UNSUPERVISED MONOCULAR DEPTH ESTIMATION BASED ON DUAL ATTENTION MECHANISM AND DEPTH-AWARE LOSS

Xinchen Ye[1], Mingliang Zhang[1-2], Rui Xu[1], Wei Zhong[1], Xin Fan[1], Zhu Liu[1], Jiaao Zhang[1]*

DUT-RU International School of Information Science & Engineering, Dalian University of Technology
Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China[1]
School of Mathematical Sciences, Dalian University of Technology of Liaoning Province, China[2]

ABSTRACT

Most existing monocular depth estimation approaches are supervised, but enough quantities of ground truth depth data are required during training. To cope with this, recent techniques deal with the depth estimation task in an unsupervised manner, i.e., replacing the use of depth data with easily obtained stereo images for training. Based on this, we propose a novel unsupervised learning architecture, which integrates dual attention mechanism into the framework and designs a depth-aware loss for better depth estimation. Specifically, to enhance the ability of feature representations, we introduce a dual attention module to capture global feature dependencies in spatial and channel dimensions for scene understanding and depth estimation. Meanwhile, we propose a depth-aware loss that fully addresses the occlusion problem in brightness constancy assumption, the intrinsic characteristics of depth map, and the left-right consistency problem, respectively. Besides, an adversarial loss is employed to discriminate synthetic or realistic depth maps by training a discriminator so as to produce better results. Extensive experiments on KITTI dataset show that our approach achieves state-of-the-art performance compared with other monocular depth estimation methods.

Index Terms— Unsupervised, Monocular, Depth estimation, Dual attention, Depth-aware loss

1. INTRODUCTION

Monocular depth estimation is an important visual task in computer vision. It has many applications in a number of challenging problems, such as augmented reality, object detection and visual odometry. With the rapid development of deep learning, many approaches utilize convolutional neural networks (CNNs) to estimate depth map due to their ability to represent complex relationship between depth maps and RGB images. In recent years, supervised learning methods demonstrate promising results for monocular depth prediction [1, 2]. However, those methods need large quantities of annotated data that are usually sparse or not easy to be captured by depth-sensing equipment. On the contrary, unsupervised learning methods [3, 4, 5, 6, 7] do not require any depth data, and transform depth estimation as image reconstruction problem that uses epipolar geometry constraints to generate target images during training. Specifically, given a pair of rectified left and right images, these unsupervised methods attempt to learn a disparity map of the current view, then use the image from the other view with the learned disparity map to synthesize the image of current view. The photometric loss is established between synthetic image and original image of the same view, which can be regarded as a supervised loss function to guide the network training. Given the camera parameters, the learned disparity map can be converted to the depth map immediately.

In this paper, we stand on the unsupervised techniques, and propose to integrate dual attention mechanism [8] in our framework and design a depth-aware loss for better monocular depth estimation. First, self-attention mechanism aims at extracting the relative features with contextual interdependency to enhance the ability of feature representations for pixel-level regression. So, we introduce a dual attention module to exploit nonlocal dependencies along the spatial and channel dimensions for better understanding the scene geometry. Then, we propose a depth-aware loss that simultaneously addresses 1) the occlusion problem to maintain the brightness constancy assumption; 2) the statistical property of depth signals that mainly contains smooth regions separated by additive step discontinuities, and 3) the left-right consistency problem [3]. Besides, an adversarial loss [9] is employed to discriminate synthetic or realistic depth maps by training a discriminator so as to produce better results. Our main contributions can be summarized as follows:

1) A dual attention module is employed to capture feature interdependencies in spatial and channel dimensions for better scene understanding.

2) A depth-aware loss is proposed to identify occlusion regions when imposing brightness constancy assumption, and to capture the statistical characteristics of depth map as priors

*This work is supported by the National Natural Science Foundation of China (Nos. 61702078, 61733002 and 61572096), and the Fundamental Research Funds for the Central Universities. Corresponding author: Xinchen Ye. E-mail: yexch@dlut.edu.cn.

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00037

to regularize the generated depth map.

3) Through experimental evaluation on public KITTI dataset [10], we demonstrate the effectiveness of our dual attention module and depth-aware loss, and show the proposed approach outperforms other state-of-the-art methods.

2. RELATED WORK

In the past decades, significant efforts of monocular depth estimating have been made in the research community. Previous approaches mainly focus on the graphical models with hand-crafted geometry priors [11, 12]. Recently, CNNs have been extensively applied into the depth estimation task. Eigen *et al.* [13] proposed the first multi-scale CNN framework in a coarse-to-fine manner. Liu *et al.* [14] considered learning the unary and pairwise terms of continuous conditional random field (CRF) combined with deep CNN. However, those above methods are supervised methods, and large numbers of annotated data that are generally difficult to obtain, are required. To deal with this issue, some advanced approaches are unsupervised and avoid the use of costly labelled ground-truth of depth map during training. Garg *et al.* [5] first proposed a deep unsupervised network based on image synthetic by linearizing the reconstruction loss to make it differentiable. Zhou *et al.* [6] simultaneously estimated monocular depth and camera pose from unlabelled video sequences. Owing to the superior practicability of the unsupervised techniques, we also develop a unsupervised learning framework as our backbone, and design a depth-aware loss for better monocular depth estimation.

Recently, self-attention mechanism has been widely used in computer vision by capturing long-range dependencies over feature maps. Hu *et al.* [15] proposed the "Squeeze-and-Excitation" (SE) block designed to perform channel-wise feature recalibration. Woo *et al.* [16] introduced the convolutional block attention module (CBAM) which is applied to refine dependencies between spatial and channel features. Wang *et al.* [8] computed the response at a position of current pixel by attending to all positions in a neighborhood of the current pixel and taking their weighted average in an embedding space to capture spatial nonlocal relationship. Motivated by the attention mechanism in [8], we design spatial attention and channel attention modules, termed as dual attention scheme, applied to our unsupervised learning architecture.

3. PROPOSED METHOD

In this section, we first give the problem statement for depth estimation. Then, we introduce the dual attention mechanism that is employed in our network. Finally, we present our proposed depth-aware loss for the depth estimation task. Fig. 1(a) illustrates the overview of the proposed network. Our fully convolutional architecture is constructed based on an encoder-decoder U-net structure [17]. We apply the dual attention scheme to the end of the encoder to model the high-level nonlocal dependencies, as marked by red dash box.

3.1. Problem Statement

Given a pair of calibrated stereo images $\{I_l, I_r\}$, our goal is to learn a nonlinear mapping $\mathcal{F}$: $\mathcal{C} \to \mathcal{D}$ by training the network, where $\mathcal{C}$ is the color image space, $\mathcal{D}$ is the depth map space. The left image I_l is used as input of the network and produces two disparity map d_l, d_r that are employed to synthesize the right image $\hat{I}_r$ and left image $\hat{I}_l$ by warping functions, respectively. The proposed depth-aware loss is used to penalize the difference between the synthesized image and the corresponding groundtruth. Then, the synthesized images are put into the discriminator to differentiate whether they are true or fake. In the test phase, for an arbitrary color image in space $\mathcal{C}$, our network can output the corresponding disparity d, which indicates an offset of two corresponding points for a pair of rectified left I_l and right image I_r. If the baseline distance b between the cameras and the camera focal length f are given, the depth map z is obtained by $z = bf/d$.

Similar to [5], we utilize epipolar geometry as constraint to establish the supervised photometric loss, thus the depth estimation can be converted into image reconstruction problem. Given the training data of rectified stereo pairs $\{I_l^i, I_r^i\}_{i=i}^N$ and the predicted disparity maps $\{d_l^i, d_r^i\}_{i=1}^N$, we can synthesize the left image and the right image by using the warping function $f_\omega(\cdot)$ as follows:

$$\hat{I}_l^i = f_\omega(I_r^i, d_l^i), \quad \hat{I}_r^i = f_\omega(I_l^i, d_r^i) \tag{1}$$

By adopting the bilinear sampler [18] in the warping function, the above formula is fully differentiable to facilitate back-propagation in network training. After obtaining the synthesized images $\hat{I}_l^i, \hat{I}_r^i$, the photometric loss has the following form:

$$L_p = L_p^l + L_p^r \tag{2}$$

where

$$L_p^t = \frac{1}{N}\sum_{i=1}^{N} ||I_t^i - \hat{I}_t^i||, \tag{3}$$

where $t = \{l, r\}$. For simplicity, we mainly give the loss function about left image, namely $t = l$ in the following sections. Note that, collecting dense depth ground-truth is very difficult and time consuming. Instead of using the real data of depth map, the unsupervised method simply requires stereo images for optimizing the network during training.

3.2. Dual Attention Module

Through our observation, neighbouring pixels in an image with similar appearances should have close depth, explicitly utilizing the relationship among pixels in neighbourhood

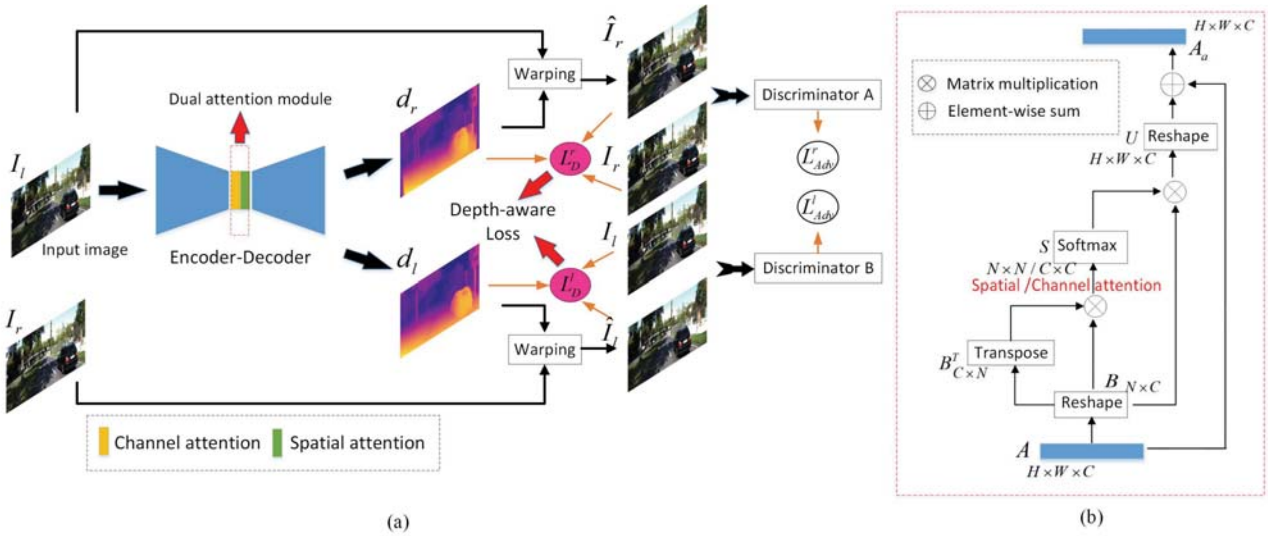

Fig. 1. (a) Overview of the proposed framework; (b) Illustration of channel or spatial attention module. Note that we apply the dual attention scheme to the end of the encoder. The depth-aware loss is proposed to guide the network training.

could contribute to the task of depth estimation. Therefore, the spatial attention module is exploited to capture nonlocal dependencies over local features of depth map. Besides, we also leverage the channel attention module considering the feature dependencies in the channel dimension. Both attention modules, namely dual attention, are connected to the end of the encoder part in a sequential manner.

As shown in Fig. 1(b), for the feature map $A \in \mathbb{R}^{H\times W\times C}$, we reshape A to $B \in \mathbb{R}^{N\times C}$, where H, W, C are the height, width and channels respectively, and $N = H \times W$. Next, we apply the operation of matrix multiplication between the B and its transpose. A softmax layer is employed to compute the spatial attention map $S \in \mathbb{R}^{N\times N}$, i.e., $S = softmax(BB^T)$. Then, we perform another matrix multiplication between S and B and reshape it to $U \in \mathbb{R}^{H\times W\times C}$. At last, the local feature A is added (element-wise sum) to U to obtain the final output A_a. Note that, the channel attention module has the same flow chart as the spatial attention module, except that the channel attention map $S \in \mathbb{R}^{C\times C}$ is calculated in the channel dimension ($S = softmax(B^T B)$).

3.3. Depth-Aware Loss

Due to the existence of parallax between the stereo images, it will appear some occlusion regions in which pixels from left image and right image cannot be matched using the epipolar geometry constraint. Although the photometric loss in Eq. 3 is held under the brightness constancy assumption, it is invalid for the occlusion region which may lead to erroneous results on the reconstructed depth map. Therefore, we design an occlusion mask whose values indicate whether the corresponding pixels are occluded. In particular, the occlusion mask is defined as $M = \mathbb{I}(\delta \geq 0)$ [20], where $\mathbb{I}(\cdot)$ is an indicator function and δ is defined as:

$$\delta = ||I_l - \hat{I}_l||^2 - (\eta_1||I_l||^2 + \eta_1||\hat{I}_l||^2 + \eta_2). \tag{4}$$

where η_1 and η_2 are parameters. Let $M_* = 1 - M$, so we have the revised photometric loss:

$$L^l_p = \frac{1}{N}\sum_{i=1}^{N} M^i_* \circ ||I^i_l - \hat{I}^i_l||, \tag{5}$$

where $\circ$ means pixel-wise multiplication.

Next, through our observation, depth map usually contains smooth regions curved by step discontinuities. The gradients of the natural depth map conform to a heavy-tailed distribution, which can be approximately modeled by leveraging total variation (TV) filter as a edge preserving regularizer. However, TV is sub-optimal in terms of producing sparsity. By contrast, we resort to use total generalized variation (TGV) filter [21] with second-order gradients to portray features of depth map, resulting the following regularization term:

$$L^l_s = \frac{1}{N}\sum_{i=1}^{N}(||\nabla^2_x d^i_l||e^{-||\nabla^2_x I^i_l||} + ||\nabla^2_y d^i_l||e^{-||\nabla^2_y I^i_l||}). \tag{6}$$

Besides, we also add left-right consistency check [3] to our depth-aware loss to ensure coherence for the left image:

$$L^l_d = \frac{1}{N}\sum_{i=1}^{N} ||d^i_l - \hat{d}^i_l||, \tag{7}$$

where $\hat{d}^i_l$ is a warped disparity map of left view, which can be obtained by applying warping function on the generated left disparity map, i.e., $\hat{d}^i_l = f_\omega(d^i_r, d^i_l)$.

Table 1. Quantitative results using the Eigen split [13] on KITTI [10]. For training, K means stereo input pairs and K(M) denotes the input unlabeled monocular video, which recasts the successive two frames as a stereo pair. Resnet means the use of Resnet50 architecture as the backbone. Sup indicates the training manner: supervised (Yes) or unsupervised (No). The results from Grag *et al.* [5] are capped at 50m and we separately list them for comparison. The best results for supervised setting are in italics with underline and for unsupervised marked as bold-face. Note that, the higher the better for the accuracy δ, and the lower the better for the metrics of Abs Rel, Sq Rel, RMSE, RMSE log.

Method	Dataset	Sup	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Saxena *et al.* [12]	K	Yes	0.280	-	8.734	-	0.601	0.820	0.926
Eigen *et al.* [2]	K	Yes	0.203	1.548	*6.307*	0.246	0.702	0.890	0.958
Liu *et al.* [14]	K	Yes	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Xu *et al.* [1]	K	Yes	*0.132*	*0.911*	-	*0.162*	*0.804*	*0.945*	*0.981*
Pilzer *et al.* [4]	K	No	0.152	1.388	6.016	0.247	0.789	0.918	0.965
Godard *et al.* [3]	K	No	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhou *et al.* [6]	K(M)	No	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yin *et al.* [7] Resnet	K(M)	No	0.155	1.296	5.857	0.233	0.793	0.931	**0.973**
Zou *et al.* [19] Resnet	K(M)	No	0.150	1.124	**5.507**	**0.223**	0.806	0.933	**0.973**
Ours	K	No	0.147	1.299	5.882	0.245	0.804	0.923	0.969
Ours Resnet	K	No	**0.135**	**1.120**	5.560	0.230	**0.808**	**0.934**	**0.973**
Garg *et al.* [5], 50m	K	No	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Pilzer *et al.* [4], 50m	K	No	0.144	1.007	4.660	0.240	0.793	0.923	0.968
Godard *et al.* [3], 50m	K	No	0.140	0.976	4.471	0.232	**0.818**	0.931	**0.969**
Ours, 50m	K	No	**0.138**	**0.965**	**4.456**	**0.231**	**0.818**	**0.932**	**0.969**

In summary, our proposed depth-aware loss L_D is written as:

$$L_D = \alpha_1(L_p^l + L_p^r) + \alpha_2(L_s^l + L_s^r) + \alpha_3(L_d^l + L_d^r). \quad (8)$$

where α_1, α_2, α_3 are the weighting factors.

Besides, an adversarial loss is employed to discriminate synthetic or realistic depth maps by training a discriminator in the proposed framework. The above mentioned encoder-decoder network together with dual attention scheme can be regarded as a generator. Then, we add a discriminator D in the end of the generator, which takes the synthesized left/right images produced by generator as input, and distinguishes them from the the corresponding realistic images. The adversarial loss for the left image can be expressed as follows:

$$\begin{aligned} L_{Adv}^l = \frac{1}{N}\sum_{i=1}^{N}\{&\mathbb{E}_{I_l^i \sim \mathcal{P}(I_l^i)}[\log D(I_l^i)] \\ &+\mathbb{E}_{\hat{I}_l^i \sim \mathcal{P}(\hat{I}_l^i)}[\log(1 - D(\hat{I}_l^i))]\}, \end{aligned} \quad (9)$$

where $\mathcal{P}(z)$ denotes the probability distribution over data z. Notably, this adversarial loss promotes the generator to learn a mapping from synthetic to realistic data, which make the synthetic image similar to the realistic image. To conclude, the final training objective has the following formula:

$$L_{final} = L_D + \alpha_4(L_{Adv}^l + L_{Adv}^r) \quad (10)$$

where α_4 is the weighting factor.

4. EXPERIMENTS

We evaluate the proposed framework on public KITTI dataset [10] to compare with the existing methods and present the qualitative and quantitative results.

Training details. We use VGG-16 architecture as the backbone for the encoder component. For the decoder component, we flip the encoder, and replace the downsampling layers with deconvolution layers. The skip connection is used between feature maps with the same spatial dimension from both encoder and decoder components to obtain more effective feature representation. The discriminator D has five 3×3 convolutions with the downsampling rate 2, and the batch normalization is added after each convolution. The proposed method is implemented using the TensorFlow framework. For training, the batch size is set to 8 using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-6}$. The initial learning rate is 10^{-4}. The weighting factors $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are set to 0.85, 0.1, 1 and 0.1, respectively. In Eq. 4, η_1 and η_2 are set to 0.01 and 0.5 respectively.

KITTI dataset. The KITTI dataset contain outdoor scenes collected by moving vehicles. We use the Eigen split [13] for KITTI dataset, which have 22600 image pairs for training and 697 for testing. Meanwhile, the KITTI split [3] that keeps 29000 pairs for training and 200 for testing is employed to perform the ablation study. Several measures commonly used in [13] are applied for quantitative evaluation: mean relative error (Abs Rel): $\frac{1}{N}\sum_{i=1}^{N}\frac{||d_i^{gt} - \hat{d}_i||}{d_i^{gt}}$,

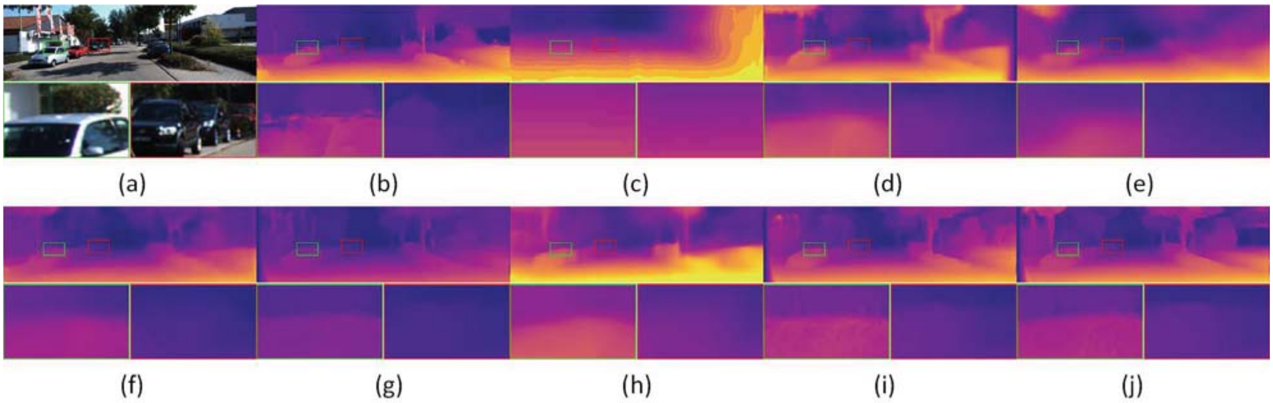

Fig. 2. Examples of monocular depth estimation result on KITTI dataset [10]. (a) color image, (b) GT, (c) Eigen *et al.* [2], (d) Garg *et al.* [5], (e) Zhou *et al.* [6], (f) Yin *et al.* [7], (g) Godard *et al.* [3], (h) Zou *et al.* [19], (i) Ours, (j) Ours Resnet.

Table 2. Comparison of different variants of the proposed method. Results are evaluated according to KITTI split [3], including 200 stereo pairs.

Method	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
baseline	K	0.124	1.396	6.137	0.217	30.352	0.841	0.937	0.975
baseline+attention	K	0.119	1.315	6.060	0.202	28.409	0.849	0.938	0.978
baseline+L_D	K	0.121	1.314	6.049	0.200	25.882	0.844	0.937	0.976
baseline+L_D+attention	K	0.114	1.287	5.790	0.196	23.429	0.857	0.942	0.978
baseline+L_{final}+attention	K	**0.104**	**1.269**	**5.481**	**0.188**	**22.449**	**0.864**	**0.946**	**0.981**

squared relative error (Sq Rel): $\frac{1}{N}\sum_{i=1}^{N}\frac{||d_i^{gt}-\hat{d}_i||^2}{d_i^{gt}}$, root mean squared error (RMSE): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i^{gt}-\hat{d}_i)^2}$, mean log10 error (RMSE log): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}||\log d_i^{gt}-\log \hat{d}_i||^2}$, accuracy with threshold t: $\delta = \max(\frac{d_i^{gt}}{\hat{d}_i}, \frac{\hat{d}_i}{d_i^{gt}}) < t$, for $t \in [1.25, 1.25^2, 1.25^3]$, where N is the total number of pixels in the test set and d_i^{gt} and $\hat{d}_i$ are the ground truth and predicted depth values for pixel i.

4.1. Performance Comparison

Table 1 shows quantitative results compared with the state-of-the-art methods, including supervised setting trained with ground-truth depth map and unsupervised setting. Among the supervised methods, our results have achieved competitive performance and are slightly inferior to Xu *et al.* [1] that designs multi-scale CRF as the post-processing. Since the results of Yin *et al.* [7] and Zou *et al.* [19] are based on Resnet50 for the encoder part, we give the estimation marked as 'Ours Resnet' in the same setting for fairly comparison. It can be seen that our approach almost surpasses all unsupervised methods. We also evaluate with a maximum depth cap of 50 meters according to Grag *et al.* [5] and demonstrate our superiority. We achieves the best performance for all the metrics. Fig. 2 shows the visual comparison. In particular, the zoomed zones show that the proposed method can capture more detailed scene structures and, in contrast, the compared methods are difficult to obtain sharp boundaries, and fail to infer the scene geometry.

4.2. Ablation Study

To discover the vital elements in our proposed method, we conduct ablation study by gradually integrating each component into our framework. In Table 2, the 'baseline' that uses VGG-16 as the backbone in an encoder-decoder fashion cannot obtain a good results. When adding the dual attention module, the modification of network structure leads to obvious performance improvement, e.g. in case of *RMSE* from 6.137 to 6.060. Similar results can be obtained by replacing the standard photometric loss in 'baseline' with our proposed depth-aware loss L_D. Specifically, *D1-all* means the percentage of bad pixels (including occlusion pixels) in all pixels for ground truth disparity map [13], which is decreased from 30.352 for 'baseline' to 25.882 for 'baseline+L_D'. This clearly demonstrates that the proposed depth-aware loss L_D can handle occlusion regions well. When both attention module and depth-aware loss are employed, the performance are largely improved, e.g., *RMSE* is decreased to 5.790. Finally, the adversarial loss also improves the performance meaningfully. As a result, the complete proposed algorithm provides the state-of-the-art performance.

5. CONCLUSION

This paper proposes a novel unsupervised learning architecture, which integrates dual attention mechanism into the framework and designs a depth-aware loss for better depth estimation. Besides, an adversarial loss is employed to discriminate synthetic or realistic depth maps by training a discriminator so as to produce better results. Extensive experiments on KITTI dataset show that our approach achieves state-of-the-art performance compared with other monocular depth estimation methods.

6. REFERENCES

[1] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe, “Multi-scale continuous crfs as sequential deep networks for monocular depth estimation,” in *IEEE CVPR*, 2017, vol. 1.

[2] David Eigen and Rob Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *IEEE CVPR*, 2015, pp. 2650–2658.

[3] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *IEEE CVPR*, 2017, vol. 2, p. 7.

[4] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe, “Unsupervised adversarial depth estimation using cycled generative networks,” in *International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 587–595.

[5] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid, “Unsupervised cnn for single view depth estimation: Geometry to the rescue,” in *ECCV*. Springer, 2016, pp. 740–756.

[6] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe, “Unsupervised learning of depth and ego-motion from video,” in *IEEE CVPR*, 2017, vol. 2, p. 7.

[7] Zhichao Yin and Jianping Shi, “Geonet: Unsupervised learning of dense depth, optical flow and camera pose,” in *IEEE CVPR*, 2018, vol. 2.

[8] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, “Non-local neural networks,” in *IEEE CVPR*, 2018.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *NIPS*, 2014, pp. 2672–2680.

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[11] Kevin Karsch, Ce Liu, and Sing Bing Kang, “Depth transfer: Depth extraction from video using non-parametric sampling,” *IEEE TPAMI*, vol. 36, no. 11, pp. 2144–2158, 2014.

[12] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng, “3-d depth reconstruction from a single still image,” *International journal of computer vision*, vol. 76, no. 1, pp. 53–69, 2008.

[13] David Eigen, Christian Puhrsch, and Rob Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *NIPS*, 2014, pp. 2366–2374.

[14] Fayao Liu, Chunhua Shen, and Guosheng Lin, “Deep convolutional neural fields for depth estimation from a single image,” in *IEEE CVPR*, 2015, pp. 5162–5170.

[15] Jie Hu, Li Shen, and Gang Sun, “Squeeze-and-excitation networks,” *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[16] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, “Cbam: Convolutional block attention module,” in *ECCV*, 2018.

[17] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, “Learning deconvolution network for semantic segmentation,” in *IEEE CVPR*, 2015, pp. 1520–1528.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., “Spatial transformer networks,” in *NIPS*, 2015, pp. 2017–2025.

[19] Yuliang Zou, Zelun Luo, and Jia-Bin Huang, “Df-net: Unsupervised joint learning of depth and flow using cross-task consistency,” in *ECCV*. Springer, 2018, pp. 38–55.

[20] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer, “Dense point trajectories by gpu-accelerated large displacement optical flow,” in *ECCV*. Springer, 2010, pp. 438–451.

[21] Kristian Bredies, Karl Kunisch, and Thomas Pock, “Total generalized variation,” *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 492–526, 2010.