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

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

Abstract

State-of-the-art stereo matching networks trained only on synthetic data often fail to generalize to more challenging real data domains. In this paper, we attempt to unfold an important factor that hinders the networks from generalizing across domains: through the lens of shortcut learning. We demonstrate that the learning of feature representations in stereo matching networks is heavily influenced by synthetic data artefacts (shortcut attributes). To mitigate this issue, we propose an Information-Theoretic Shortcut Avoidance (ITSA) approach to automatically restrict shortcut-related information from being encoded into the feature representations. As a result, our proposed method learns robust and shortcut-invariant features by minimizing the sensitivity of latent features to input variations. To avoid the prohibitive computational cost of direct input sensitivity optimization, we propose an effective yet feasible algorithm to achieve robustness. We show that using this method, state-of-the-art stereo matching networks that are trained purely on synthetic data can effectively generalize to challenging and previously unseen real data scenarios. Importantly, the proposed method enhances the robustness of the synthetic trained networks to the point that they outperform their fine-tuned counterparts (on real data) for challenging out-of-domain stereo datasets.

1. Introduction

Stereo matching is a fundamental task in computer vision and is widely used for depth sensing in various applications such as augmented reality (AR), robotics and autonomous driving. In recent years, end-to-end trained Convolutional Neural Networks (CNNs) have achieved impressive results for this task as quantified by the performance on several publicly available stereo-matching benchmarks [5,15,17,44,51].

Generally, end-to-end stereo-matching networks require a large amount of labelled data for training. To overcome this challenge, many state-of-the-art networks are initially trained on labelled synthetic data, commonly generated using game engines. However, models trained using synthetic data do not generalize well to unseen realistic domains. For example, the PSMNet [5] pre-trained on the Scene Flow dataset [24] performs poorly when tested on unseen realistic domains as illustrated in Fig. 1. Therefore, in practice, the networks trained with synthetic data are fine-tuned using labelled data from the relevant target domain. However, collecting even a relatively small amount of dense ground truth data in the real-world can be challenging for tasks like stereo-matching [21,40]. Furthermore, to be practically useful in many applications, a stereo-matching model should be able to generalize effortlessly to different do-

![Figure 1](image-url)
mains like day and night times, varying weather conditions, etc. Collecting data for fine-tuning that cover all possible situations is both difficult and expensive. It is therefore highly desirable to remove the fine-tuning requirement.

It is known that neural networks, including stereo matching networks, can learn superficial shortcut features (or spurious correlations with the target labels), which prevent them from generalizing across different domains [2,11]. We found that stereo matching networks trained on synthetic data are susceptible to exploiting shortcuts in synthetic data such as (1) consistent local statistics (RGB color features) between the left and right stereo images and (2) over-reliance on local chromaticity features (e.g. color, illumination, texture) of the reference stereo viewpoint. Detailed analysis and discussion are included in Sec. 4.2. Dependency on these shortcut cues, instead of the desirable semantic and structural representations, means that these networks would fail drastically when the spurious correlations between shortcuts and labels do not exist in a new (unseen) domain [32]. While several shortcut-removal approaches have been previously proposed [3,16,37], most of these methods are manually designed (e.g. carefully selected data augmentations [3,16]) and rely on the assumption that the shortcuts could be identified in advance. However, shortcuts can be non-intuitive, task-specific, and difficult to identify [8,26].

Our goal is to train a stereo matching network on synthetic data that can generalize to realistic scenes without the need for fine-tuning. To achieve this, we propose an information-theoretic approach to automatically restrict the shortcut-related information from being encoded from the input into the feature representations. The approach is based on the well known information bottleneck (IB) principle that proposes to optimize the following objective [1,39]:

$$\arg\max_{\theta} I(Y;Z;\theta) - \beta I(X;Z;\theta)$$  \hspace{1cm} (1)$$

where $Z$ is the encoding of input $X$, $Y$ is the target, $I$ is mutual information and $\beta \in [0, 1]$ is the hyperparameter that controls the size of the information bottleneck. While optimizing the IB objective leads to compressed feature representations, our empirical experiments showed that these compressed features are neither robust nor shortcut-invariant (details are provided in Sec. 3.3.1). Consequently, the IB optimized networks may still incorporate shortcuts and remain fragile when tested in unseen domains. The recently introduced robust IB criterion [30] encourages the learning of both robust and compressive features by replacing the mutual information in IB with statistical Fisher information. Robust IB is presented in the context of learning features that are robust to adversarial attacks and to the best of our knowledge it has not been used for domain generalization.

In our approach, we combine the task loss (e.g. smooth L1 loss) with Fisher information to learn a generalizable stereo matching model. Although such an objective can work in theory, straightforward optimization of the Fisher information by gradient descent requires computation of the second-order derivatives and is therefore computationally expensive for tasks with high dimensional inputs such as stereo matching and semantic segmentation. To overcome this shortcoming, we propose ITSA which consists of a novel loss term and perturbation technique to approximate the optimization of the Fisher information loss. The proposed ITSA is computationally efficient, and as we show by extensive experiments, it can promote the learning of shortcut-invariant features. Unlike the existing domain-invariant stereo matching networks [36,52], the proposed ITSA does not involve significant network alteration and is model-agnostic. Therefore, as shown in the experiments section, it can be easily integrated with different stereo matching networks.

The empirical results show that stereo-matching networks trained on synthetic data, with the proposed ITSA, can generalise to realistic data without fine-tuning. Additional experiments on challenging out-of-domain stereo datasets (e.g. different adverse weathers and night scenes) show that our method also improves the overall robustness of the stereo matching networks and importantly even outperforms the networks fine-tuned on realistic domains when tested on these challenging datasets. The main contributions of this paper include:

- We show that learning feature representations that are less sensitive to input variations can significantly enhance the synthetic to realistic domain generalization, and robustness in stereo matching networks.
- We introduce a novel loss function that enables us to minimize the Fisher information, without computing the second-order derivatives.
- We also show that the application of the proposed framework is not limited to stereo matching task, and can be used in training models for non-geometry based vision problems such as semantic segmentation.

The rest of the paper is organized as follows. Sec. 2 describes the related work in the field of learning-based stereo matching networks, domain generalization and shortcut learning. Sec. 3 presents the proposed method for automatic shortcut avoidance and domain generalization. Experimental results and discussions are presented in Sec. 4, and Sec. 6 concludes the paper.

2. Related Work

Learning-based Stereo Matching Networks

In recent years, end-to-end learned deep stereo match-
ing networks have excelled in most datasets and benchmarks [5, 17, 44, 51]. These networks generally have three sub-modules (1) feature extraction sub-network, (2) cost-volume generator, and (3) cost aggregation and refinement sub-network. There are two main types of stereo matching networks based on how the cost volume is generated.

Correlation-based stereo matching networks construct the cost volume by correlating the features extracted from the two views. Previously proposed correlation-based methods include DispNetC [24], iResNet [20], CRL [29], SegStereo [47], and AANet [44]. Although these methods are usually computationally efficient, semantics and structural information in the feature representations are lost due to the correlation operation [15]. As a result, the correlation-based stereo matching methods usually have inferior performance compared to the concatenated-based methods.

Concatenation-based methods use a cost volume that is a simple assembly of features extracted from the two views. Examples of the state-of-the-art concatenation-based stereo matching networks include PSMNet [5], GANet [51], GCNet [17], StereoDrNet [4] and EMCUA [27]. While these networks can achieve superior performance in stereo matching, they require labelled samples from the target environments, for fine-tuning. Without fine-tuning, these networks cannot generalize to unseen test data.

To overcome this problem, Zhang et al. [52] proposed DSMNet, which employs Domain Normalization and non-local graph-based filtering layers to enforce the learning of structural features that are domain-invariant. Similarly, Shen et al. [36] introduced CFNet, an efficient network architecture with multi-scale cost volume fusion and refinement, to enforce the learning of robust structural representation for stereo matching. In contrast, we have identified shortcut learning [12] as a major factor that hinders stereo matching networks from generalizing across domains. In this work, we show that avoiding shortcut learning can effectively enhance the robustness of the stereo matching networks and enables a model to generalize across domains. This is evidenced by showing networks’ superior performance on challenging realistic data without fine-tuning.

Single Domain Generalization

Domain generalization typically involves forcing DNNs to learn domain-invariant features, using data sampled from multiple source domains [19, 31]. On the other hand, single domain generalization is a more challenging problem because only one source domain is available for training. To solve this problem, Volpi et al. [41] proposed adversarial data augmentation (ADA), which aims to expand and diversify the distribution of training data. Specifically, ADA creates “fictitious” yet “challenging” new populations, simulating data sampled from novel domains, using adversarial training. In a similar fashion, Qiao et al. [31] proposed a novel framework that employs ADA and meta-learning to enforce the learning of domain-invariant features. While these works focus on minimizing the domain differences, we are interested in learning robust and shortcut-invariant features that are transferable across different domains. To this end, we propose ITSA, an information-theoretic approach to prevent shortcut learning (see next section), particularly in the stereo matching networks.

Shortcut Learning

Geirhos et al. [11] coined the term shortcut learning as a phenomenon where DNNs learn trivial solutions by relying on superficial features (shortcuts). These features are spuriously correlated with the target labels, without contributing to transferability across contexts. For example, image classification networks tend to rely on shortcuts such as backgrounds [2, 11] and textures [15, 42] to improve their performance. However, these networks fail to generalize to unseen domains, where the spurious correlations between shortcuts and labels are violated [32]. Similarly, we observed that stereo matching networks trained on synthetic data also have a tendency to exploit shortcuts to produce accurate depth results in synthetic domains. Consequently, these networks fail drastically when tested in unseen realistic environments.

Several attempts have been made to restrict the learning of identified shortcuts and generalize DNNs across domains [3, 6, 16, 37, 42]. These methods reply on having some shortcut-related prior knowledge and usually include data augmentations [3, 16], whitening transformation [6] or dropout-based regularization [37] as part of their solutions. However, shortcuts are non-trivial, task-specific, and are often difficult to be identified a priori [8, 26]. In contrast, our proposed method automatically avoid shortcut learning without requiring shortcut-related knowledge in advance.

3. Methodology

3.1. Problem Definition

In this work, we focus on the synthetic-to-realistic domain generalization for stereo matching. Given a synthetic stereo data set $D_{syn}$ consisting of stereo image pairs $\{x_{syn,i}, y_{syn,i}\}_{i=1}^{n}$ with corresponding ground-truth disparity $\{y_{syn,i}\}_{i=1}^{n}$, the goal is to design a robust and shortcut-invariant stereo matching network that can accurately predict disparity map $\hat{y}_{i}$ for unseen realistic environments $D_{real}$.

Our approach to achieve synthetic-to-realistic domain generalization is to use an information-theoretic measure to automatically restrict the shortcut-related information from being included in feature representations.
3.2. Model

A typical stereo-matching network can be represented by the following equation:

\[ \hat{y}^{(i)} = m_\psi \left( C \left( f_\theta \left( x_1^{(i)} \right), f_\theta \left( x_r^{(i)} \right) \right) \right) \]  

(2)

where \( f_\theta (\cdot) \) is the feature extraction sub-network, \( C (\cdot) \) the cost volume and \( m_\psi (\cdot) \) the cost aggregation and refinement sub-network. The refined cost volumes are converted to disparity maps \( \hat{y} \) via the soft argmin [17] operation.

Our proposed method (ITSA) can be applied to any stereo-matching network that has the above structure. In the experiments section, we show the result of applying the proposed algorithm to different stereo-matching networks with concatenation cost (we observed similar results with correlation-based methods) volumes [5, 15, 36]. The high-level structure of the network including the proposed shortcut avoidance strategy is shown in Fig. 2.

3.3. Loss function

Our main contribution is the loss function devised to automatically restrict the shortcut-related information from being encoded in the learning process. As we explained earlier, the information bottleneck (IB) principle [1, 39] is typically used to compress features and would be a natural choice to achieve this objective.

The standard \( \mathcal{L}_{\text{IB}} \) loss defined in Eq. (1), which uses mutual information to quantify information content, was designed to extract features that are both concise and relevant for prediction. However, models trained by this loss are not robust to existence of artefacts that can generate shortcuts (similar to adversarial distortions mentioned in [30]).

To demonstrate the above point, we conducted a toy experiment. In this experiment, we investigated the efficacy of using IB loss for helping digit recognition networks (DRNs) to generalize from MNIST (source) [18] to MNIST-M [9] (target) dataset. The former contains images of handwritten digits with black background, and the latter is created by combining the MNIST digits with randomly extracted from color patches as their background. All networks were trained on the MNIST training set only and the top-1 accuracy (%) was employed for evaluation. The details of the experiment are included in the supplementary document. As shown in Tab. 1, the standard IB can effectively reduce overfitting, and achieves the best performance in the source domain. However, it fails to generalize its performance to the unseen domain. Importantly, it even performs worse than the baseline networks in the unseen target domain.

3.3.1 Robust Information Bottleneck and Fisher Information

As our aim is to develop an IB based cost function that is not susceptible to existence of shortcuts in source data, we take inspiration from the robust IB principle [30]. Robust IB utilizes the statistical Fisher information \( \Phi(Z|X) \) of the extracted features \( Z \) parameterized by the inputs \( X \) as a more robust measure of information (in place of \( I(Z, X) \)). The Fisher information \( \Phi(Z|X) \) is defined as:

\[ \Phi(Z|X) = \int_X \Phi(Z|X = x)p_X(x)dx, \]  

(3)
where
\[ \Phi(Z|X = x) = \int_Z \left| \nabla_x \log p_Z|X(z|x) \right|^2 p_Z|X(z|x) dz. \tag{4} \]

The term \( \Phi(Z|X = x) \) in Eq. (3, 4) can be regarded as the sensitivity of the latent distribution \( p_Z|X(\cdot|x) \), with respect to changes at the input \( x \). Therefore, optimizing the Fisher information, \( \Phi(Z|X) \), will minimize the average sensitivity of the latent distribution with respect to change of inputs \( X \). As shortcuts are generated by data artefacts that are transient \(^1\) by nature, they are sensitive to perturbations of input data \([11]\). As such, minimizing the Fisher information is a step towards promoting the learning of shortcut-invariant features. Our conjecture is supported by the results of the toy experiment included in Tab. 1. The DRNs constrained by the Fisher information (RIB) achieved better results of the toy experiment included in Tab. 1. The DRNs in the target domain.

In order to minimize the Fisher information expressed in Eq. (4), one has to compute second order derivatives of the overall loss function, such as \( \nabla_z \log p_Z|X(z|x) \), which is computationally prohibitive for tasks with large dimensional inputs such as stereo matching, semantic segmentation, etc. \([38]\). To overcome this issue, we propose ITSA, a simple yet computationally feasible approach to promote the learning of shortcut-invariant features.

### 3.3.2 Approximating Fisher information

Optimizing the Fisher information \( \Phi(Z|X) \) measure defined in Eq. (3) is related to minimizing \( \Phi(Z|X = x) \). By adding a regularization term such as \( \Phi(Z|X = x) \) to the loss function, we can penalize the transient features and discourages networks from learning shortcuts. To calculate this term, we employ a first order approximation as described below.

**Lemma 3.1.** If \( \epsilon > 0 \), \( u \) is a unit vector (i.e. \( \|u\| = 1 \), we refer to as the shortcut perturbation) and \( x^* = x + \epsilon u \), then, subject to first order approximation:

\[ \Phi(Z|X = x) = \frac{E_z \left[ |p_Z|X=x^*_z(z) - p_Z|X=x(z) |^2 \right]}{\epsilon^2 \cos^2 \psi} + V \left[ \left\| \nabla_z \log p_Z|X=X(z) \right\|_2 \right] \tag{5} \]

where \( E_z[v] \) and \( V[v] \) are the expectation and variance of \( v \), and \( \psi \) is the angle between \( u \) and \( \nabla_z p_Z|X=x \).

Proof is given in the supplementary material.

The first term in the RHS of Eq. (5) will be minimized when the divergence (distance) between the two distributions, \( p_Z|X=x \) and \( p_Z|X=x+\epsilon u \), is reduced. There are many popular divergence measures between distributions, such as Kullback-Leibler divergence, Jensen-Shannon divergence, Total Variation, the Wasserstein distance, etc. In this work, we choose the Wasserstein distance: as the distributions \( p_Z|X=x \) and \( p_Z|X=x+\epsilon u \) may not have common supports and it leads to a simpler loss function.

In the case of a deterministic feature extractor, which is common in stereo matching networks, the distributions \( p_Z|X=x \) and \( p_Z|X=x^* \) can be seen as two degenerate distributions (i.e. Dirac delta distributions) located at points \( z = f_0(x) \) and \( z^* = f_0(x^*) \). Furthermore, the \( V[\cdot] \) in Eq. (5) will be zero. In this case, the Wasserstein-\( p \) distance can be simplified as:

\[ W_p(p_Z|X=x^*,p_Z|X=x) = \left( \|z^* - z\|_p^p \right)^{1/p}. \tag{6} \]

Using the above insights, we can see that minimizing \( \|z^* - z\|_p \) is a step towards minimizing \( \Phi(Z|X=x) \) (for \( p = 1 \)). Thus, we propose to promote the learning of robust and shortcut-invariant features in stereo matching networks, by optimizing the overall loss function defined below:

\[ L = L_{smooth L_1}(\hat{y}, y) + \frac{\lambda}{2} \left( L_{FI}(z_i, \hat{z}_i) + L_{FI}(z_i, \hat{z}_i) \right) \tag{7} \]

where \( \hat{y} \) and \( y \) are the estimated and ground-truth disparity maps. \( L_{FI} \) is our proposed Fisher information loss function defined as:

\[ L_{FI} = \sum_{i=1}^n \left\| z^{(i)} - \hat{z}^{(i)} \right\|_2 \tag{8} \]

and \( L_{smooth L_1} \) is the smooth-L1 loss function commonly employed for optimizing stereo matching networks \([5, 15, 51, 52]\).

### 3.3.3 Shortcut Perturbation (SCP)

In order to compute \( L_{FI} \), we need to define \( u \) (refer as shortcut perturbation and is introduced in Lemma 3.1): \( u = \frac{\nabla_z z^{(i)}}{\|\nabla_z z^{(i)}\|_2} \) where \( \nabla_z z^{(i)} \) is the gradient of the extracted features \( z \) with respect to input. The shortcut-perturbed image can then be expressed as:

\[ x^{(i)} = x^{(i)} + \epsilon \frac{\nabla_z z^{(i)}}{\|\nabla_z z^{(i)}\|_2} \tag{9} \]

The above perturbation will put more weight on pixels that are sensitive to changes in the input. Intuitively, pixels with large absolute value of \( \nabla_z z \) will have significant impact in altering the statistics of encoded latent distributions and the extracted latent feature representations. Moreover, these pixels are also likely to include shortcuts as shortcuts are highly sensitive to perturbations of the input \([11]\).

To examine the accuracy of the above approximations, we trained the digit recognition network of our toy experiment with the proposed SCP and \( L_{FI} \) (ITSA). As the proposed method is specifically designed for domain generalization, our method can effectively generalize the network.
to unseen domains and achieve better performance (4\%) than the robust information bottleneck as shown in Tab. 1.

4. Experiments

4.1. Experimental Settings

Datasets and Metrics: Scene Flow [24] is a large collection of synthetic stereo images with dense disparity ground truth. It contains FlyingThings3D, Driving and Monkaa subsets, and provides 35,454 training and 4,370 testing images. In our experiments, all stereo matching networks are trained on the Scene Flow dataset only.

The realistic datasets used in our experiments include KITTI2012 [10] and KITTI2015 [25] containing 193 and 200 stereo images of outdoor driving scenes, Middlebury [34] containing 15 images of high resolution indoor scenes, and ETH3D [35] containing 27 low resolution, greyscale stereo images of both indoor and outdoor scenes. Furthermore, datasets covering different weather conditions provided by the DrivingStereo [46] dataset, and night-time provided by Oxford Robotcar [23]) were also included to evaluate the robustness of our proposed method. All the above datasets come with with sparse ground truth.

We evaluated the performance of disparity estimation using the D1 error rate (\%), with different pixel threshold. The D1 metric computes the percentage of bad pixels (disparity end-point error larger than the threshold) in the left frame. Following the advice of data originators, a threshold of 3 pixels is selected for KITTI and DrivingStereo, 2 pixels for Middlebury, and 1 pixel for ETH3D.

Baseline networks for our experiments. We have selected these networks namely PSMNet [5], GwcNet [15] and CFNet [36] as the baseline networks for our experiments. We have selected these networks mainly due to the fact that PSMNet and GwcNet are well-studied, and commonly employed as a baseline in many prior works [43, 49, 53]; and CFNet is one of the recently proposed state-of-the-art stereo matching networks. The networks are implemented using PyTorch framework and are trained end-to-end with Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) optimizer. Similar to the original implementations of the selected networks, our data processing includes color normalization and random cropping the input images to size $H = 256$ and $W = 512$. Following the original implementation of CFNet, asymmetric chromatic augmentation and asymmetric occlusion [45] are also employed for data augmentation in CFNet. The maximum disparity for PSMNet and GwcNet is set to 192, and for CFNet is set to 256. All models are trained from scratch for 20 epochs with learning rate set to 0.001 for the first 10 epochs and decreased by half for another 10 epochs. The batch size is set to 12 for training on 2 NVIDIA RTX 8000 Quadro GPUs. The models are trained using synthetic data only and directly tested using data from different realistic datasets. For all experiments included in the following sections, the hyper-parameters $\lambda$ and $\epsilon$ were set to 0.1 and 0.5 respectively. The hyper-parameter tuning experiments are detailed in the supplementary document. The code of our implementations is available at: https://anonymous.4open.science/r/ITSA-D281

4.2. Shortcuts in stereo matching networks

Our hypothesis is that the baseline stereo matching networks naively trained on synthetic data only, learn to exploit common artefacts of synthetic stereo images as shortcut features. These artefacts include (1) consistent local statistics (RGB color features) between between the left and right stereo images and (2) over-reliance on local chromaticity features of the reference stereo viewpoint.

To empirically verify the above, we tested three baseline networks trained only with synthetic data (i.e. Scene flow), using augmented stereo inputs images. The augmented stereo images were derived from the Scene Flow

| Inputs         | No Augment (%) | PSMNet [5] | GwcNet [15] | CFNet [36] |
|----------------|---------------|------------|-------------|------------|
| ACJ            | 13.98         | 3.13       | 1.34        |
| GrayScale (L)  | 37.68         | 8.41       | 1.32        |
| GrayScale (R)  | 9.82          | 2.25       | 1.09        |
| SCP            | 5.84          | 2.90       | 2.55        |

Table 2. Analysis of the effect of data augmentation on the performance of stereo matching networks. All networks are only trained on the Scene Flow training set and the EPE metric is employed for evaluation. The results show that removing shortcut related artefacts (by data augmentation) negatively impact the performance of these networks. In particular, our proposed augmentation can even significantly impact robust methods (e.g. CFNet).
Table 3. Ablation results on PSMNet [5] and GwcNet [15]. SCP is the proposed shortcut perturbations and $L_{FI}$ is the proposed loss function in Eq. (7). The D1 metric was used for evaluation.

| SCP | $L_{FI}$ | PSMNet KITTI-2012 | GwcNet KITTI-2015 | PSMNet KITTI-2015 | GwcNet KITTI-2015 |
|-----|----------|--------------------|-------------------|-------------------|-------------------|
| ✓   | ✓        | 27.4               | 11.7              | 29.3              | 12.8              |
| ✓   | ✓        | 8.1                | 5.3               | 8.6               | 5.9               |
| ✓   | ✓        | 5.2                | 4.9               | 5.8               | 5.4               |

Table 4. Synthetic-to-realistic domain generalization evaluation using KITTI, Middlebury and ETH3D training sets. All methods are trained on the Scene Flow dataset and directly tested on the three real datasets. Pixel error rate with different threshold are employed: KITTI 3-pixel, Middlebury 2-pixel and ETH3D 1-pixel.

| Methods          | KITTI 2012 | KITTI 2015 | Middlebury Full | Middlebury Half | Middlebury Quarter | ETH3D 4.9 |
|------------------|------------|------------|-----------------|-----------------|---------------------|----------|
| HD* [48]         | 23.6       | 26.5       | 50.3            | 37.9            | 20.3                | 54.2     |
| PSMNet [5]       | 27.4       | 29.3       | 60.4            | 29.1            | 19.6                | 16.1     |
| GwcNet [15]      | 11.7       | 12.8       | 45.5            | 18.1            | 10.9                | 9.0      |
| CasStereo [14]   | 11.8       | 11.9       | 40.6            | -               | -                   | 7.8      |
| GANet [51]       | 10.1       | 11.7       | 32.2            | 20.3            | 11.2                | 14.1     |
| DSMNet [52]      | 6.2        | 6.5        | 21.8            | 13.8            | 8.1                 | 6.2      |
| CFNet [50]       | 4.7        | 5.8        | 28.2            | 13.5            | 9.4                 | 5.8      |
| ITSA-PSMNet      | 5.2        | 5.8        | 28.4            | 12.7            | 9.6                 | 9.8      |
| ITSA-GwcNet      | 4.9        | 5.4        | 26.8            | 11.4            | 9.3                 | 7.1      |
| ITSA-CFNet       | 4.2        | 4.7        | 20.7            | 10.4            | 8.5                 | 5.1      |

4.4. Synthetic-to-Realistic Domain Generalization Evaluation

In Tab. 4, we compare the synthetic-to-realistic domain generalization performance of our method with the state-of-the-art stereo matching networks [5, 14, 15, 36, 48, 51, 52] on the four realistic datasets. All networks are trained on the synthetic Scene Flow training set only. We found that the proposed ITSA substantially improved the domain generalization performance (6.8% − 23.5%) of the selected stereo networks (PSMNet [5] and GwcNet [15]), outperforming the state-of-the-art stereo matching networks in the realistic datasets. The improved networks also outperform DSMNet [52] on the KITTI 2012 [10] and KITTI 2015 [25] datasets, and achieve comparable performance as the CFNet on the Middlebury [34] datasets. In addition, we show that ITSA is even capable of further enhancing the robustness and cross-domain performance of CFNet [36], which was the best performing stereo matching networks in the Robust Vision Challenge 2020. Comparison of qualitative results generated by the baseline networks and our methods are included in Fig. 4.

4.5. Robustness to Anomalous Scenarios

Here, we analyze the robustness to anomalous conditions of a network trained on synthetic data with the proposed ITSA. The anomalous conditions include night-time, foggy and rainy weather conditions. In this comparison, we train the same network twice: (1) pre-train using synthetic data followed by fine-tuning on realistic KITTI 2015 dataset (common strategy), (2) train only using synthetic data with the proposed SCP and $L_{FI}$ (ITSA). We also included the pre-trained counterpart of CFNet [36] to illustrate the efficacy of our method in further enhancing the network robustness.

In Tab. 5, we show that the fine-tuned (FT) networks generally has better performance when tested on data similar to...
the KITTI training data (sunny and cloudy). In contrast, our method (ITSA) consistently enhances the robustness of selected stereo matching networks and outperforms the fine-tuned (FT) models in the real-world anomalous scenarios including rainy and foggy weather and night-time. The performances were evaluated using the D1 metric.

5. Limitations

Although the proposed method can significantly improve the performance of stereo matching networks without fine-tuning and even outperform their fine-tuned counterpart when tested in unseen challenging environments (e.g. rain and night-time), its performance remains fragile under extreme conditions (e.g. heavy rain and extreme low light), which may occur in real scenarios. This is reflected by the large errors reported in Tab. 5. By looking at samples with large errors, we noticed that those inaccuracies are largely due to having insufficient light source, lens glare/flares and reflection on specular surfaces (wet grounds). In our future work, we aim to address these issues and develop a robust system that can handle these extreme scenarios in real-world applications.

6. Conclusion

In this work, we have presented ITSA: a novel information theory-based approach for domain generalization in stereo matching networks. To address the shortcut learning challenge, we propose to minimize the sensitivity of the extracted feature representations to the input perturbations, measured via the Fisher information. We further proposed an efficient algorithm to optimize the Fisher information objective. Experimental results show that the proposed method consistently promotes the learning of robust and shortcut-invariant features, and substantially enhances the performance of existing stereo matching networks in cross-domain generalization, even outperforming their fine-tuned counterparts in challenging scenarios. We also show that the proposed method can be easily extended for non-geometry based vision problems such as semantic segmentation.
References

[1] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In ICLR, 2017.
[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In Proceedings of the European Conference on Computer Vision (ECCV), September 2018.
[3] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2229–2238, 2019.
[4] Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, and Henry Fuchs. Stereodrmn: Dilated residual stereonet. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 11786–11795, 2019.
[5] 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.
[6] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11580–11590, 2021.
[7] 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.
[8] Nikolay Dagaev, Brett D Roads, Xiaoliang Luo, Daniel N Barry, Kaustubh R Patil, and Bradley C Love. A too-good-to-be-true prior to reduce shortcut reliance. arXiv preprint arXiv:2102.06406, 2021.
[9] 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.
[10] 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.
[11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. arXiv preprint arXiv:2004.07780, 2020.
[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11):665–673, 2020.
[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231, 2018.
[14] 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, pages 2495–2504, 2020.
[15] 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.
[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations, 2019.
[17] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision, pages 66–75, 2017.
[18] 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.
[19] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5400–5409, 2018.
[20] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2811–2820, 2018.
[21] Rui Liu, Chengxi Yang, Wexiu 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, pages 12757–12766, 2020.
[22] 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.
[23] 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):3–17, 2017.
[24] Nikolaus Mayer, Eddy Ilg, Philipp 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4040–4048, 2016.
[25] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. ISPRS Journal of Photogrammetry and Remote Sensing (JPRS), 2018.
[50] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2100–2110, 2019. 8

[51] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 185–194, 2019. 1, 3, 5, 7

[52] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In European Conference on Computer Vision, pages 420–439. Springer, 2020. 2, 3, 5, 7

[53] Youmin Zhang, Yimin Chen, Xiao Bai, Jun Zhou, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. arXiv preprint arXiv:1909.03751, 2019. 6