2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506283

CO-TEACHING: AN ARK TO UNSUPERVISED STEREO MATCHING

Hengli Wang[⋆] *Rui Fan*[†] *Ming Liu*[⋆]

[⋆] Hong Kong Unviersity of Science and Technology, Hong Kong SAR, China
[†] Tongji University, Shanghai 201804, China
hwangdf@connect.ust.hk, rui.fan@ieee.org, eelium@ust.hk

ABSTRACT

Stereo matching is a key component of autonomous driving perception. Recent unsupervised stereo matching approaches have received adequate attention due to their advantage of not requiring disparity ground truth. These approaches, however, perform poorly near occlusions. To overcome this drawback, in this paper, we propose CoT-Stereo, a novel unsupervised stereo matching approach. Specifically, we adopt a co-teaching framework where two networks interactively teach each other about the occlusions in an unsupervised fashion, which greatly improves the robustness of unsupervised stereo matching. Extensive experiments on the KITTI Stereo benchmarks demonstrate the superior performance of CoT-Stereo over all other state-of-the-art unsupervised stereo matching approaches in terms of both accuracy and speed. Our project webpage is https://sites.google.com/view/cot-stereo.

Index Terms— stereo matching, unsupervised learning, co-teaching strategy.

1. INTRODUCTION

Stereo matching is a fundamental problem in computer vision and robotics. This important technique has been widely employed in many tasks, such as robot vision [1, 2, 3] and autonomous driving [4, 5]. The goal of stereo matching is to estimate dense correspondences between a pair of stereo images and further generate a dense disparity image [6].

Traditional and data-driven approaches are two major types of stereo matching algorithms [6, 7]. Traditional algorithms formulate stereo matching as either a local block matching problem or a global energy minimization problem [6]. Data-driven approaches [8, 9, 10] utilize convolutional neural networks (CNNs) to extract informative visual features and create a 3D cost volume, by analyzing which a dense disparity image can be estimated. Among data-driven approaches, PSMNet [8] adopts 3D CNNs to regularize cost volumes for disparity estimation, while GwcNet [9] further utilizes group-wise correlation to provide efficient representations for visual feature similarity measurement. Meanwhile, LEAStereo [10] uses a neural architecture search framework to search an effective and efficient network for stereo matching. However, such supervised stereo matching approaches typically require a large amount of training data with disparity ground truth, often making them difficult to apply in practice.

With the limitation of the supervised approaches in mind, many researchers [12, 13, 14, 15] have resorted to unsupervised techniques, which do not require disparity ground truth to realize stereo matching. These approaches generally train networks by minimizing a hybrid loss, *e.g.*, combining a photometric loss and a smoothness loss [12, 13]. Some approaches also incorporate occlusion reasoning into the

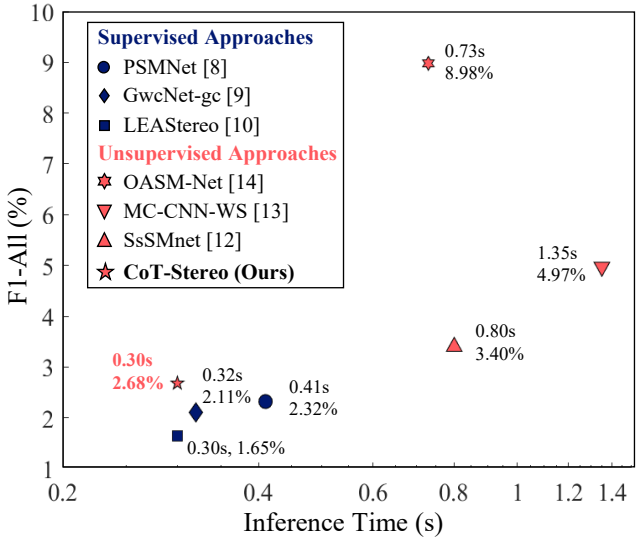

Fig. 1. Evaluation results on the KITTI Stereo 2015 benchmark [11], where "F1-All" denotes the percentage of erroneous pixels measured over all regions. Our CoT-Stereo outperforms all other state-of-the-art unsupervised stereo matching approaches in terms of both accuracy and speed.

This work was supported in part by the Collaborative Research Fund by Research Grants Council Hong Kong under Project C4063-18G, in part by the Department of Science and Technology of Guangdong Province Fund under Project GDST20EG54, and in part by the Zhongshan Municipal Science and Technology Bureau Fund under project ZSST21EG06, awarded to Prof. Ming Liu.

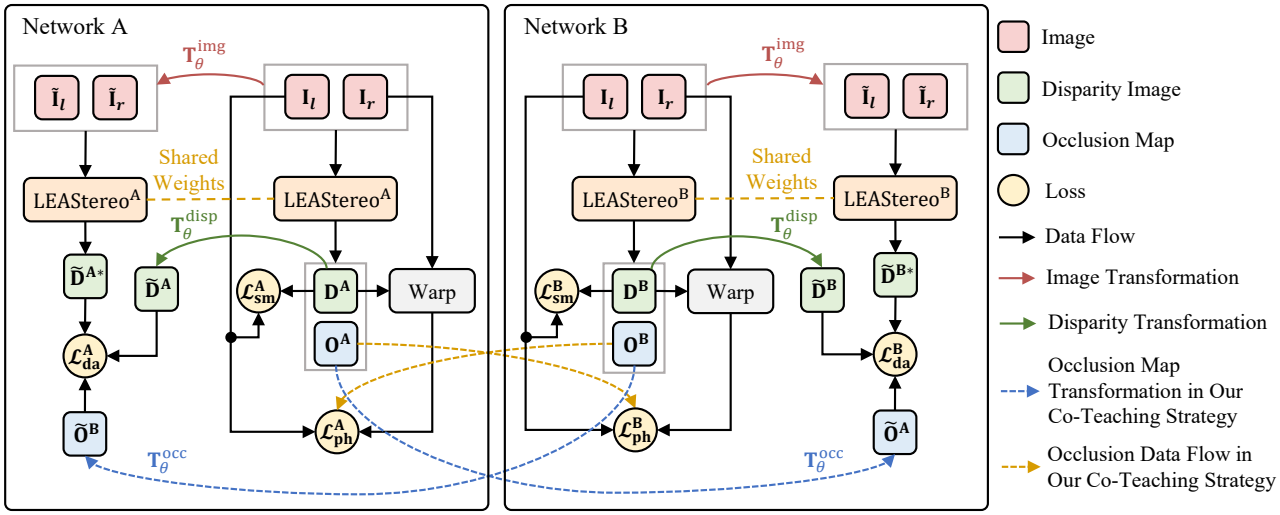

Fig. 2. An overview of our CoT-Stereo architecture, where two LEAStereo [10] networks with different initializations teach each other about the occlusions interactively.

training paradigm to further improve the stereo matching performance [14, 15]. However, such unsupervised approaches still perform unstably in some regions, especially near occlusions, because a single network can be sensitive to outliers when the disparity ground truth is inaccessible.

To address the instability problem, we propose CoT-Stereo, an unsupervised stereo matching approach. It outperforms all other state-of-the-art unsupervised stereo matching approaches in terms of both accuracy and speed on the KITTI Stereo benchmarks [16, 11], as illustrated in Fig. 1. Our CoT-Stereo employs a co-teaching framework, as shown in Fig. 2, where two networks (LEAStereo [10] is used as the backbone network) with different initializations interactively teach each other about the occlusions. Our previous work has adopted this co-teaching framework for unsupervised optical flow estimation [17], and in this paper, we employ this framework for unsupervised stereo matching. This framework can significantly improve model's robustness against outliers and further enhance the overall performance of unsupervised stereo matching.

2. METHODOLOGY

2.1. Preliminaries and Loss Functions

Given a pair of stereo images $\mathbf{I}_l$ and $\mathbf{I}_r$, the objective of stereo matching is to produce a dense disparity image $\mathbf{D}$. This can be achieved by an off-the-shelf stereo matching network, *e.g.*, LEAStereo [10]. An occlusion map $\mathbf{O}$ indicating each pixel's probability of belonging to the occluded regions can also be computed using the technique proposed in [18]. Now the problem becomes how to train the network without direct supervision from the disparity ground truth. Following the paradigm of unsupervised stereo matching, we employ a hybrid loss, which combines (a) a photometric loss $\mathcal{L}_{\text{ph}}$, (b) a smoothness loss $\mathcal{L}_{\text{sm}}$, and (c) a data-augmentation loss $\mathcal{L}_{\text{da}}$, to train our CoT-Stereo, as illustrated in Fig. 2. The photometric loss $\mathcal{L}_{\text{ph}}$ can be formulated as a combination of an SSIM term [19] and an L1 norm term:

$$\mathcal{L}_{\text{ph}}(\mathbf{I}_l, \mathbf{I}_r, \mathbf{D}, \mathbf{O}) = \frac{1}{\mathcal{N}} \sum_{\mathbf{p}} \Bigg(\alpha \frac{1 - \text{SSIM}\left(\mathbf{I}_l(\mathbf{p}), \widehat{\mathbf{I}}_l(\mathbf{p})\right)}{2} + (1-\alpha) \left\| \mathbf{I}_l(\mathbf{p}) - \widehat{\mathbf{I}}_l(\mathbf{p}) \right\|_1 \Bigg) \cdot \mathcal{S}\left(\overline{\mathbf{O}}(\mathbf{p})\right), \quad (1)$$

where $\widehat{\mathbf{I}}_l = \omega(\mathbf{I}_r, \mathbf{D})$ denotes the warped image from $\mathbf{I}_r$ based on $\mathbf{D}$; $\overline{\mathbf{O}}(\mathbf{p}) = 1 - \mathbf{O}(\mathbf{p})$; $\|\cdot\|_1$ denotes the L1 norm; $\mathcal{S}(\cdot)$ denotes the stop-gradient; and $\mathcal{N} = \sum_{\mathbf{p}} \mathcal{S}\left(\overline{\mathbf{O}}(\mathbf{p})\right)$ is a normalizer. Equation (1) shows that $\mathcal{L}_{\text{ph}}$ is an occlusion-aware loss used to penalize the photometric error. Following [14], we also adopt a smoothness loss $\mathcal{L}_{\text{sm}}$ to smooth the disparity estimations:

$$\mathcal{L}_{\text{sm}}(\mathbf{I}_l, \mathbf{D}) = \frac{1}{N_{\mathbf{p}}} \sum_{\mathbf{p}} \sum_{d \in \{x,y\}} |\nabla_d \mathbf{D}(\mathbf{p})| \, e^{-\|\nabla_d \mathbf{I}_l(\mathbf{p})\|_1}, \quad (2)$$

where $N_{\mathbf{p}}$ denotes the number of pixels. Moreover, inspired by [20], we adopt a data-augmentation scheme to enable networks to better handle occlusions. Specifically, we first perform transformations $\mathbf{T}_\theta^{\text{img}}$, $\mathbf{T}_\theta^{\text{disp}}$ and $\mathbf{T}_\theta^{\text{occ}}$ (*e.g.*, spatial, occlusion and appearance transformations [20]) on $(\mathbf{I}_l, \mathbf{I}_r)$, $\mathbf{D}$ and $\mathbf{O}$ respectively to obtain the augmented samples $\widetilde{\mathbf{I}}_l$, $\widetilde{\mathbf{I}}_r$, $\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{O}}$. Please note that, different from $\mathbf{O}$, a higher value in $\widetilde{\mathbf{O}}$ indicates that the pixel is less likely to be occluded in $\widetilde{\mathbf{D}}$ but more likely to be occluded in $\widetilde{\mathbf{D}}^*$. Given $\widetilde{\mathbf{I}}_l$ and $\widetilde{\mathbf{I}}_r$, we can also use LEAStereo [10] to get a disparity estimation $\widetilde{\mathbf{D}}^*$.

Algorithm 1: Co-Teaching Strategy

Input: Ω^{A} and Ω^{B}, learning rate η, constant threshold τ, epoch T_k and $T_{\max}$, iteration $N_{\max}$.
Output: Ω^{A} and Ω^{B}.

1 **for** $T = 1 \rightarrow T_{\max}$ **do**
2 **Shuffle** training set $\mathcal{D}$
3 **for** $N = 1 \rightarrow N_{\max}$ **do**
4 **Forward** individually to get $\mathbf{D}^i$, $\mathbf{O}^i$, $\widetilde{\mathbf{D}}^i$, $\widetilde{\mathbf{D}}^{i*}$ and $\widetilde{\mathbf{O}}^i$, $i \in \{\mathrm{A}, \mathrm{B}\}$
5 **Set** $\mathbf{O}^i \left(\mathbf{O}^i > \mathcal{R}(T)\right) = 1, i \in \{\mathrm{A}, \mathrm{B}\}$ ▷ Omit the pixels with high probability to be occluded
6 **Compute** $\mathcal{L}^{\mathrm{A}} = \mathcal{L}^{\mathrm{A}}_{\mathrm{ph}}(\mathbf{I}_l, \mathbf{I}_r, \mathbf{D}^{\mathrm{A}}, \mathbf{O}^{\mathrm{B}}) + \lambda_1 \cdot \mathcal{L}^{\mathrm{A}}_{\mathrm{sm}}(\mathbf{I}_l, \mathbf{D}^{\mathrm{A}}) + \lambda_2 \cdot \mathcal{L}^{\mathrm{A}}_{\mathrm{da}}(\widetilde{\mathbf{D}}^{\mathrm{A}}, \widetilde{\mathbf{D}}^{\mathrm{A}*}, \widetilde{\mathbf{O}}^{\mathrm{B}})$
7 **Compute** $\mathcal{L}^{\mathrm{B}} = \mathcal{L}^{\mathrm{B}}_{\mathrm{ph}}(\mathbf{I}_l, \mathbf{I}_r, \mathbf{D}^{\mathrm{B}}, \mathbf{O}^{\mathrm{A}}) + \lambda_1 \cdot \mathcal{L}^{\mathrm{B}}_{\mathrm{sm}}(\mathbf{I}_l, \mathbf{D}^{\mathrm{B}}) + \lambda_2 \cdot \mathcal{L}^{\mathrm{B}}_{\mathrm{da}}(\widetilde{\mathbf{D}}^{\mathrm{B}}, \widetilde{\mathbf{D}}^{\mathrm{B}*}, \widetilde{\mathbf{O}}^{\mathrm{A}})$
8 **Update** $\Omega^i = \Omega^i - \eta \nabla \mathcal{L}^i, i \in \{\mathrm{A}, \mathrm{B}\}$
9 **end**
10 **Update** $\mathcal{R}(T) = 1 - \tau \cdot \min\left\{\frac{T}{T_k}, 1\right\}$
11 **end**

Our data-augmentation loss $\mathcal{L}_{\mathrm{da}}$ is then defined as follows:

$$\mathcal{L}_{\mathrm{da}}(\widetilde{\mathbf{D}}, \widetilde{\mathbf{D}}^*, \widetilde{\mathbf{O}}) = \frac{\sum_{\mathbf{p}} l\left(\left|\mathcal{S}\left(\widetilde{\mathbf{D}}(\mathbf{p})\right) - \widetilde{\mathbf{D}}^*(\mathbf{p})\right|\right) \cdot \mathcal{S}\left(\widetilde{\mathbf{O}}(\mathbf{p})\right)}{\sum_{\mathbf{p}} \mathcal{S}\left(\widetilde{\mathbf{O}}(\mathbf{p})\right)},$$
$$l(x) = \begin{cases} x - 0.5, & x \geq 1 \\ x^2/2, & x < 1 \end{cases}, \tag{3}$$

where $l(\cdot)$ denotes the smooth L1 loss.

2.2. Co-Teaching Strategy

Fig. 2 and Algorithm 1 present the overview of our introduced co-teaching framework, where we simultaneously train two LEAStereo networks: (a) network A (with parameter Ω^{A}) and (b) network B (with parameter Ω^{B}). In each mini-batch, the two networks first forward individually to get several outputs (Line 4). Then, we use a dynamic threshold $\mathcal{R}(T)$ to omit the pixels with high occlusion probability (Line 5). $\mathcal{R}(T)$ is designed based on the network memorization mechanism. Specifically, during training, the networks will first learn stereo matching from clear patterns, and then will be gradually affected by outliers [21]. Therefore, $\mathcal{R}(T)$ is initialized as 1 and it decreases gradually as the epochs increase. This helps the networks avoid memorizing outliers (possible inaccurate occlusion estimations) and further improves the performance of unsupervised stereo matching.

Afterwards, we let the two networks swap their estimated occlusion maps and compute their loss functions (Line 6 and 7). Since different networks can learn different types of occlusion and disparity estimations, swapping the occlusion estimations enables the two networks to adaptively correct the inaccurate occlusion estimations, which can further improve the performance of unsupervised stereo matching. Please note that since deep neural networks are highly non-convex, we use two LEAStereo [10] networks with different initializations in our CoT-Stereo. Finally, we update both the parameters of these two networks as well as the dynamic threshold $\mathcal{R}(T)$ (Line 8 and 10).

3. EXPERIMENTAL RESULTS

3.1. Datasets and Implementation Details

For the implementation, we set $\alpha = 0.85$ in Equation (1). In addition, we set $T_k = 0.2 \cdot T_{\max}$ and $\tau = 0.7$ in Algorithm 1. Moreover, we adopt the Adam optimizer and use a learning rate $\eta = 10^{-4}$ with an exponential decay scheme. Since the two networks present similar performance after convergence, we simply adopt network A for performance evaluation.

We use three public datasets, (a) the Scene Flow [22], (b) the KITTI Stereo 2012 [16], and (c) the KITTI Stereo 2015 [11] datasets, to validate the effectiveness of our CoT-Stereo. The Scene Flow dataset [22] is collected in three different synthetic scenes, while the two KITTI Stereo datasets [16, 11] are collected in real-world driving scenarios and have public benchmarks. Two evaluation metrics, (a) the average end-point error (AEPE) that measures the difference between the disparity estimations and ground-truth labels and (b) the percentage of bad pixels (tolerance: 3 pixels) (F1) [16, 11], are adopted for accuracy comparison.

In our experiments, we first conduct ablation studies on the Scene Flow dataset [22] to demonstrate the effectiveness of our adopted loss functions and proposed co-teaching strategy, as illustrated in Section 3.2. Then, we evaluate our CoT-Stereo on the two KITTI Stereo benchmarks [16, 11], as presented in Section 3.3.

3.2. Ablation Study

Table 1 presents the evaluation results of our CoT-Stereo with different setups on the Scene Flow dataset [22]. For our proposed co-teaching strategy, (a)–(c) and (g) of Table 1 demon-

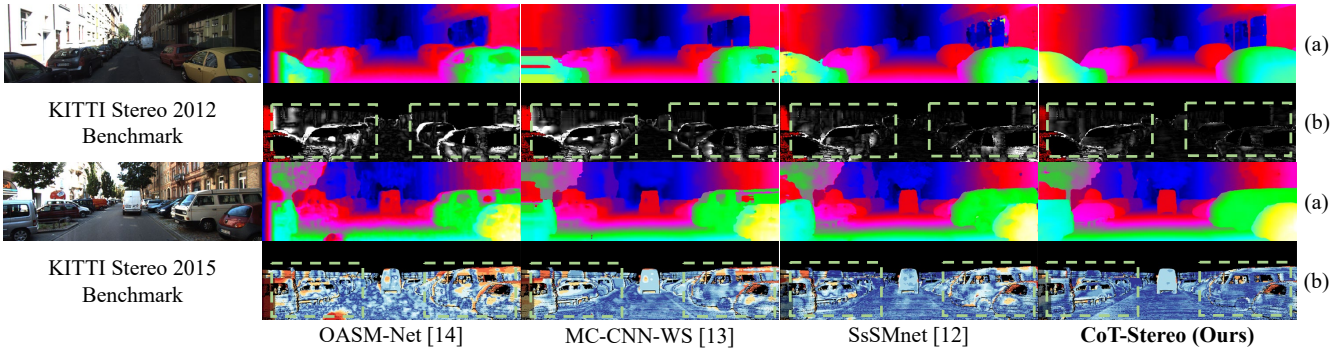

Fig. 3. Examples on the KITTI Stereo benchmarks [16, 11], where rows (a) and (b) show the disparity estimations and the corresponding error maps, respectively. Significantly improved regions are marked with green dashed boxes.

Table 1. Evaluation results of our CoT-Stereo with different setups on the Scene Flow dataset [22]. "Swap" and "DT" denote the occlusion estimation swapping operation and the dynamic threshold selection scheme, respectively. The adopted setup (the best result) is shown in bold type.

No.	Swap	DT	$\mathcal{L}_{\text{ph}}$	$\mathcal{L}_{\text{sm}}$	$\mathcal{L}_{\text{da}}$	AEPE (px)
(a)	–	–	✓	✓	✓	3.68
(b)	✓	–	✓	✓	✓	2.35
(c)	–	✓	✓	✓	✓	3.10
(d)	✓	✓	✓	–	–	4.72
(e)	✓	✓	✓	✓	–	3.97
(f)	✓	✓	✓	–	✓	1.86
(g)	✓	✓	✓	✓	✓	**1.31**

strate the effectiveness of the occlusion estimation swapping operation and the dynamic threshold selection scheme, which can effectively improve unsupervised stereo matching. Additionally, we can clearly observe that the combination of the three loss functions can effectively improve the performance, as shown in (d)–(g) of Table 1. Moreover, (g) in Table 1 denotes the adopted setup, which validates the effectiveness of our adopted loss functions and proposed co-teaching strategy.

3.3. Evaluations on the Public Benchmarks

Table 2 shows the online leaderboards of the KITTI Stereo 2012 [16] and Stereo 2015 [11] benchmarks, and Fig. 1 visualizes the results on the KITTI Stereo 2015 benchmark. We can observe that our CoT-Stereo outperforms all other state-of-the-art unsupervised stereo matching approaches in terms of both accuracy and speed, which demonstrates the effectiveness of the occlusion estimation swapping operation and the dynamic threshold selection scheme for unsupervised stereo matching. Excitingly, our CoT-Stereo can even present competitive performance compared with the state-of-the-art supervised approaches. Examples on the KITTI Stereo benchmarks are shown in Fig. 3, where it is evident that our CoT-Stereo can generate more robust and accurate disparity estimations. All the analysis proves the excellent performance of our CoT-Stereo for unsupervised stereo matching.

Table 2. Evaluation results (%) on the KITTI Stereo 2012[1] [16] and Stereo 2015[2] [11] benchmarks. "S" denotes supervised approaches. "Noc" and "All" represent the F1 for non-occluded pixels and all pixels, respectively [16, 11]. Best results for supervised and unsupervised approaches are both shown in bold type.

Approach	S	KITTI 2012		KITTI 2015	
		Noc	All	Noc	All
PSMNet [8]	✓	1.49	1.89	2.14	2.32
GwcNet-gc [9]	✓	1.32	1.70	1.92	2.11
LEAStereo [10]	✓	**1.13**	**1.45**	**1.51**	**1.65**
OASM-Net [14]	–	6.39	8.60	7.39	8.98
Flow2Stereo [15]	–	4.58	5.11	6.29	6.61
MC-CNN-WS [13]	–	3.02	4.45	4.11	4.97
SsSMnet [12]	–	2.30	3.00	3.06	3.40
CoT-Stereo (Ours)	–	**1.82**	**2.32**	**2.43**	**2.68**

4. CONCLUSION AND FUTURE WORK

This paper proposed a novel co-teaching strategy for unsupervised stereo matching, which consists of a dynamic threshold selection scheme and an occlusion estimation swapping operation. The former ensures that the networks do not memorize possible outliers, while the latter enables the two networks to adaptively correct the inaccurate occlusion estimations and further improve the performance of unsupervised stereo matching. Extensive experimental results on the KITTI Stereo benchmarks showed that our approach, CoT-Stereo, outperforms all other state-of-the-art unsupervised stereo matching approaches in terms of both accuracy and speed.

[1] http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo

[2] http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo

5. REFERENCES

[1] Hengli Wang, Yuxiang Sun, and Ming Liu, "Self-supervised drivable area and road anomaly segmentation using rgb-d data for robotic wheelchairs," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4386–4393, 2019.

[2] Hengli Wang, Rui Fan, Yuxiang Sun, and Ming Liu, "Applying surface normal information in drivable area and road anomaly detection for ground mobile robots," in *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2020.

[3] Hengli Wang, Rui Fan, Yuxiang Sun, and Ming Liu, "Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms," *IEEE Trans. Cybern.*, 2021.

[4] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vision (ECCV)*. Springer, 2020, pp. 340–356.

[5] Rui Fan, Hengli Wang, Peide Cai, Jin Wu, Junaid Bocus, Lei Qiao, and Ming Liu, "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE/ASME Trans. Mechatronics*, 2021.

[6] Rui Fan, Xiao Ai, and Naim Dahnoun, "Road surface 3d reconstruction based on dense subpixel disparity map estimation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3025–3035, 2018.

[7] Hengli Wang, Rui Fan, Peide Cai, and Ming Liu, "PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4353–4360, 2021.

[8] Jia-Ren Chang and Yong-Sheng Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 5410–5418.

[9] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li, "Group-wise correlation stereo network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2019, pp. 3273–3282.

[10] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge, "Hierarchical neural architecture search for deep stereo matching," in *Adv. Neural Inf. Process. Syst. (NIPS)*, 2020.

[11] Moritz Menze and Andreas Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2015, pp. 3061–3070.

[12] Yiran Zhong, Yuchao Dai, and Hongdong Li, "Self-supervised learning for stereo matching with self-improving ability," *CoRR*, 2017.

[13] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret, "Weakly supervised learning of deep metrics for stereo reconstruction," in *Proc. IEEE Inter. Conf. Comput. Vision (ICCV)*, 2017, pp. 1339–1348.

[14] Ang Li and Zejian Yuan, "Occlusion aware stereo matching via cooperative unsupervised learning," in *ACCV*. Springer, 2018, pp. 197–213.

[15] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu, "Flow2stereo: Effective self-supervised learning of optical flow and stereo matching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 6648–6657.

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*. IEEE, 2012, pp. 3354–3361.

[17] Hengli Wang, Rui Fan, and Ming Liu, "CoT-AMFlow: Adaptive modulation network with co-teaching strategy for unsupervised optical flow estimation," in *Conf. Robot Learn. (CoRL)*, 2020.

[18] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu, "Occlusion aware unsupervised learning of optical flow," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 4884–4893.

[19] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[20] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang, "Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 6489–6498.

[21] Devansh Arpit, Stanislaw K Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron C Courville, Yoshua Bengio, et al., "A closer look at memorization in deep networks," in *ICML*, 2017.

[22] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2016, pp. 4040–4048.