# ITSA: An Information-Theoretic Approach to Automatic Shortcut Avoidance and Domain Generalization in Stereo Matching Networks Supplementary Materials

WeiQin Chuah* Ruwan Tennakoon* Reza Hoseinnezhad* Alireza Bab-Hadiashar* David Suter†
RMIT University, Australia* Edith Cowan University (ECU), Australia†
{wei.qin.chuah,ruwan.tennakoon,rezah,abh}@rmit.edu.au, d.suter@ecu.edu.au

| Methods | Time ($s$) | GPU Memory ($Mb$) |
|---|---|---|
| RIB [12] | 2.451 | $5,931$ |
| ITSA | 0.007 | $1,225$ |

Table 1. Comparison of GPU memory requirement and training time per iteration between the robust information bottleneck (RIB) and the proposed ITSA methods. The RIB method has significantly higher GPU memory requirement and longer training time as compared to the ITSA.

## 1. Implementation details

### 1.1. Toy Experiment

In this section, we discuss the implementation details of our toy experiment (refer to Section 3.3 of the main manuscript). Following [1], we modelled the digit recognition network as a variational auto-encoder. Our encoder has the following structure: *conv-pool-conv-pool* where it consists of two $5 \times 5$ convolutional layers with 64 and 128 channels respectively. Each convolutional layer is followed by a ReLU activation function and a max pooling operation with kernel size of $2 \times 2$. We utilize two Fully-Connected (FC) layers to estimate the parameters mean $\mu$ and standard deviation $\sigma$ of the latent distributions. The dimension of the latent features is set to 256. Meanwhile, the decoder has the following structure: $fc$-$fc$-$softmax$. For the decoder, the input dimension of the two FC layers was set to 1024 and the size of the softmax layer is 10 (number of classes). The batch size was set to 128 and Adam was selected as the optimizer. The networks were trained for 100 epochs with a constant learning rate of $1e^{-4}$. We apply the same settings for all four models, namely the baseline, the standard Information Bottleneck [1], the Robust Information Bottleneck [12] and our proposed method. Following [13], we train the networks using the first $10,000$ samples in the MNIST [8] training set. The optimized networks were evaluated on the MNIST and MNIST-M [4] test sets. The hyper-parameter $\beta$ in the standard and robust information bottleneck networks was set to $0.01$ respectively. For the proposed method, the hyper-parameter $\epsilon$ and $\lambda$ were set to $0.8$ and $0.1$ respectively. The comparison of training time per iteration (s) and GPU memory requirement are included in Tab. 1. The comparison clearly shows that our proposed ITSA method requires significantly lower training time and GPU memory consumption as compared to the robust information bottleneck (RIB) method.

| Networks | KITTI | Baseline | $\epsilon$ | | | |
|---|---|---|---|---|---|---|
| | | | 0.1 | 0.2 | 0.5 | 1 |
| PSMNet | 2012 | 27.8 | 6.3 | 5.5 | **5.2** | 6.0 |
| GwcNet | | 11.7 | 5.6 | 5.1 | **4.9** | 5.6 |
| PSMNet | 2015 | 30.7 | 6.9 | 6.1 | **5.8** | 6.4 |
| GwcNet | | 12.8 | 5.9 | 5.5 | **5.4** | 5.7 |

Table 2. The relationship between the perturbation strength $\epsilon$ in SCP and the performance of stereo disparity estimation in PSMNet [2] and GwcNet [6]. The performances were evaluated on the KITTI-2012 [5] and KITTI-2015 datasets [11], using the D1 metric. The hyper-parameter $\lambda$ was set to 0.1 in these experiments.

### 1.2. Semantic Segmentation

For the semantic segmentation task discussed in Section 4.6 of the main manuscript, we adopted the Fully Convolutional Networks (FCN) [9] backboned with the ImageNet pre-trained ResNet-50 [7] as our model. The network was trained using the GTAV [14] synthetic dataset, which consists of $24,966$ samples with annotations compatibles with the Cityscapes [3] dataset. We randomly selected $23,466$ samples as training set and the remaining $1500$ as validation set. The optimized model was evaluated on the Cityscapes validation set, using the mean intersection over union (mIoU) metric. The mIoU is the average of all IoU values over all classes. Training was conducted for 20 epochs using the Adam optimizer with an initial learning rate of 1e-4. We also adopted the polynomial learning rate scheduling with the power of 0.9. The batch size was set to 12 and the hyper-parameter $\epsilon$ and $\lambda$ were set to 0.2 and

| Networks | KITTI | λ |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | 0.0 | 0.01 | 0.1 | 0.5 | 1.0 | 2.0 |
| PSMNet | 2012 | 8.1 | 7.2 | **5.2** | 7.1 | 6.8 | 8.4 |
| GwcNet |  | 5.3 | 5.5 | **4.9** | 5.5 | 5.8 | 6.1 |
| PSMNet | 2015 | 8.6 | 6.6 | **5.8** | 7.2 | 6.6 | 8.3 |
| GwcNet |  | 5.9 | 6.0 | **5.4** | 5.9 | 6.2 | 6.9 |

Table 3. Relationship between hyper-parameter $\lambda$ and the performance of stereo disparity estimation in PSMNet [2] and GwcNet [6]. The performances were evaluated using the D1 metric, and the $\epsilon$ was set to 0.5.

| Methods | $\nabla_x z$ | $\lambda$ |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | 0.0 | 0.01 | 0.1 | 0.5 | 1.0 | 2.0 |
| PSMNet [2] | Max | 43.1 | 39.7 | 16.9 | 11.5 | 9.1 | 8.4 |
|  | Min | -39.4 | -41.4 | -24.5 | -12.0 | -7.4 | -8.7 |
| GwcNet [6] | Max | 49.5 | 29.8 | 17.8 | 14.1 | 12.2 | 9.2 |
|  | Min | -33.4 | -31.7 | -17.7 | -14.3 | -11.2 | -8.5 |

Table 4. Relationship between hyper-parameter $\lambda$ and the range of the gradient $\nabla_x z$. The gradients are computed using the Scene Flow testing set.

0.01 respectively. To prevent the model from overfitting, data augmentations such as color jittering, Gaussian blur, random grey-scaling and random cropping were conducted.

## 2. Hyper-parameters Selection

In this section, we discuss the effect of the hyper-parameters $\epsilon$ and $\lambda$ on the performance of stereo matching networks. The hyper-parameter $\epsilon$ controls the strength of perturbation in the proposed shortcut perturbation (SCP) augmentation method. As shown in Tab. 2, the proposed SCP method can consistently improve the performance of stereo matching in both PSMNet and GwcNet. The best performance was achieved when $\epsilon$ was set to 0.5.

Meanwhile, the hyper-parameter $\lambda$ controls the trade-off between learning task-relevant and shortcut-invariant features. As shown in Tab. 3, for both PSMNet and GwcNet, the best results was obtained when $\lambda = 0.1$. Furthermore, in Tab. 4, we observe that the sensitivity of the extracted features with respect to the changes in the inputs (represented by the magnitude of the gradient $\nabla_x z$) decreases as $\lambda$ increases. We speculate that large values of $\lambda$ will cause the network to over-prioritize the learning of shortcut-invariant features and harm the performance of stereo matching. Conversely, the network will focus on learning task-relevant features which may include shortcuts and deteriorate the domain generalization performance, when the $\lambda$ is small.

## 3. Lemma 3.1 - Proof

In this section, we provide the proof to Lemma 3.1 discussed in the main manuscript. We first take the first order

| Models | FT | ITSA | Sun | Cloud | Rain | Fog | Night | Avg |
|---|---|---|---|---|---|---|---|---|
| PSMNet [2] | ✓ | ✗ | 3.94 | 2.82 | 11.51 | 6.50 | 16.66 | 8.28 |
|  | ✗ | ✓ | 4.78 | 3.24 | 9.43 | 6.31 | 8.56 | 6.46 |
|  | ✓ | ✓ | 1.94 | 1.61 | 4.12 | 1.72 | 8.51 | 3.58 |
| GwcNet [6] | ✓ | ✗ | 3.10 | 2.46 | 12.34 | 5.98 | 25.33 | 9.84 |
|  | ✗ | ✓ | 4.35 | 3.31 | 9.78 | 5.88 | 9.41 | 6.55 |
|  | ✓ | ✓ | 2.18 | 2.07 | 9.21 | 2.16 | 8.37 | 4.80 |
| CFNet [16] | ✓ | ✗ | 1.79 | 1.65 | 5.20 | 1.59 | 11.56 | 4.36 |
|  | ✗ | ✓ | 3.42 | 2.87 | 5.32 | 4.32 | 8.95 | 4.98 |
|  | ✓ | ✓ | 1.84 | 1.55 | 2.40 | 1.58 | 5.69 | 2.61 |

Table 5. Robustness evaluation on anomalous scenarios. Fine-tuning the synthetically pre-trained stereo matching networks using the proposed ITSA method can significantly enhances the robustness of the fine-tuned models in the real-world anomalous scenarios including rainy and foggy weather and night-time. The performances were evaluated using the D1 metric.

approximation of $p_{(Z|X=x^*)}$, which is defined as:

$$p_{(Z|X=x^*)} = p_{(Z|X=x)} + \epsilon u^\top \cdot \nabla_x p_{(Z|X=x)}$$

Given $A \cdot B = \|A\| \|B\| \cos\theta$, we have:

$$p_{(Z|X=x^*)} - p_{(Z|X=x)} = \epsilon \, \|u\| \left\| \nabla_x p_{Z|X=x} \right\|_2 \cos\psi$$

Rearrange and we have:

$$\begin{aligned} \left| \frac{p_{Z|X=x^*} - p_{Z|X=x}}{\epsilon \cos\psi} \right| &= \left\| \nabla_x p_{Z|X=x} \right\|_2 \\ \left| \frac{p_{Z|X=x^*} - p_{Z|X=x}}{\epsilon \cos\psi} \right| &= \left\| \nabla_x \log p_{Z|X=x} \right\|_2 p_{Z|X=x} \end{aligned}$$

The expectation is then written as:

$$\mathbb{E}_z \left[ \left| \frac{p_{Z|X=x^*} - p_{Z|X=x}}{\epsilon \cos\psi} \right| \right] = \mathbb{E}_z \left[ \left\| \nabla_x \log p_{Z|X=x} \right\|_2 \right]$$

and by taking the square on both sides, we can expand it as follow:

$$\begin{aligned} &\mathbb{E}_z \left[ \left| \frac{p_{Z|X=x^*} - p_{Z|X=x}}{\epsilon \cos\psi} \right| \right]^2 \\ &= \mathbb{E}_z \left[ \left\| \nabla_x \log p_{Z|X=x} \right\|_2 \right]^2 \\ &= \mathbb{E}_z \left[ \left\| \nabla_x \log p_{Z|X=x} \right\|_2^2 \right] - \mathcal{V} \left[ \left\| \nabla_x \log p_{Z|X=x} \right\|_2 \right] \\ &= \Phi \left( Z \mid X = x \right) - \mathcal{V} \left[ \left\| \nabla_x \log p_{Z|X=x} \right\|_2 \right] \end{aligned}$$

where $\mathcal{V}[\cdot]$ is the variance. Therefore,

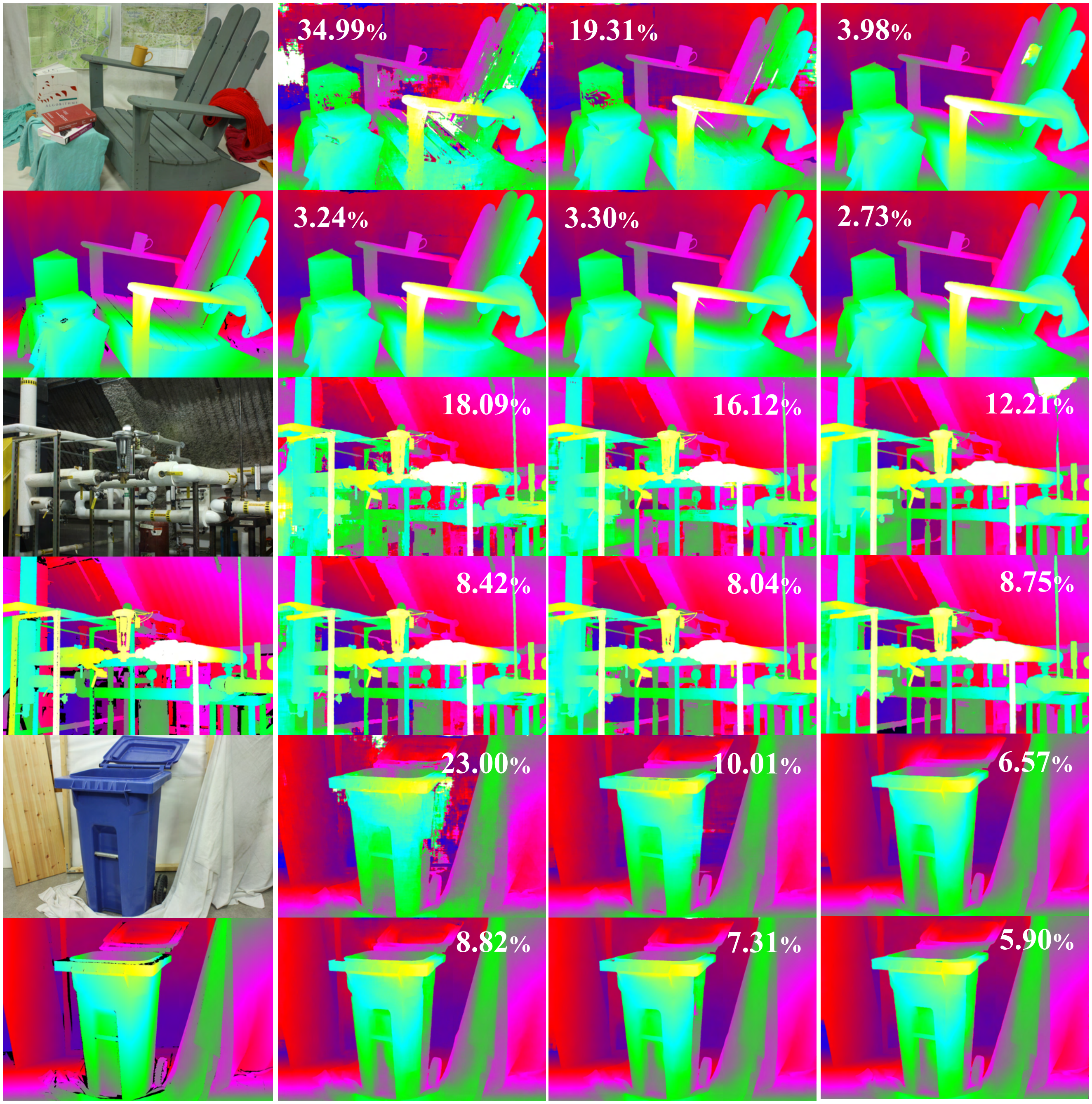
$$\begin{aligned} \Phi \left( Z \mid X = x \right) = & \frac{\mathbb{E}_z \left[ \left| p_{Z|X=x^*} - p_{Z|X=x} \right| \right]^2}{\epsilon^2 \cos^2\psi} \\ & + \mathcal{V} \left[ \left\| \nabla_x \log p_{Z|X=x} \right\|_2 \right] \end{aligned}$$

## 4. ITSA-Finetuned Networks Robust Analysis

In this section, we evaluate the robustness of stereo matching networks fine-tuned using the proposed ITSA approach. Our experimental results showed that using the proposed ITSA for fine-tuning the selected stereo matching networks can further enhance the networks' robustness to anomalous scenarios. As shown in Tab. 5, PSMNet [2] and GwcNet [6] that are fine-tuned on the KITTI-2015 train set, using the ITSA method have achieved an overall improvement of $4.70\%$ and $5.04\%$. Furthermore, ITSA can also improve the robustness of the top-performing CFNet [16] in the challenging real-world scenarios ($1.75\%$ overall improvement).

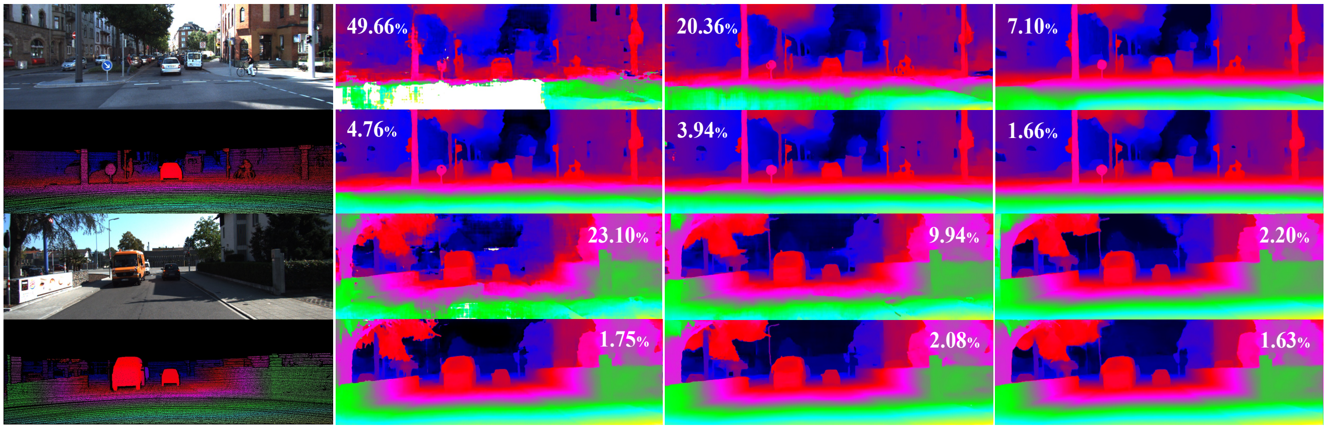
## 5. Additional Qualitative Results

Additional qualitative results of stereo matching tasks on indoor and outdoor realistic data are included in Fig. 1 and Fig. 2. We also included the qualitative results for anomalous scenarios (e.g. rainy weather and night-time) in Fig. 3 and Fig. 4. Furthermore, additional qualitative results of semantic segmentation task are also included in Fig. 5.



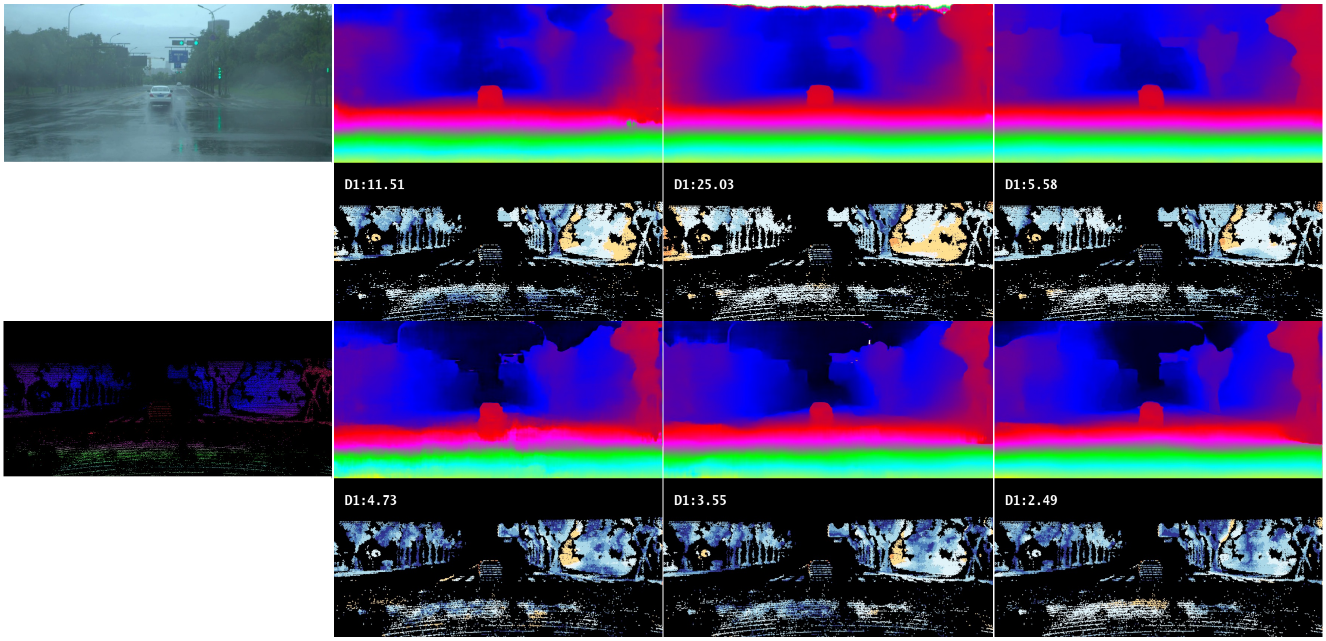
(a) Left (b) PSMNet (c) GwcNet (d) CFNet

Figure 1. Qualitative comparison on the Middlebury (half resolution) dataset [15], using the PSMNet [2], GwcNet [6] and CFNet [16]. For each example, the left stereo image and the ground truth disparity map are included in the left column. Moreover, the disparity maps estimated by the baseline networks (pretrained on Scene Flow) are included on the top row and the results of our method are included in the bottom row. The corresponding 2-pixel threshold error rate is also included on the predicted disparity map.



(a) Left (b) PSMNet (c) GwcNet (d) CFNet

Figure 2. Qualitative comparison on the KITTI 2015 [11] training set, using the PSMNet [2], GwcNet [6] and CFNet [16]. For each example, the left stereo image and the ground truth disparity map are included in the left column. Moreover, the disparity maps estimated by the baseline networks (pretrained on Scene Flow) are included on the top row and the results of our method are included in the bottom row. The corresponding D1 error rate is also included on the predicted disparity map.



(a) Left (b) PSMNet (c) GwcNet (d) CFNet

Figure 3. Qualitative comparison on out-of-distribution rainy day data provided by the DrivingStereo [17]. The estimated disparity maps are generated using the PSMNet [2], GwcNet [6] and CFNet [16]. The left stereo image and the ground-truth disparity map are included in the left column. Moreover, the disparity maps estimated by the KITTI-2015 [11] fine-tuned networks are included on the first row (b-d), and the results of our method (ITSA) are included in the third row (b-d). The corresponding D1 error rate is also superimposed on the included error map. Our method can significantly improve the performance of these stereo matching networks in challenging unseen domains, despite training on the synthetic data only..

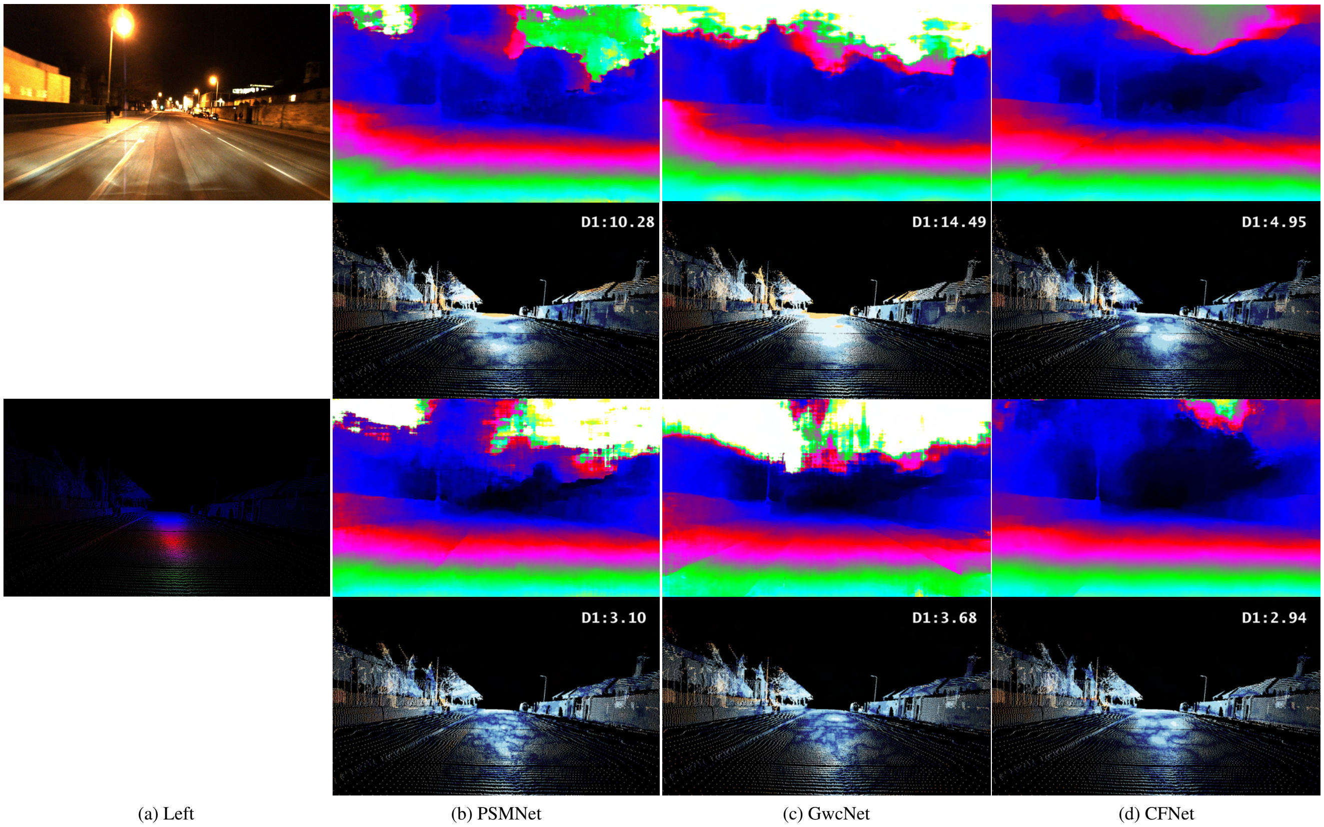


Figure 4. Qualitative comparison on out-of-distribution night-time data provided by the Oxford Robotcar [10]. The estimated disparity maps are generated using the PSMNet [2], GwcNet [6] and CFNet [16]. The left stereo image and the ground-truth disparity map are included in the left column. Moreover, the disparity maps estimated by the KITTI-2015 [11] fine-tuned networks are included on the first row (b-d) and the results of our method (ITSA) are included in the third row (b-d). The corresponding D1 error rate is also superimposed on the included error map. Our method can significantly improve the performance of these stereo matching networks in challenging unseen domains.

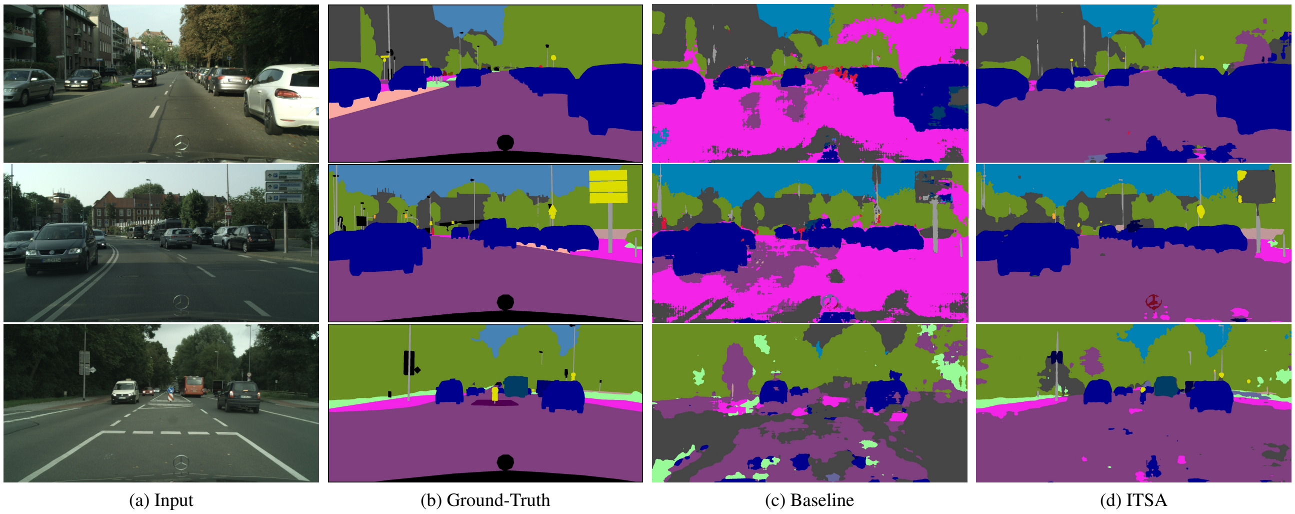


Figure 5. Qualitative results of semantic segmentation. We have employed the FCN-8 backboned with the ResNet-50 as our network. Our method (ITSA) can be extended to semantic segmentation networks and enhance the synthetic-to-real generalization performance. Best view in color and zoom in for details.

## References

[1] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. 1

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1, 2, 3, 4, 5, 6

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[6] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 1, 2, 3, 4, 5, 6

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[10] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 6

[11] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018. 1, 5, 6

[12] Ankit Pensia, Varun Jog, and Po-Ling Loh. Extracting robust and accurate features via a robust information bottleneck. *IEEE Journal on Selected Areas in Information Theory*, 1(1):131–144, 2020. 1

[13] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 1

[14] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 1

[15] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014. 4

[16] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, June 2021. 2, 3, 4, 5, 6

[17] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 5