UNDERWATER STEREO MATCHING VIA UNSUPERVISED APPEARANCE AND FEATURE ADAPTATION NETWORKS

Wei Zhong *Yazhi Yuan* *Xinchen Ye** *Dian Zheng* *Rui Xu*

Dalian University of Technology, Liaoning, China

ABSTRACT

Stereo matching has been widely used to estimate depth maps in terrestrial environments. However, it is difficult to achieve appealing performance in underwater environments, since adequate underwater stereo data with groundtruth depth information is not easily available for training an underwater depth estimation model. In addition, the domain gap also leads to the failure of directly applying existing models of terrestrial scenes to underwater scenes. Therefore, this paper proposes a novel underwater depth estimation network which can infer depth maps from real underwater stereo images in an unsupervised adaptation manner. The proposed learning pipeline contains style adaptation (SA) in appearance space and feature adaptation (FA) in semantic space to progressively adapt the depth estimation models to underwater domain. Experimental results show that by integrating the proposed adaptation modules into the off-the-shelf stereo matching backbones, our method achieves a superior performance of underwater depth estimation compared to other state-of-the-art methods.

Index Terms— Stereo matching, domain adaptation, style translation, Underwater

1. INTRODUCTION

As underwater environment perception plays an important role in ocean exploration, underwater vision techniques have been put forward for higher requirements. Among them, efficient depth estimation is helpful for underwater close-range perception, which can offer distance information to benefit a variety of underwater computer vision tasks such as underwater detection [1], localization [2], navigation [3] and so on. However, compared to the depth perception techniques, e.g, stereo matching [4], in terrestrial environments, existing deep learning based methods encounter many bottlenecks due to the weak illumination conditions and geometrical distortion in underwater environments. To be specific, it is extremely impractical to obtain enough amounts of underwater paired RGB and depth data as supervision to train an underwater depth estimation network due to the inevitable high-cost for specialized depth sensing equipment in underwater environments. Even if underwater RGB images can be captured by color cameras, degraded imaging quality, e.g., color change and blur, are also obstacles to accurate depth estimation. Besides, the domain gap between data captured from different scenes also leads to the failure of directly applying existing terrestrial depth estimation models [4, 5, 6, 7] to the underwater domain, impeding the performance of underwater depth estimation to go a step further.

To cope with the limitation of training data in underwater depth estimation, the work [8] uses color transfer techniques to synthesize underwater data and trains an extra depth estimation network based on the synthesized data with underwater style and their corresponding depth maps under the stacked mode. However, this method is constructed based on the mode of single-view depth estimation, which is a highly ill-posed problem due to the inherent ambiguity and the lack of geometry constraints when mapping monocular RGB measurement into depth. Another promising avenue is the stereo-based unsupervised learning technique [9, 10], which takes stereo pairs as input and leverages the photometric consistency between stereo pairs to assist the training. A recent work [11] follows this pipeline to address the underwater depth estimation, and achieves relatively good results. However, they just simply use the captured underwater stereo data to train an off-the-shelf unsupervised learning backbone without any consideration of the essential degradation characteristics of underwater data, resulting in unsatisfactory performance.

To address the above problems, we continue to research stereo-based depth estimation. As shown in Fig. 1, our core idea is to make full use of the labeled terrestrial stereo dataset (e.g., KITTI [12]) and unlabeled real underwater stereo dataset (USD) [13], and infer depth maps by gradually adapting a stereo matching network (SMN) trained on a terrestrial dataset to the underwater domain on the premise of fully considering the degradation characteristics of underwater scenes. The proposed learning pipeline mainly contains style adaptation (SA) in appearance space and feature adaptation (FA) in semantic space to resolve domain discrepancy in an unsupervised adaptation manner. Specifically, in SA, we construct an unsupervised stereo translation network (USTN), which aims at synthesizing stereo images with un-

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61906029, 61976037 and 62027826. * Corresponding author. E-mail: yexch@dlut.edu.cn

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

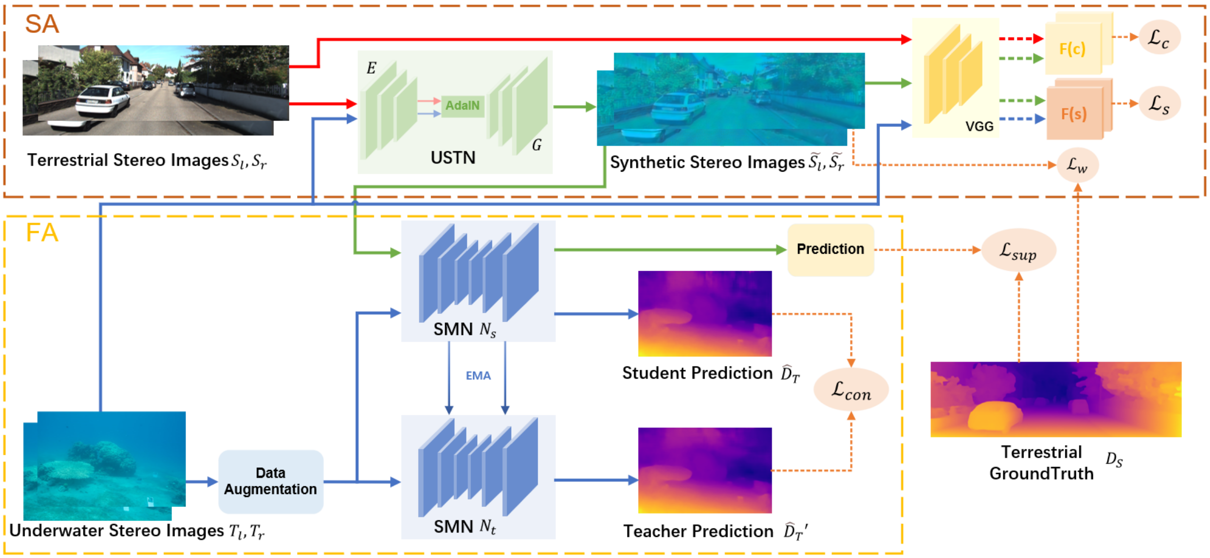
SA
Terrestrial Stereo Images S_l, S_r
E
AdaIN
USTN
G
Synthetic Stereo Images $\widetilde{S}_l, \widetilde{S}_r$
VGG
F(c)
F(s)
$\mathcal{L}_c$
$\mathcal{L}_s$
$\mathcal{L}_w$
FA
Prediction
$\mathcal{L}_{sup}$
SMN N_s
Student Prediction $\widehat{D}_T$
EMA
$\mathcal{L}_{con}$
Terrestrial GroundTruth D_S
Data Augmentation
SMN N_t
Underwater Stereo Images T_l, T_r
Teacher Prediction $\widehat{D}_T{}'$

Fig. 1. Illustration of the proposed framework, which contains style adaptation (SA) and feature adaptation (FA) modules. At test phase, the student SMN model N_s is used to predict depth maps from real underwater stereo images.

derwater style from the terrestrial dataset, i.e., combining scene structure of terrestrial images and style of underwater images in the appearance space, to benefit the effective training of SMN. Besides, we introduce a warping regularization to synthesize geometry-consistent stereo images with underwater style, which strengthens the left-right consistency between both views. Then, considering the weak generalization to real underwater images when only trained on synthetic data, an FA module is needed to address this problem. Due to the large domain discrepancy of semantic content between terrestrial and underwater scenes, aligning features via global adversarial approaches [8] may trigger a negative transfer, which enforces the alignment in an incorrect semantic category, leading to even worse performance than a network trained solely on the terrestrial domain. Instead, we present a teacher-student self-ensembling technique [14] for feature adaptation, which can incrementally minimize the domain discrepancy by compelling the student to produce consistent predictions provided by the teacher on target underwater data.

Experimental results show that our method successfully achieves a superior performance of underwater depth estimation compared to other state-of-the-art methods [8, 11]. To sum up, the contributions in this paper are listed as follows: 1) We propose a novel underwater stereo matching network, which can infer depth maps from real underwater stereo images in an unsupervised adaptation manner. 2) Our SA can generate underwater-like geometry-consistent synthetic stereo data to address the scarcity of training data. 3) We adopt the self-ensembling technique to realize the feature adaptation, thus reducing the large domain discrepancy of semantic content between synthetic data and real underwater data.

2. PROPOSED METHOD

Fig. 1 shows our whole framework, which contains style adaptation (SA) and feature adaptation (FA) modules. Given a collection of stereo RGB images with groundtruth disparity maps $\{S_l^i, S_r^i, D_S^i\}_{i=1}^N$ in the source terrestrial domain, and underwater stereo images $\{T_l^i, T_r^i\}_{i=1}^M$ in the target underwater domain, where M, N is the number of training data, our goal is to estimate the disparity maps $\{D_T^i\}_{i=1}^M$ for each underwater image pair. Specifically, $\{S_l, S_r\}$ are fed into our unsupervised stereo translation network (USTN) to generate the corresponding synthetic stereo images $\{\widetilde{S}_l, \widetilde{S}_r\}$. Then, $\{\widetilde{S}_l, \widetilde{S}_r\}$ together with underwater stereo images $\{T_l, T_r\}$ are sent into the proposed a teacher-student self-ensembling module to minimize the domain discrepancy between synthetic data and real data. At test phase, the student SMN model N_s is used to predict the disparity map $\widehat{D}_T$ from real underwater stereo images $\{T_l, T_r\}$.

2.1. Style Adaptation

Different from GAN-based approaches [8] that need to train a specific style adaptation model for each given underwater scene, we follow the Adaptive Instance Normalization (AdaIN) [15] approach and integrate with perceptual constraint to achieve content preservation and style transfer simultaneously. We first separately send each of the stereo images $\{S_l, S_r\}$ to the encoder E to obtain the corresponding features $\{E(S_l), E(S_r)\}$. Meanwhile, we choose one of the underwater stereo images as the target image (e.g., T_l), and feed it into E to obtain the feature $E(T_l)$. Then, we can obtain the stylized features by recombining the features of

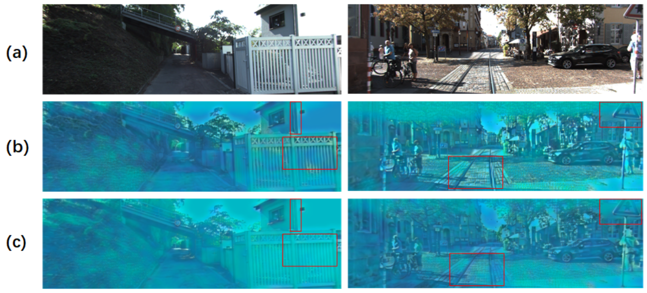

Fig. 2. Example underwater-style synthesized images. (a) Terrestrial images[12]; (b) Synthetic images from USTN w/o warp; (c)Synthetic images from USTN.

both domains from the encoder E:

$$\widetilde{S}'_l = \sigma(E(T_l))(\frac{E(S_l) - \mu(E(S_l))}{\sigma(E(S_l))}) + \mu(E(T_l)). \quad (1)$$

where $\widetilde{S}'_l$ is the stylized feature of left image[1]. μ, σ are mean and standard deviation of images features. Finally, the adjusted $\widetilde{S}'_l$ is sent to a module G to generate the output $\widetilde{S}_l = G(\widetilde{S}'_l)$. To retain the scene structure of generated images while aligning style representation to real underwater images, the perceptual losses [16] containing content loss L_c and style loss L_s are defined respectively as follows:

$$\begin{aligned} L_c &= \sum_{l \in l_c} w_l \left\| F(\widetilde{S}_l) - F(S_l) \right\|_2^2, \\ L_s &= \sum_{l \in l_s} w_l \left\| \mathbb{G}(\widetilde{S}_l) - \mathbb{G}(T_l) \right\|_2^2, \mathbb{G} = \phi_{h,w}\phi_{h,w}^T. \end{aligned} \quad (2)$$

where $F(.), \phi$ are features and vector representation from the pretrained encoder VGG-19 [17]. l_c, l_s are the number of content features and Gram matrix $\mathbb{G}$ layers considered in the losses. w_l is the weight of the l-th layer. $\mathbb{G}$ is a quantitative description of the image style more comprehensively.

Note that the above process generates synthetic left image and right image separately, which may cause distortion that destroys the geometry relationship between both views. Therefore, we introduce a warping loss L_w to synthesize geometry-consistent stereo images with underwater style:

$$L_w = \left\| W(\widetilde{S}_r, D_S) - \widetilde{S}_l \right\|. \quad (3)$$

where $W(.)$ is the warp operation [18] to reconstruct the left view from the synthetic right view and disparity map. Accordingly, the final loss for training USTN is formulated as:

$$L_{SA} = \alpha_1 L_c + \alpha_2 L_s + \alpha_3 L_w, \quad (4)$$

where α_1, α_2, α_3 are adjustment parameters.

2.2. Feature Adaptation

The synthetic data can not necessarily inherit the characteristics of the underwater environment completely, so that we are motivated to further improve the performance of underwater stereo matching not only in appearance level but also perform feature domain adaptation from synthetic to real underwater domains. We introduce the self-ensembling [14] domain adaptation approach for training the depth estimation networks, which consists of training a student network N_s and a teacher network N_t.

First, the student network N_s is trained supervisedly using synthetic underwater images with labels which is optimized by an L_1 loss defined as $L_{sup} = ||\widehat{D}_S - D_S||_1$. Then the parameter transferring between N_s and N_t adopts exponential moving average (EMA) [19] method, that is, the parameters of student network are multiplied by a certain attenuation rate and added into the parameters of teacher network as a part of teacher network, and participate in the update of teacher network in each iteration process. The parameters Ω_t^i of N_t at training step i are updated by the student's parameters Ω_s^i:

$$\Omega_t^i = \lambda \Omega_t^{i-1} + (1 - \lambda)\Omega_s^i. \quad (5)$$

where λ is an adjustment parameter.

To adapt to the real underwater domain, each input underwater image is augmented stochastically. Then, the predicted disparity maps $\widehat{D}_T$, $\widehat{D}'_T$ from N_s and N_t respectively are used to formulate the self-ensembling consistency loss L_{con} as:

$$L_{con} = ||\widehat{D}_T - \widehat{D}'_T||_1. \quad (6)$$

The output from teacher N_t can be regarded as a pseudo label to supervise the student N_s. To sum up, the total loss of feature adaptation is formulated as:

$$L_{FA} = \beta_1 L_{sup} + \beta_2 L_{con}. \quad (7)$$

where β_1, β_2 are the adjustment parameters.

3. EXPERIMENTS

We conduct thorough experiments of our proposed method on KITTI[12] (source domain) and USD[13] (target domain). KITTI[12] is a real-world terrestrial dataset with 194 stereo images with ground-truth disparity maps. USD[13] collected 57 underwater stereo pairs with ground-truth disparity maps from four different sites (Katzaa, Mikhmoret, Nachsholim and Satil) with different types of characteristic attributes. We adopt [15] as the backbone of our USTN and GwcNet[5] as the backbone of SMN. The metrics EPE and D1 [5] are used to evaluate the results. For training, we first train USTN to generate the synthetic underwater stereo images. Then, for depth estimation, we initialize the SMN backbone with the released terrestrial model and train the student with synthetic

[1] The right stylized feature $\widetilde{S}'_r$ can be obtained by the similar function.

Table 1. Ablation on style adaptation based on the backbone Gwc-Net. The three data lines respectively show the performance of underwater depth estimation under different settings. K, M, N, S represent four datasets from USD.

Model$_T$	EPE$_K$	D1$_K$	EPE$_M$	D1$_M$	EPE$_N$	D1$_N$	EPE$_S$	D1$_S$
Gwc-Net[5]	8.2064	**0.6545**	15.7979	0.6567	13.1062	0.5673	21.7819	0.7250
USTN w/o warp	7.5232	0.6747	8.6108	0.5834	6.8688	**0.4897**	16.2889	0.6723
USTN	**7.0659**	0.6804	**8.3918**	**0.5661**	**6.6404**	0.5173	**12.8664**	**0.6582**

Table 2. Ablation on feature adaptation. "SA" and "FA" stand for style adaptation and feature adaptation respectively.

Backbone	Adaptation	EPE$_K$	D1$_K$	EPE$_M$	D1$_M$	EPE$_N$	D1$_N$	EPE$_S$	D1$_S$
Gwc-net[5]	SA	7.0659	**0.6804**	8.3918	**0.5661**	6.6404	0.5173	12.8664	0.6582
Gwc-net[5]	SA+FA	**6.6597**	0.6932	**8.1247**	0.5789	**6.3310**	**0.4589**	**10.1833**	**0.6133**

Table 3. Performance comparison. The results with lower EPE and D1 values are better.

Method	EPE$_K$	D1$_K$	EPE$_M$	D1$_M$	EPE$_N$	D1$_N$	EPE$_S$	D1$_S$
MU-Net[8]	7.0974	**0.6794**	9.4024	**0.5588**	7.9691	0.4948	14.5912	0.6460
SU-Net[11]	12.5766	0.7780	26.9227	0.8939	19.7091	0.8981	12.6444	0.8600
Ours	**6.6597**	0.6932	**8.1247**	0.5789	**6.3310**	**0.4589**	**10.1833**	**0.6133**

data by L_{sup} and simultaneously update the teacher model. Then, we further optimize N_s by L_{FA} with real data for 300 epochs. The learning rate is set to 0.0001 and is down-scaled by 10 after epoch 200. The adjustment parameters are set as $\lambda = 0.99$, $\alpha_1 = 5.0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.1$, $\beta_1 = 1$, $\beta_2 = 0.5$. Throughout the training period, we do not use the ground-truths from USD[13], but just use them to evaluate our method at test phase.

3.1. Ablation Study

Effectiveness of SA.

Based on USTN, we can generate synthetic images with underwater-style presenting in Fig. 2 (b). We can also observe that the regularization effect of warp loss on the image style transfer and geometry-consistent of the stereo image in Fig. 2 (c). It can be seen intuitively that warp loss is effective in maintaining and restoring the edges of objects (eg. vehicles, roads) and repeated textures (eg. fences).

The verification conclusions can be obtained through the performance of stereo matching networks. As shown in Table 1, the performance of the network model obtained from training on data without SA is relatively worse than the model obtained from supervised training using synthetic data. And on the basis of perceptual features, the model trained with integrating warp loss can bring a progressive performance improvement not only generate a better USTN model that retains more correct structural information, but also achieve a further reduction of stereo matching prediction error.

Effectiveness of FA. We also analyze the effect of self-ensembling domain adaptation method for stereo matching. Table 2 reveals the benefit of applying self-ensembling training strategy to the existing stereo matching in different underwater scenarios.

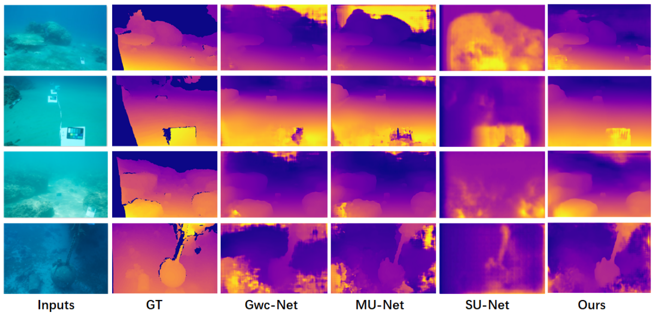

Fig. 3. Prediction results of USD[13] using different training models in the four underwater datasets.

3.2. Performance Comparisons

As shown in Table 3, we compare our proposed method with a monocular unsupervised network(MU-Net) [8] and an stereo unsupervised network(SU-Net) [11] for underwater depth estimation. We can observe that in all four different underwater scenarios, our method is superior to the previous monocular or stereo underwater depth estimation methods. Fig. 3 show the disparity maps from different methods. It's pretty intuitive that our method can get the closest result to GT. Among them, compared with the unsupervised method, the improvement of our method is particularly large thanks to our style adaptation and feature adaptation based on stereo vision.

4. CONCLUSION

We proposed the effective stereo adaptation networks for underwater stereo matching to effectively solve the problem of domain shift and improve the performance of stereo matching network on the real underwater stereo images. Experimental results demonstrate the effectiveness of the proposed method.

5. REFERENCES

[1] Pep Lluis Negre, Francisco Bonin-Font, and Gabriel Oliver, "Cluster-based loop closing detection for underwater slam in feature-poor regions," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2589–2595.

[2] Chee Sheng Tan, Rosmiwati Mohd-Mokhtar, and Mohd Rizal Arshad, "Fast fourier transform overlap approach for underwater acoustic positioning system," in *2018 IEEE 8th International Conference on Underwater System Technology: Theory and Applications (USYS)*, 2018, pp. 1–5.

[3] A. Yu. Rodionov, F. S. Dubrovin, P. P. Unru, and S. Yu. Kulik, "Experimental research of distance estimation accuracy using underwater acoustic modems to provide navigation of underwater objects," in *2017 24th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS)*, 2017, pp. 1–4.

[4] Jia-Ren Chang and Yong-Sheng Chen, "Pyramid stereo matching network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li, "Group-wise correlation stereo network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2495–2504.

[7] Haofei Xu and Juyong Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1959–1968.

[8] Xinchen Ye, Zheng Li, Baoli Sun, Zhihui Wang, Rui Xu, Haojie Li, and Xin Fan, "Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks," *IEEE Transactions on Circuits and Systems for Video Technology(TCSVT)*, vol. 30, no. 11, pp. 3995–4008, 2019.

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6602–6611.

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow, "Digging into self-supervised monocular depth prediction," in *The International Conference on Computer Vision (ICCV)*, October 2019.

[11] Katherine A. Skinner, Junming Zhang, Elizabeth A. Olson, and Matthew Johnson-Roberson, "Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 7947–7954.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving The kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[14] Geoff French, Michal Mackiewicz, and Mark Fisher, "Self-ensembling for visual domain adaptation," in *International Conference on Learning Representations (ICLR)*, 2018.

[15] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1510–1519.

[16] Justin Johnson, Alexandre Alahi, editor="Leibe Bastian Fei-Fei, Li", Jiri Matas, Nicu Sebe, and Max Welling, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*, 2016.

[17] K. Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2017.

[18] Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, and Hongsheng Li, "Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020, pp. 12757–12766.

[19] Jaehoon Choi, Taekyung Kim, and Changick Kim, "Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6830–6840.