPR*, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Neural Information Processing Systems*, 2014.

Maxime Ferrera and et al. Fast stereo disparity maps refinement by fusion of data-based and model-based estimations. In *3DV*, 2019.

Huan Fu and et al. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

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, 2016.

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*, pp. 270–279, 2017.

Clément Godard and et al. Digging into self-supervised monocular depth estimation. In *CVPR*, 2019.

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 *CVPR*, 2017.

Taeksoo Kim and et al. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017.

Alex Krizhevsky and et al. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

Jae-Han Lee and et al. Monocular depth estimation using relative depth maps. In *CVPR*, 2019.

Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *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 Transactions on Pattern Analysisand Machine Intelligence*, 38(10):2024–2039, 2016.

Chenxu Luo and et al. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *TPAMI*, 2019.

Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5695–5703, 2016.

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

Nikolaus Mayer and et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

Moritz Menze and et al. Object scene flow for autonomous vehicles. In *CVPR*, 2015.

Ji-Hun Mun and et al. Temporally consistence depth estimation from stereo video sequences. In *PRCM*, 2015.

Matthias Ochs, Adrian Kretz, and Rudolf Mester. Sdnet: Semantically guided depth estimation network. In *GCPR*, 2019.

Andrea Pilzer and et al. Unsupervised adversarial depth estimation using cycled generative networks. In *3DV*, 2018.

Matteo Poggi and et al. Learning monocular depth estimation with unsupervised trinocular assumptions. In *3DV*, 2018.

Pierluigi Zama Ramirez and et al. Geometry meets semantics for semi-supervised monocular depth estimation. In *ACCV*, 2018.

Anurag Ranjan and et al. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Learning 3-d scene structure from a single still image. In *ICCV*, 2007.

Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 155–171, 2018.

Fabio Tosi and et al. Learning confidence measures in the wild. In *BMVC*, 2017.

Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zili Yi and et al. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Huangying Zhan and et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018.

Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pp. 286–301. Springer, 2016.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851–1858, 2017.

Jun-Yan Zhu and et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pp. 2223–2232, 2017.

Yuliang Zou and et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, 2018.

Laurent Zwald and Sophie Lambert-Lacroix. The berhu penalty and the grouped effect. *arXiv preprint arXiv:1207.6868*, 2012.