CROSS-MODALITY DEPTH ESTIMATION VIA UNSUPERVISED STEREO RGB-TO-INFRARED TRANSLATION

Shi Tang[a] *Xinchen Ye*[b*] *Fei Xue*[b] *Rui Xu*[b]

[a]School of Software, Tsinghua University, Beijing
[b]DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Dalian

ABSTRACT

Existing depth estimation methods infer scene depth only from stereo visible light (RGB) images. Since RGB imaging is sensitive to changes in light, it's difficult to estimate depth information accurately in some degraded visibility conditions. In contrast, infrared (IR) imaging captures thermal radiation and is not affected by brightness changing, providing extra clues for depth estimation. However, most datasets used for training in depth estimation do not have IR images paired with RGB-D data. Therefore, how to obtain the paired IR images and exploit the respective advantages of RGB and IR images to improve the performance of depth estimation, is of vital importance. Our core idea is to first develop an unsupervised RGB-to-IR translation (RIT) network with proposed Fourier domain adaptation and multi-space warping regularization to synthesize stereo IR images from their corresponding stereo RGB images. And then modified depth estimation backbones can be used as the cross-modality depth estimation (CDE) network to infer disparity maps from cross-modal RGB-IR stereo pairs. Assisted by the synthetic stereo IR images, we obtain superior performance just by flexibly deploying our framework to several off-the-shelf depth estimation backbones of single-modality (RGB) based methods.

Index Terms— Depth estimation, Stereo, Cross-modality, Infrared, Image translation

1. INTRODUCTION

Depth estimation is a fundamental task for many computer vision applications [1, 2, 3]. With the development of CNN, many algorithms step further on improving the performance of depth estimation. A mainstream way is to estimate scene depth from stereo images. According to different ways of training, the supervised methods [4, 5, 6] use true disparity maps to guide the training directly, while the unsupervised ones [7, 8, 9] reformulate depth estimation into an image reconstruction problem which leverages the stereo relationship as supervision to assist the training.

In general, most existing methods [4, 5, 6, 10, 11, 7, 8, 9, 12, 13] use only visible light (RGB) images for estimation. Although RGB images have rich textures and high contrast, the imaging is very sensitive to changes in light, resulting in the difficulty to estimate accurately in degraded visibility conditions. On the contrary, infrared (IR) imaging captures thermal radiation of objects with less textures and low contrast, but has its own advantages against RGB images, e.g., invariance to light changes, higher penetration on foggy/rainy/snowy weather and further detection distance, which can provide more significant clues for depth estimation. However, limited by the characteristics of low contrast and less textures, methods based on infrared images [14, 15, 16] cannot estimate depth with sufficient accuracy. Thus, an RGB-IR cross-modal fashion of depth estimation is urgently needed for estimating in a more robust way. However, a huge obstacle lies in the acquisition of IR images. Some commonly used datasets like KITTI [17] do not have paired stereo IR images together with RGB-D data. Although there are some hardware solutions on the market, additional IR sensors are usually required. And comes also the challenges of image collection and registration. Therefore, how to obtain the paired stereo IR images effectively becomes the primary problem to be solved in this work.

Image translation [18] is a feasible way to solve this problem by generating IR images from the corresponding RGB images considering the strong structural similarity between RGB-IR pairs. However, our difficulties lie in: 1) Due to the absence of paired RGB-IR data in KITTI, it is hard to accurately express the intrinsic characteristics of IR images as regularization and preserve the scene semantic consistency for efficiently training the translation network in an unsupervised way. 2) We have to preserve the stereo consistency between left and right views when generating stereo IR images for binocular depth estimation methods. Existing single-view translation methods [18, 19] often fail to produce geometry-consistent textures across both the generated views. In this paper, we successfully address these two challenges by properly designing the unsupervised learning scheme and stereo constraint.

Specifically, as shown in Fig. 1, we propose a framework consisting of an RGB-to-IR translation (RIT) network to synthesize stereo IR images from their corresponding stereo RGB images and a cross-modality depth estimation (CDE) network to estimate disparities from the cross-modal RGB-IR stereo pairs. Our core idea is to learn the single-view RGB-to-IR relationship from an available paired RGB-IR dataset[1], and then generalize/adapt it to the target KITTI stereo dataset. Therefore, we propose an unsupervised Fourier domain adaptation strategy to facilitate the translation task, which adopts self-ensembling to align features extracted from the source and target data, transferring the RGB-to-IR relationship to the target domain. Since IR images have the properties of low contrast and containing many smooth regions, we reformulate them with frequency statistics, and apply domain alignment in the Fourier domain. For stereo constraint, we introduce a multi-space warping regularization to generate geometry-consistent stereo IR images, which strengthens the left-right consistency between both views from the output, feature and affinity spaces. Our main contributions are listed as follows:

- We break away the shackles of general paradigms and firstly introduce a novel cross-modality depth estimation scheme,

*Corresponding author, email address: yexch@dlut.edu.cn.

[1]Some public datasets with single-view paired RGB-IR images (but without paired depth data), e.g., INO [20], can be easily obtained.

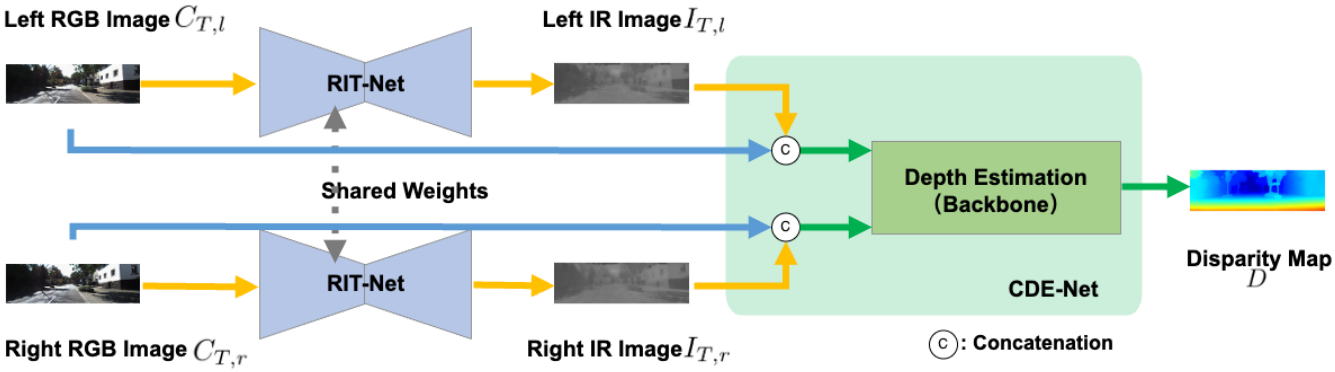

Fig. 1. Network Overview. It consists of an RGB-to-IR translation (RIT) network to synthesize stereo IR images from their corresponding stereo RGB images and a cross-modality depth estimation (CDE) network to infer disparities from the cross-modal RGB-IR stereo pairs.

which can both overcome the weakness of single-modality based methods and keep the low requirement on the training dataset.

- We propose a Fourier domain adaptation and a multi-space warping regularization to make the unsupervised stereo translation be feasible.
- Assisted by the synthetic stereo IR images, we obtain superior performance on KITTI compared with existing single-modality (RGB) based methods. Our scheme can be flexibly deployed to most off-the-shelf depth estimation backbones, including unsupervised [13, 21] and supervised methods [11], to improve their performance.

2. METHOD

As shown in Fig. 1, given a collection of stereo RGB images $\{C_{T,l}^i, C_{T,r}^i\}_{i=1}^N$ in the target domain (KITTI dataset [17]), and single-view RGB-IR pairs $\{C_S^i, I_S^i\}_{i=1}^N$ in the source domain (INO dataset [20]), where N is the number of training data, our goal is to estimate the disparity maps $\{D^i\}_{i=1}^N$ for each image pair. Specifically, $C_{T,l}$, $C_{T,r}$ are separately fed into RIT to generate the corresponding left and right IR images $I_{T,l}$, $I_{T,r}$. Then, the RGB pairs $\{C_{T,l}, C_{T,r}\}$ and the generated IR pairs $\{I_{T,l}, I_{T,r}\}$ are concatenated as the input of CDE $\{concat(C_{T,l}, I_{T,l}), concat(C_{T,r}, I_{T,r})\}$.

2.1. Fourier Domain Self-Ensembling Adaptation

We construct our RIT by adopting the self-ensembling [22] strategy, which can simultaneously accomplish the RGB-to-IR relationship modeling in the source domain, and the adaptation towards target domain through the feature alignment in the Fourier domain. As shown in Fig. 2, the framework contains a student network G_S and a teacher network G_T, whose structures are exactly the same. G_S aims to learn an RGB-to-IR translation in the source domain by a supervised loss L_{sup}[2] containing a pixel-wise loss and an adversarial loss to penalize the difference between the predicted IR image $\widehat{I}_S$ and the groundtruth I_S. L_{sup} enables G_S to produce semantically consistent predictions for the source samples. The parameters Ω_T^i of the teacher G_T at training step i are updated by the student's parameters Ω_S^i, i.e., the student is exponential moving averaged to form the teacher:

$$\Omega_T^i = \alpha\Omega_T^{i-1} + (1-\alpha)\Omega_S^i, \tag{1}$$

where α is an adjustment parameter.

[2] L_{sup} is the same with that in Pix2Pix [18] and not presented here for saving space.

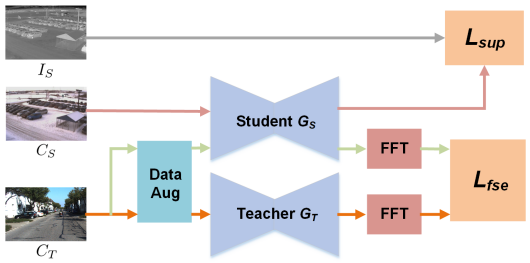

Fig. 2. Illustration of Fourier domain self-ensembling adaptation. 'DataAug' means stochastic data augmentations, while 'FFT' is the operation of the fast Fourier transform. The student network G_S is finally used for our RIT.

To adapt to the target domain, each input target sample C_T is passed through both G_S and G_T, generating predicted IR images $\widehat{I}_{T\to S}$ and $\widehat{I}_T$. Stochastic data augmentations (DataAug), e.g., dropout, noise and image flipping, are used for both pathways. Predictions from G_T can be thought of as the pseudo labels since G_T is an ensembled model that averages the student's weights. Additionally, we reformulate the traditional self-ensembling loss in the spatial domain by the Fourier transform to balance the penalties for low- and high- frequency parts:

$$\begin{aligned} L_{fse}(\widehat{I}_{T\to S}, \widehat{I}_T) = \mathbf{M}_\beta \circ ||\mathrm{FFT}(\widehat{I}_{T\to S}) - \mathrm{FFT}(\widehat{I}_T)||^2 + \\ \gamma(1-\mathbf{M}_\beta) \circ ||\mathrm{FFT}(\widehat{I}_{T\to S}) - \mathrm{FFT}(\widehat{I}_T)||^2, \end{aligned} \tag{2}$$

where FFT denotes the fast Fourier transform. $\mathbf{M}_\beta$ denotes a binary matrix containing all zeros in its center region with a ratio β of the maximum image size to mask out the low-frequency part. And '$\circ$' denotes the element-wise multiplication. γ is a parameter to trade-off the importance between both parts.

2.2. Multi-Space Warping Regularization

To generate geometry-consistent stereo IR images, the left and right RGB images $\{C_{T,l}, C_{T,r}\}$ are separately sent into G_S. And the warping regularization is imposed on output, feature and affinity spaces, to make the generated features and outputs consistent between the left and right views.

Output space. The most direct embodiment of the left-right consistency in stereo images is that pixels from both views corresponding to the same position in the scene should have similar pixel values:

$$L_O(I_{T,l}, I_{T,r}) = ||(\mathcal{W}(I_{T,r}, D)) - I_{T,l}||_1, \tag{3}$$

where $\mathcal{W}$ is the backward warping function that warps $I_{T,r}$ to the left view using D via bilinear interpolation [23].

Feature space. The loss for feature space enforces the left-right consistency from the perspective of feature representation, which is defined as:

$$L_F(F^i_{T,l}, F^i_{T,r}) = \sum_{i \in \mathcal{F}} ||(\mathcal{W}(F^i_{T,l}, D^i_{\downarrow})) - F^i_{T,r}||_1, \quad (4)$$

where $\mathcal{F}$ is a set of multi-scale layers whose feature maps F^i at ith-layer in the decoder are used to compute the loss. $D^i_{\downarrow}$ denotes the downsampled disparity map.

Affinity space. We also align the pair-wise similarities (affinity) to keep the nonlocal semantic information consistent between both views. Let $A^{j,k}_{T,l}$ denote the similarity between the j-th and the k-th pixels computed from the left-view feature map and $A^{j',k'}_{T,r}$ denote the similarity of matching point in the right view. The affinity loss is formulated as:

$$L_A = \sum_{i \in \mathcal{F}} \sum_{j} \sum_{k} (A^{i,j,k}_{T,l} - A^{i,j',k'}_{T,r})^2, \quad (5)$$

where i is the index that indicates the affinity map computed at i-th layer. The similarity $A^{i,j,k}$ is simply computed as:

$$A^{i,j,k} = \vec{F}^{i,j\mathrm{T}} \vec{F}^{i,k} / (||\vec{F}^{i,j}||_2 ||\vec{F}^{i,k}||_2). \quad (6)$$

where $\vec{F}^{i,j}$ and $\vec{F}^{i,k}$ denote the feature vector at j-th and k-th positions of i-th feature map, respectively. Thus, the overall loss is defined as:

$$L_{warp} = L_O + \rho_1 L_F + \rho_2 L_A, \quad (7)$$

where ρ_1 and ρ_2 are adjustment parameters. The final loss for RIT is defined as:

$$L_{RIT} = L_{sup} + \alpha_1 L_{fse} + \alpha_2 L_{warp}, \quad (8)$$

where α_1, α_2 are the adjustment parameters.

After the RGB-to-IR translation, the cross-modal RGB-IR pairs could be fed into CDE which can be modified from most off-the-shelf depth estimation networks [24, 11, 13] by replacing the first layer.

3. EXPERIMENTS

Datasets and Backbones. We train and evaluate our method using INO dataset [20] and KITTI dataset [17]. For the source domain, we extracted 10 frams per second from INO with a total number of 10819 samples. As our target datasets, KITTI 2012 and KITTI 2015 are real-world datasets with street views from a driving car, containing 20000+ raw stereo image data, and 194/200 stereo images with sparse groundtruth disparities for 2012/2015, respectively.

In the experiments, Pix2Pix [18] is used as our backbone of RIT. For the generator, we use the encoder-decoder mentioned in [18] with 9 resnet blocks. Three representative state-of-the-art depth estimation networks, i.e., GWCNet [11], Monodepth2 [13] and EPCDepth [21], are used as backbones to evaluate our performance, where GWCNet is a supervised method while the rest are unsupervised. Note that, all the experiment settings, i.e., data spliting on KITTI, loss functions L_{de} and evaluation metrics, are set the same way with those in the backbones for fair and easy comparison.

Training Details. For RIT, we first pre-train it from scratch with L_{sup} and L_{fse} for 150 epochs and then fine-tune it together with L_{warp} for 50 epochs. We adopt the Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The learning rate is 0.0002. For CDE, we randomly initialize the first layer and load their released models for subsequent layers, then train CDE for 300 epochs. The initial learning rate is set to 1e-4 and is down-scaled by 10 after epoch 200.

3.1. Performance Comparison

RGB-to-Infrared Translation (RIT). We compared our RIT with four translation methods, i.e., Pix2Pix [18], Pix2PixHD [25], CycleGAN [19] and GDWCT [26]. As shown in Fig. 3, Pix2Pix cannot adapt well to KITTI, which presents many messy artifacts (orange regions) and introduces too many false textures from RGB. Pix2PixHD is subject to the interference of shadows in RGB since many black areas corresponding to the shadow regions of RGB images appear on IR images (red regions). CycleGAN produces many wrong stylized textures. GDWCT preserves some image details but presents over-smooth results. Our method generates the most realistic IR images with reasonable image textures.

To further validate our RIT, we train our CDE with IR images generated by different translation methods as input and evaluate the performance of depth estimation for each method based on GWCNet [11]. The wrong stylized textures brought by CycleGAN [19] even have a negative impact on the performance and our method achieves the lowest errors for both metrics as shown in Table 1.

Table 1. Quantitative depth estimation results of our CDE based on GWCNet backbone [11] using IR images generated from different translation methods as input (error metrics, the lower the better).

Method	D1-all(%)	EPE
w/o IR (GWCNet [11])	2.202	0.657
Pix2Pix [18]	2.149	0.655
Pix2PixHD [25]	2.146	0.647
CycleGAN [19]	2.267	0.664
GDWCT [26]	2.157	0.646
Ours	**2.067**	**0.637**

Cross-Modality Depth Estimation (CDE). We first evaluate our method based on GWCNet [11], Monodepth2 [13] and EPCDepth [21] separately, and verify the effectiveness of our cross-modal strategy against these original single-modal depth estimation backbones. Table 1 and Table 3 show the numerical comparisons. Our CDE achieves the best results for all the metrics, e.g., about 6.2% error reduction on *D1-all* against GWCNet and 5.5% error reduction on *Abs Rel* against Monodepth2, respectively, demonstrating the effectiveness of our method in correcting disparity outliers and reducing the average estimation error guided by IR. Some qualitative results are shown in Fig. 4. For clear observation, we display the error maps between predicted disparities and groundtruths. Obviously, the results based on cross-modal inputs in (e) present fewer estimation errors. To further illustrate the promotion brought by introducing IR images and demonstrate the important role of IR imaging in depth estimation, we also visualize some depth estimation results in Fig. 7 to show the improvements brought by IR.

Table 2. Quantitative depth estimation results using IR images generated by different domain adaptation strategies (top) and different settings of warping regularization (bottom) based on GWCNet [11].

Setting	D1-all(%)	EPE
w/o Domain Adaptation	2.149	0.655
Self-ensembling	2.138	0.653
Fourier domain self-ensembling (FSE)	**2.123**	**0.652**
FSE + Output space regularization	2.108	0.649
FSE + Feature space regularization	2.111	0.649
FSE + Affinity space regularization	2.107	0.651
FSE + Output + Feature + Affinity	**2.067**	**0.637**

3.2. Ablation Study

We choose GWCNet [11] as the backbone of depth estimation for the ablation study, and verify the effectiveness of our key modules.

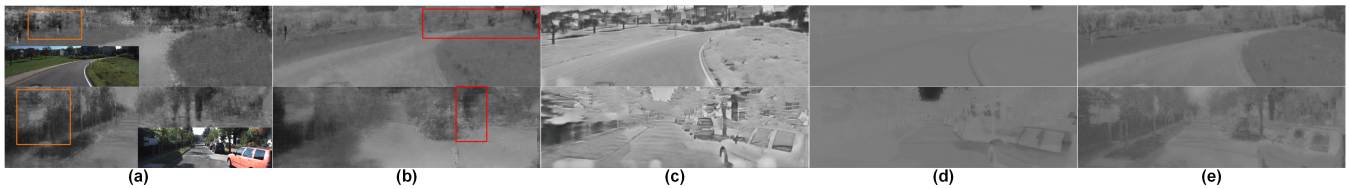

Fig. 3. Visualization of generated IR images from different translation methods on target KITTI dataset. (a) Pix2Pix [18]; (b) Pix2PixHD [25]; (c) CycleGAN [19]; (d) GDWCT [26]; (e) Ours. Color images are zoomed out and superimposed on (a) for saving space.

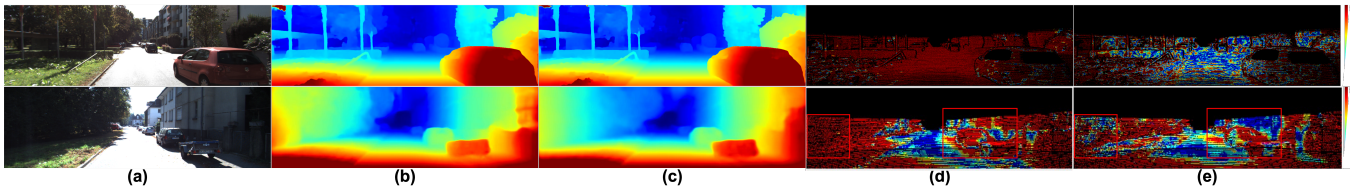

Fig. 4. Qualitative depth estimation results and error maps computed between predicted depth maps and groundtruths. (a) color images; Depth maps (blue color for remote region) estimated from (b) Original backbones and (c) Ours. (d) and (e) are error maps (blue color for lower error) corresponding to (b) and (c), respectively.

Fig. 5. Visualization of IR images generated by different domain adaptation strategies. (a) color image; (b) Baseline (Pix2Pix); (c) Spatial domain adaptation; Fourier domain adaptation with γ set at (d) 0.9 (our final choice), (e) 0.6, (f) 1.4, respectively.

Table 3. Quantitative depth estimation results based Monodepth2 [13] and EPCDepth [21] using the Eigen split [27].

Method	Error Metric (lower is better)				Accuracy Metric (higher is better)		
	RMSE	RMSE log	Abs Rel	Sq Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [13]	4.960	0.209	0.109	0.879	0.864	0.948	**0.975**
Ours	**4.918**	**0.203**	**0.103**	**0.873**	**0.875**	**0.949**	**0.975**
EPCDepth [21]	0.0969	0.669	4.304	0.183	0.893	0.963	0.982
Ours	**0.0919**	**0.665**	**4.250**	**0.178**	**0.899**	**0.966**	**0.983**

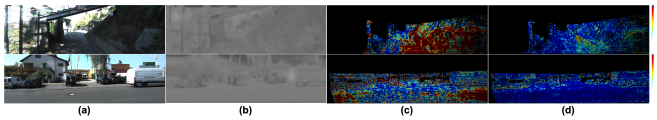

Fig. 6. Visualization of warping errors (blue for lower errors) computed by the L1 error loss between IR images generated from left and warped right views. (a) RGB images (b) Generated IR images; Results from the cases (c) 'w/o Regularization'; (d) 'Multi-Space'.

Effectiveness for Fourier domain adaptation. We present the IR images generated by different domain adaptation strategies in Fig. 5. Compared with Pix2Pix [18] that cannot generalize well to the target dataset (many artifacts chequered with black and white), the spatial domain adaptation reduces the artifacts, but leads to blurred boundaries and loss of details. Our Fourier domain adaptation solves this by tuning a balance parameter γ between the high- and low-frequency parts of the generated IR images in the Fourier domain. Overemphasizing high-frequency information (Fig. 5(e)) will lead to the introduction of wrong color textures, otherwise it will reduce the image contrast and lose details (Fig. 5(f)). We set $\gamma = 0.9$ according to the experimental test, which can obtain the superior visual performance and benefit the final estimation as shown in Table 2.

Effectiveness for multi-space warping regularization. We also conduct an ablation study in Table 2 to analyse the elements in our multi-space warping regularization. The case without warping regularization obtains a relatively poor result. When adding the constraints of output, feature, and affinity space into 'w/o Regularization', improvements can be observed. With all three constraints, the errors are the lowest. Fig. 6 further demonstrates the effectiveness of our multi-space warping regularization through visualizing the warping error maps.

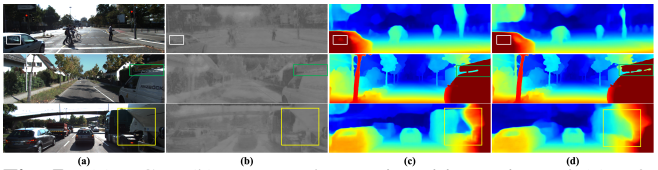

Fig. 7. (a) RGB; (b) generated IR; Disparities estimated (c) w/o IR and (d) with IR (ours). Benefit from IR's robustness to reflection, light changes and shadows, improvements can be observed when encountering: 1) Reflection (white rectangles); 2) Brightness changes (green); and 3) object contours in shadow (yellow).

4. CONCLUSION

The core idea of this work is to develop an RIT network to synthesize stereo IR images from their corresponding stereo RGB images and build a CDE network to infer disparity maps from cross-modal RGB-IR stereo pairs, which can both overcome the weakness of single-modality based methods and keep the low requirement on training dataset. Experiments demonstrate our superiority on the tasks of stereo IR images generation and depth estimation.

5. REFERENCES

[1] Dimitrios S. Alexiadis, Anargyros Chatzitofis, Nikolaos Zioulis, Olga Zoidi, Georgios Louizis, Dimitrios Zarpalas, and Petros Daras, "An integrated platform for live 3d human reconstruction and motion capturing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 798–813, 2017.

[2] Felipe Codevilla, Eder Santana, Antonio M. Lopez, and Adrien Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[3] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, and Wanli Ouyang, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2018.

[4] Lincheng Li, Shunli Zhang, Xin Yu, and Li Zhang, "Pmsc: Patchmatch-based superpixel cut for accurate stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 679–692, 2018.

[5] He Dai, Xuchong Zhang, Yongli Zhao, Hongbin Sun, and Nanning Zheng, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[6] Baiyu Pan, Liming Zhang, and Hanzi Wang, "Multi-stage feature pyramid stereo network-based disparity estimation approach for two to three-dimensional video conversion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1862–1875, 2021.

[7] Fangzheng Tian, Yongbin Gao, Zhijun Fang, Yuming Fang, Jia Gu, Hamido Fujita, and Jenq-Neng Hwang, "Depth estimation using a self-supervised network based on cross-layer feature fusion and the quadtree constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[8] Shu Chen, Zhengdong Pu, Xiang Fan, and Beiji Zou, "Fixing defect of photometric loss for self-supervised monocular depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[9] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

[11] 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.

[12] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

[14] X. I. Lin, S. Y. Sun, L. I. Lin-Na, and F. Y. Zou, "Depth estimation from monocular infrared images based on svm model," *Laser & Infrared*, 2012.

[15] S. Sun, L. Li, and H. Zhao, "Depth estimation from monocular vehicle infrared images based on KPCA and BP neural network," *Hongwai yu Jiguang Gongcheng/Infrared and Laser Engineering*, vol. 42, pp. 2348–2352, 09 2013.

[16] Shouchuan Wu, Haitao Zhao, and Shaoyuan Sun, "Depth estimation from infrared video using local-feature-flow neural network," *International Journal of Machine Learning and Cybernetics*, vol. 10, 09 2019.

[17] 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.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[19] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] "INO dataset. `https://www.ino.ca/en/video-analytics-dataset/`,".

[21] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *International Conference on Computer Vision (ICCV)*, October 2021.

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

[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[24] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo, "Image-to-image translation via group-wise deep whitening-and-coloring transformation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *IEEE International Conference on Computer Vision (ICCV)*, December 2015.