An Information-Theoretic Method to Automatic Shortcut Avoidance and Domain Generalization for Dense Prediction Tasks

WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, David Suter, and Alireza Bab-Hadiashar, *Senior Member, IEEE*

Abstract—Deep convolutional neural networks for dense prediction tasks are commonly optimized using synthetic data, as generating pixel-wise annotations for real-world data is laborious. However, the synthetically trained models do not generalize well to real-world environments. This poor "synthetic to real" (S2R) generalization we address through the lens of shortcut learning. We demonstrate that the learning of feature representations in deep convolutional networks is heavily influenced by synthetic data artifacts (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. Specifically, our proposed method minimizes the sensitivity of latent features to input variations: to regularize the learning of robust and shortcut-invariant features in synthetically trained models. To avoid the prohibitive computational cost of direct input sensitivity optimization, we propose a practical yet feasible algorithm to achieve robustness. Our results show that the proposed method can effectively improve S2R generalization in multiple distinct dense prediction tasks, such as stereo matching, optical flow, and semantic segmentation. Importantly, the proposed method enhances the robustness of the synthetically trained networks and outperforms their fine-tuned counterparts (on real data) for challenging out-of-domain applications.

Index Terms—Domain Generalization, Shortcut Learning, Dense Prediction Tasks, Stereo Matching, Semantic Segmentation, Optical Flow.

1 Introduction

Deep convolutional neural networks (DCNNs) have achieved state-of-the-art performance in many *dense prediction tasks* such as semantic segmentation [1], [2], [3], motion correspondences estimation (stereo matching [4], [5], [6] and optical flow [7], [8], [9], [10]). Despite their successes, these top performing models usually require large amounts of labelled data samples for training. As such, their performance strongly depends on the availability of labelled training data. However, the process of generating pixel-wise annotated ground-truth for dense prediction applications is both challenging and expensive. Furthermore, it is certainly infeasible to generate labelled data sufficient to cover all scenarios of the real-world (e.g. different weathers, seasons, towns, day/night, urban/rural, etc). Meanwhile, synthetic data, generated using game engines, offers an alternative solution for training DCNNs for dense prediction applications. As different scenarios can be simulated in game engines, large amounts of densely annotated synthetic data can be easily generated at a much lower cost.

On the other hand, synthetically generated images unavoidably have some visual artefacts such as unrealistic textures, fake appearances, simplified lighting conditions and unrealistic scene layouts [11]. As such, models that are trained on synthetic data often show poor generalization on the real-world domain, due to the existence of the minor visual differences between synthetic and real images (see Figure 1). For example, under loose constraints, convolutional neural networks (CNNs) tend to be biased towards background [12] and local textures [13], [14], instead of learning the true concepts (e.g. shapes and image context) from the training data. Geihos *et al.* [15] coined the process of learning these biases from the training data as shortcut learning.

While several methods have been previously proposed [14], [16], [17] to mitigate shortcut learning, these methods are manually designed and rely on the assumption that the shortcuts can be identified in advance. However, shortcut cues exploited by CNNs can be non-intuitive, task-specific, and difficult to identify [18], [19]. Therefore, a universal approach that can effectively mitigate shortcut learning for different tasks and network architectures is desirable.

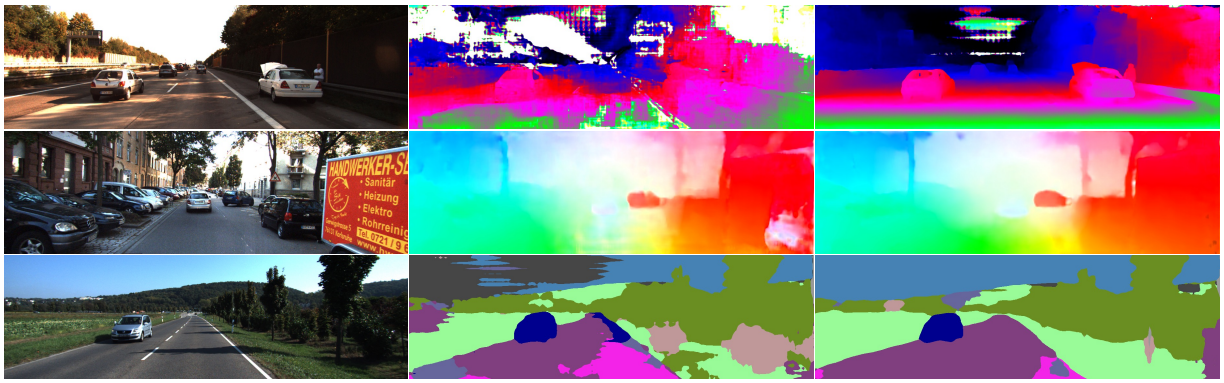
A straightforward method to avoid the learning of undesirable attributes (e.g. shortcuts) from the input data is to restrict the amount of input-relevant information encoded in the feature representation. This goal can be achieved by using the Information Bottleneck (IB) principle [20] by seeking parameters θ that optimize the following objective:

$$\arg\max_{\theta} I\left(Y, Z; \theta\right) - \beta I\left(X, Z; \theta\right) \tag{1}$$

where Z is the representation of input X, Y is the target label, $I(\cdot)$ is mutual information and $\beta \in [0, 1]$ is the hyper-

- *W. Chuah, R. Hoseinnezhad and A. Bab-Hadiashar are with the School of Engineering, RMIT University, Melbourne, Australia. Email: wei.qin.chuah@student.rmit.edu.au, {rezah,abh}@rmit.edu.au.*
- *R. Tennakoon is with School of Science, RMIT University, Melbourne, Australia. Email: ruwan.tennakoon@rmit.edu.au.*
- *D. Suter is with School of Science, Edith Cowan University, Joondalup, WA 6027, Australia. Email: d.suter@ecu.edu.au*

Manuscript received April 19, 2005; revised August 26, 2015.



(a) Input (b) Baseline (c) ITSA

Fig. 1: Synthetic-to-realistic (S2R) performance comparison for different dense prediction tasks: stereo matching (top row), optical flow (middle row) and semantic segmentation (bottom row). Baseline task networks that are naively trained on synthetic data only (b) failed to generalize to unseen realistic domain. The proposed ITSA method can substantially improve S2R performance in different tasks, despite training using synthetic data only.

parameter that controls the size of information bottleneck. The IB principle regularizes the learning of compressive yet expressive feature representation Z, by allowing the network to exploit only the minimum amount of input information required for accurate predictions. While the IB principle has been proven to be effective in generalizing the network between IID training and testing data, our experiments (see Section 3.2) showed that IB does not promote the generalization to out-of-domain data. This insight strongly suggests that these compressed features, constrained by the IB; are neither robust nor shortcut-invariant. As a result, the IB optimized networks still suffer from shortcut learning, and fail to generalize to unseen domains.

Recently, Pensia *et al.* [21] introduced robust IB, which is built upon the original IB principle. Robust IB was proposed to improve a networks' adversarial robustness, by minimizing the sensitivity of the learned feature representations (against small perturbation in the input space). Specifically, robust IB utilizes the Fisher information of the learned features (parameterized by the inputs), to measure the sensitivity of the features with respect to a perturbation in the input space. As a result, robust IB constrained networks are insensitive to small perturbations that are added to the input images (e.g. adversarial attack via Fast Gradient Sign Method (FGSM) [22]). As the learned shortcut-dependent features are also highly sensitive to changes in the input space [15], robust IB offers a promising tool to automatically mitigate shortcut learning and promote synthetic-to-realistic (S2R) domain generalization. To the best of our knowledge, the principle of the robust IB has not been previously explored for domain generalization.

Inspired by this, we propose to minimize the Fisher information, as an additional regularization loss term, to promote S2R domain generalization for stereo matching, optical flow estimation and semantic segmentation networks. However, direct optimization of the Fisher information, using gradient descent, requires the computation of second order derivatives. As such, it would be computationally prohibitive for deep neural networks with many parameters, and tasks that use high dimensional input images (e.g. dense prediction tasks such as segmentation and correspondences estimation). To overcome this problem, we proposed an Information-Theoretic Shortcut Avoidance (ITSA) learning algorithm, 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 and achieve high degrees of S2R domain generalization. Furthermore, ITSA is also task and model agnostic as it can be easily extended to different dense prediction tasks, without network architecture alteration.

This paper is an extension of our previous work [23], where we focused on demonstrating the efficacy of the proposed ITSA on improving S2R domain generalization in stereo matching networks. In this work we specifically:

1) investigate the use of ITSA for more broad set of problems. We show that the proposed ITSA can also substantially improve S2R domain generalization for other dense prediction tasks such as optical flow estimation (Section 4.2) and semantic segmentation (Section 4.3).
2) demonstrate that ITSA can also substantially enhance the robustness of dense prediction networks and perform favourably as compared to its fine-tune counterpart on challenging anomalous scenarios such as rainy weather and night-time (Section 4.3.5).
3) analyze the feature maps extracted by dense prediction networks trained with different settings (Section 5.2). Evidently shown by the results of the analysis, the proposed method ITSA can promote the learning of robust and shortcut-invariant features in dense prediction networks.

The rest of the paper is organized as follows. Section 2 describes the related work in the field of synthetic-to-realistic domain generalization for dense prediction vision tasks and shortcut learning. Section 3 presents the problem statement (of domain generalization in dense prediction tasks) and a motivational toy example: to illustrate the

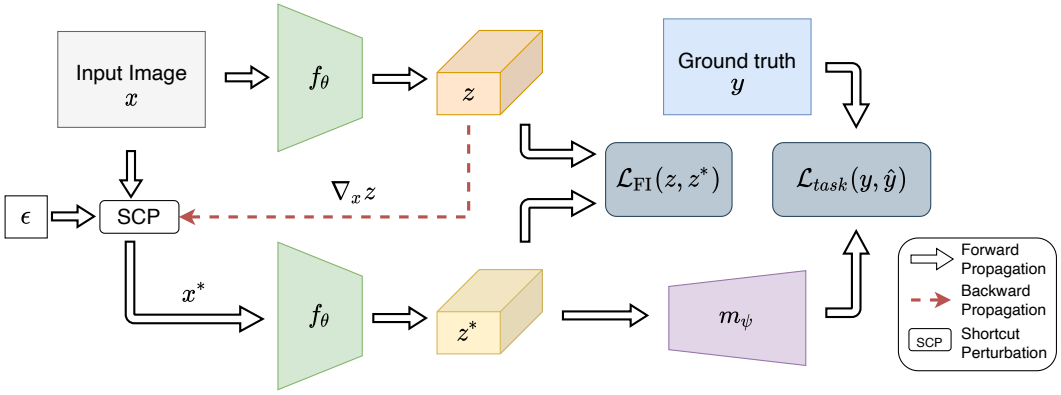


Fig. 2: An overview of the proposed shortcut-avoidance strategy to achieve domain generalization in dense prediction networks. The m_ψ network represents the task network (for example, stacked 3D hourglass in stereo matching networks, classifier in semantic segmentation networks, and convolutional decoder in optical flow networks). Furthermore, the parameters are shared across the two feature extractor networks f_θ (best viewed in color).

limitation of the previously proposed Information Bottleneck principle in promoting domain generalization. This section also includes detail of the proposed method for automatic shortcut avoidance and domain generalization. Experimental results are presented in Section 4. An ablation study of the proposed method is included in Section 5.1, and Section 5.2 provides discussion of the feature maps we derived from our models. Section 6 concludes the paper.

2 Related Work

2.1 Synthetic-to-Realistic Domain Generalization for Dense Prediction Vision Tasks

In this paper, we demonstrate an approach to promote S2R generalization across three application domains. The first two (stereo matching and optical flow) essentially depend on correspondence estimation (2.1.1); and the third is semantic segmentation (2.1.2).

2.1.1 Correspondence Estimation

In recent years, end-to-end deep neural networks designed for correspondence estimation such as stereo matching [4], [24], [25], [26] and optical flow [7], [8], [9], [27] have excelled in most datasets and benchmarks. These networks are primarily trained from scratch, using synthetic datasets generated via game simulation. However, without fine-tuning the synthetically pre-trained networks on real-world data, these networks often do not generalize to unseen realistic environments.

To promote domain generalization in stereo matching networks, Shen *et al.* [5] introduced CFNet, an efficient network architecture with multi-scale cost volume fusion and refinement, to enforce the learning of robust and domain-invariant structural representation for stereo matching. Similarly, Zhang *et al.* [28] proposed DSMNet, which employs Domain Normalization and non-local graph-based filtering layers to enforce the learning of structural features that are domain-invariant. Moreover, DSMNet can also be extended to improve the domain generalization performance in optical flow networks.

In contrast, we have identified shortcut learning [29] 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 superior performance on challenging realistic data without fine-tuning. Furthermore, unlike DSMNet, our method can be easily used to improve the domain generalization performance of optical flow networks without altering the network architecture. (Indeed, our ITSA approach can likewise be employed in stereo matching and semantic segmentation, without altering the network architecture.)

(a) MNIST (b) MNIST-M (c) SYN (d) SVHN (e) USPS

Fig. 3: Examples of five different domains in the Digits-DG dataset. The five domains include MNIST, MNIST-M, SVHN, SYN and USPS. Each image in these datasets contains one digit only.

2.1.2 Semantic Segmentation

The topic of synthetic-to-realistic domain generalization for semantic segmentation has been largely under-explored, as only a few studies have been made previously. In general, the previously proposed methods can be categorized as feature normalization approaches [30], [31] and style transfer approaches [32], [33]. The feature normalization approaches aim to remove domain-specific information embedded in the statistics of the feature representation (e.g. feature covariance [31] or feature mean and variance [30]), using whitening [34] or standardization [35], [36] techniques.

On the other hand, the style transfer approaches seek to promote the learning of domain-invariant features by randomizing the appearance and texture of the synthetic training images. For instance, Yue *et al.* proposed DRPC [33] that utilized style transfer to randomize the appearance of its synthetic data, using ImageNet [37] training images. Furthermore, DRPC enforces consistency between the features learned from the synthetic images and the augmented images to promote the learning of domain-invariant features. Meanwhile, Peng *et al.* proposed GLTR [32] that promotes the learning of domain-invariant features by alleviating the texture bias in synthetically trained semantic segmentation networks. This is achieved by randomizing the texture of the synthetic training images, both globally and locally, using abstract art paintings sampled from the "Painter by Numbers" dataset.

In comparison to existing approaches, our method does not include auxiliary datasets for appearance or texture augmentation. Also, the proposed method does not involve feature normalization techniques, which could remove important information from the feature representations [38]. Instead, our method aims to automatically avoid shortcut learning in synthetically trained semantic segmentation networks and promote the learning of robust and shortcut-invariant features to improve the domain generalization performance.

2.2 Shortcut Learning

Geirhos *et al.* [15] 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 [12], [15] and textures [13], [39] to improve their performance. However, these networks fail to generalize to unseen domains, where the spurious correlations between shortcuts and labels are violated [40]. 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 [14], [16], [31], [39], [41]. These methods rely on having some shortcut-related prior knowledge and usually include data augmentations [16], [41] or dropout-based regularization [14] as part of their solutions. However, shortcuts are non-trivial, task-specific, and are often difficult to identify a priori [18], [19]. In contrast, our proposed method automatically avoids shortcut learning without requiring shortcut-related knowledge in advance.

3 Methodology

3.1 Problem Statement

In this work, we focus on the synthetic-to-realistic domain generalization for multiple vision tasks. Without considering any specific task, given a synthetic dataset D_{syn} with N number of densely annotated samples $\{x^i_{syn}, y^i_{syn}\}^N_{i=1}$, the goal is to obtain robust task networks (e.g. DCNNs) that are domain-invariant and can generate accurate estimates $\hat{y}^{(i)}$ for unseen realistic environments D_{real}.

To this end, we aim to prevent the task networks from exploiting spurious shortcut features from the synthetic input images. As mentioned in Section 1, the Information Bottleneck (IB) principle limits the amount of input-relevant information encoded in the feature representation. Consequently, the IB-constrained networks learn to extract compressive yet relevant features for predictions. As such, the IB principle may be seen as a natural choice to achieve our objective: Avoiding shortcut learning from the synthetic data. However, it was found that the IB principle is not robust to small input perturbations (e.g. FGSM adversarial attack) [21]. As a result, the compact features extracted by the IB-constrained network remain fragile to such perturbation. Moreover, as shortcut cues exploited by DCNNs are also highly sensitive to input perturbations (e.g. data augmentations) [15], we conjecture that the IB-constrained networks would also include the spurious shortcut cues in the learned compact feature representations for prediction, and harming the domain generalization performance.

To resolve this issue, a robust version of this method, called the robust IB [21], is proposed, which minimizes the sensitivity of the features with respect to the input variations. As a result, robust IB can theoretically avoid the learning of spurious shortcut cues from synthetic data, and offers a promising approach to improve synthetic-to-realistic domain generalization performance. We demonstrate the domain generalization effect of IB approach, and its robust variant called robust IB (RIB), using a toy example included in the next section.

3.2 Motivation: Toy Example

In this toy experiment, we investigate the efficacy of the Information Bottleneck (IB) principle [20], and its robust counterpart (RIB) [21]; in improving the performance of domain generalization. As computing the mutual information in IB is computationally intractable for deep neural networks, we employed the deep variational information bottleneck (VIB) [42] to construct a lower bound on the IB objective in Equation (1). The implementation details of this toy experiment are included in Section 1 of our supplementary document. We choose the digit recognition (DR) task for this purpose. This is because DR only requires Convolutional Neural Networks (CNNs) with substantially lesser number of trainable parameters as compared to other computer vision tasks, which makes implementing RIB possible. To evaluate the performance of domain generalization for DR, we employ the commonly used Digits-DG dataset. This dataset consists of five hand-written digit datasets, namely MNIST [43], MNIST-M [44], SYN [44], SVHN [45] and USPS [46], where each subset can be regarded as a different domain. In our experiments, we follow the common practice of referring to MNIST as the source domain and the others as the unseen target domains.

We hypothesize that the trained model should be able to generalize its performance from source domain to unseen target domains if the model learned to avoid using spurious correlations (shortcuts). As shown in Table 1, standard

TABLE 1: Performance comparison of digit recognition networks optimized via empirical risk minimization (ERM), the variational information bottleneck (VIB), its robust variant (RIB) and our proposed ITSA. While VIB performs well in the in-domain tests, it performs poorly on out-of-domain tests. Top-1 accuracy (%) is reported.

Methods	MNIST	MNIST-M	SHVN	SYN	USPS	Avg
Baseline (ERM)	97.9 ± 0.17	28.5 ± 4.14	13.0 ± 1.28	14.4 ± 1.41	72.9 ± 2.96	45.4 ± 1.07
VIB	**99.1** ± 0.17	15.2 ± 0.72	10.1 ± 0.03	10.7 ± 0.27	**79.5** ± 3.09	42.9 ± 0.50
RIB	98.9 ± 0.41	46.0 ± 2.86	18.1 ± 3.20	18.1 ± 1.12	75.1 ± 2.27	51.3 ± 0.81
ITSA	98.6 ± 0.24	**51.4** ± 0.20	**24.1** ± 2.99	**21.2** ± 1.56	74.6 ± 1.08	**54.0** ± 0.62

TABLE 2: Comparison of GPU memory requirement and training time per iteration for the digital recognition networks. The batch size was set to 12. The RIB [21] has significantly higher GPU memory requirement and processing time as compared to other counterparts.

Methods	Time/Iter (s)	GPU Memory (MB)
Baseline (ERM)	0.008	1,223
VIB	0.011	1,227
RIB	13.229	19,325
ITSA	0.018	1,227

IB (VIB) can effectively reduce overfitting and achieves the best performance in the source domain (MNIST). Interestingly, VIB also achieves the best performance in the USPS dataset. We conjecture that VIB promotes the encoding of domain-specific shortcuts in the compressed feature representations. Consequently, the IB-optimized networks perform impressively on in-domain datasets (e.g. MNIST and USPS) but fail when tested on out-of-domain datasets. We show in Figure 3 that MNIST and USPS datasets can be considered as in-domain, as their data have common characteristics (e.g. white digit with black background). More importantly, VIB even displays worse performance than the baseline method in unseen domains: MNIST-M, SVHN and SYN. This indicates that the standard IB principle is not suited for mitigating shortcut learning (to promote domain generalization performance).

3.3 Robust Information Bottleneck and Fisher Information

As our aim is to develop an IB based cost function that is not susceptible to the existence of shortcuts in source data. We take inspiration from the robust IB method [21] that utilizes (in place of $I(Z,X)$) the statistical Fisher information $\Phi(Z|X)$ of the extracted features Z parameterized by the inputs X as a more robust measure of information. Fisher information $\Phi(Z|X)$ is defined as:

$$\Phi(Z|X) = \int_{\mathcal{X}} \Phi(Z|X=x) p_X(x) dx, \tag{2}$$

where

$$\Phi(Z|X=x) = \int_{\mathcal{Z}} \left\| \nabla_x \log p_{Z|X}(z|x) \right\|_2^2 p_{Z|X}(z|x) dz. \tag{3}$$

The term $\Phi(Z|X=x)$ 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 [15]. As such, minimizing the Fisher information is a step towards promoting the learning of shortcut-invariant features. This conjecture is supported by the results of the toy experiment (Table 1). The digit recognition models, regularized by Fisher information (RIB), achieve significantly better performance than the baseline and standard IB networks in the target domains.

In order to minimize the Fisher information (Equation (3)), one has to compute second order derivatives such as $\nabla_\theta \nabla_x \log p_{Z|X}(z|x)$, which is computationally prohibitive for dense prediction tasks that requires large dimensional inputs [47]. We show in Table 2 that RIB demands significant training time and GPU memory consumption as compared to the baseline and VIB approaches. To overcome this issue, we propose ITSA, a simple yet computationally feasible approach to promote the learning of shortcut-invariant features. As shown in Table 1 and 2, the proposed ITSA can achieve impressive domain generalization performance in the digit recognition task, with substantially lesser training time and GPU memory consumption than the RIB approach. The high-level overview of our proposed ITSA shortcut avoidance strategy is depicted in Figure 2.

3.4 Approximating Fisher information

Optimizing the Fisher information $\Phi\left(Z \mid X\right)$ measure (Equation (2)) is related to minimizing $\Phi\left(Z \mid X=x\right)$. By adding a regularization term such as $\Phi\left(Z \mid X=x\right)$ to the loss function, we can penalize the transient features and discourage 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:*

$$\Phi\left(Z \mid X=x\right) = \frac{\mathbb{E}_z\left[\left|p_{Z|X=x^*}(z) - p_{Z|X=x}(z)\right|\right]^2}{\epsilon^2 \cos^2 \psi} + \mathcal{V}\left[\left\|\nabla_x \log p_{Z|X=x}(z)\right\|_2\right] \tag{4}$$

where $\mathbb{E}_z[\upsilon]$ and $\mathcal{V}[\upsilon]$ are the expectation and variance of υ, and ψ is the angle between u and $\nabla_x p_{Z|X=x}$.

1. We use transient to describe image attributes that are inconsistent across domains, and spuriously correlated with the true label. These features may include backgrounds, textures, image style, etc.

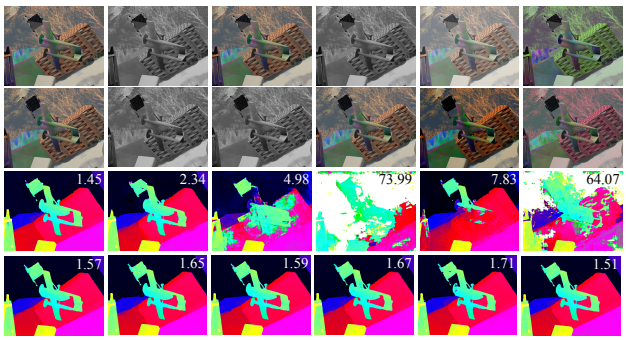


Fig. 4: Examples of shortcuts in stereo matching networks. The left and right input images are included in the top two rows. The disparity maps estimated by the baseline PSMNet [24] are included in the third row and ITSA-PSMNet in the bottom row. The performance of the baseline PSMNet deteriorates substantially when the shortcut attributes are distorted or removed from the input stereo images. The corresponding EPE is displayed on the estimated disparity map. Best viewed in color and zoom in for details.

The proof is given in the supplementary document (see Section 2).

The first term in the RHS of Equation (4) 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 $z = f_\theta(x)$ and $z^* = f_\theta(x^*)$. Furthermore, the $\mathcal{V}[\cdot]$ in Equation (4) will be zero. In this case, the Wasserstein-p distance can be simplified as:

$$W_p(p_{Z|X=x^*}, p_{Z|X=x}) = \left(\|z^* - z\|_2^p\right)^{1/p}. \tag{5}$$

Using the above insights, we can see that minimizing $\|z^* - z\|_2$ is a step towards minimizing $\Phi(Z \mid X = x)$ (for $p = 1$). Thus, we propose to promote the learning of robust and shortcut-invariant features in dense prediction networks (e.g. stereo matching, optical flow and semantic segmentation networks), by optimizing the overall loss function defined below:

$$\mathcal{L} = \mathcal{L}_{task}(\hat{y}, y) + \lambda \mathcal{L}_{\text{FI}}(z, z^*) \tag{6}$$

where $\hat{y}$ and y are the estimated and ground-truth disparity maps, $\mathcal{L}_{\text{FI}}$ is our proposed Fisher information loss function defined as:

$$\mathcal{L}_{\text{FI}} = \sum_{i=1}^{n} \left\| z^{(i)} - z^{*(i)} \right\|_2 \tag{7}$$

and $\mathcal{L}_{task}$ is the task-specific loss function. For example, the distance-based loss function (e.g. MAE, MSE, smooth-L1) is commonly employed for optimizing pixel-wise regression networks such as stereo matching networks and optical flow networks. Meanwhile, the cross-entropy loss is commonly used to train pixel-wise classification networks such as semantic segmentation networks.

TABLE 3: 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 (lower value indicates better performance). 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).

Inputs	PSMNet [24]	GwcNet [48]	CFNet [5]
No Aug (X)	1.38	0.85	1.00
ACJ	13.98	3.13	1.34
GrayScale (X_L)	37.68	8.41	1.32
GrayScale (X_R)	9.82	2.25	1.09
SCP ($\epsilon = 0.5$)	5.84	2.90	2.55

3.5 Shortcut Perturbation (SCP)

In order to compute $\mathcal{L}_{\text{FI}}$, we need to define u (referred to as shortcut perturbation, introduced in Lemma 3.1):

$$u = \frac{\nabla_x z^{(i)}}{\left\|\nabla_x z^{(i)}\right\|_2} \tag{8}$$

where $\nabla_x z^{(i)}$ is the gradient of the extracted features z with respect to input. The shortcut-perturbed image can then be expressed as:

$$x^{*(i)} = x^{(i)} + \epsilon \frac{\nabla_x z^{(i)}}{\left\|\nabla_x z^{(i)}\right\|_2} \tag{9}$$

The above perturbation will put more weight on pixels that are sensitive to changes in the input. Intuitively, pixels with large absolute value of $\nabla_x z$ 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 [15]. In other words, the role of SCP is to distort the shortcut information presented in the input images in order to suppress the learning of shortcut features. We show in Sections 4.1.3, 4.2.3, and 4.3.3 that SCP augments the shortcut artefacts in the synthetic images and substantially deteriorates the performance of baseline networks that leveraged shortcut features for predictions.

To examine the accuracy of the above approximations, we trained the digit recognition network of our toy experiment with the proposed SCP and $\mathcal{L}_{\text{FI}}$ (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 (2.7%) than the robust information bottleneck as shown in Table 1.

TABLE 4: Synthetic-to-realistic domain generalization evaluation of stereo matching networks, 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 (lower value indicates better performance).

Methods	KITTI		Middlebury			ETH3D
	2012	2015	Full	Half	Quarter	
HD^3 [49]	23.6	26.5	50.3	37.9	20.3	54.2
PSMNet [24]	27.4	29.3	60.4	29.1	19.6	16.1
GwcNet [48]	11.7	12.8	45.5	18.1	10.9	9.0
CasStereo [50]	11.8	11.9	40.6	-	-	7.8
GANet [4]	10.1	11.7	32.2	20.3	11.2	14.1
MS-PSMNet [51]	14.0	7.8	-	19.8	-	16.8
DSMNet [28]	6.2	6.5	21.8	13.8	**8.1**	6.2
MS-GCNet [51]	5.5	6.2	-	18.5	-	8.8
CFNet [5]	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 EXPERIMENTS

To demonstrate the efficacy of our proposed ITSA method in improving the synthetic-to-realistic domain generalization performance for dense prediction tasks, we have included three different dense prediction tasks in our experiments. These tasks are stereo matching and disparity estimation, optical flow estimation and semantic segmentation. For each task, the dataset description, implementation details, shortcut analysis, cross-domain generalization performance comparison and network robustness analysis are included in the next section.

4.1 Stereo Matching and Disparity Estimation

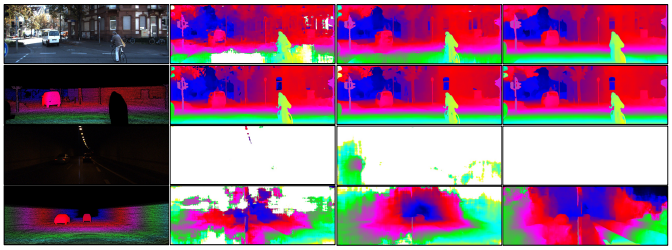
4.1.1 Datasets Description

Synthetic Dataset: *Scene Flow* [52] 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*.

Realistic Dataset: The realistic datasets used in our experiments include *KITTI2012* [53] and *KITTI2015* [54] containing 193 and 200 stereo images of outdoor driving scenes, *Middlebury* [55] containing 15 images of high resolution indoor scenes, and *ETH3D* [56] containing 27 low resolution, greyscale stereo images of both indoor and outdoor scenes. Furthermore, datasets covering different weather conditions provided by the *DrivingStereo* [57] dataset, and night-time provided by *Oxford Robotcar* [58]) were also included to evaluate the robustness of our proposed method. All the above datasets come with sparse ground-truth.

4.1.2 Implementation Details

We have selected three popular and top-performing stereo matching networks namely PSMNet [24], GwcNet [48] and CFNet [5] 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 [59], [60], [61]; 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 (β_1=0.9, β_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 [62] 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 λ and ϵ were set to 0.1 and 0.5, respectively.



(a) Input (b) PSMNet [24] (c) GwcNet [48] (d) CFNet [5]

Fig. 5: Qualitative results on KITTI 2015 stereo data. For each example, the results of the baseline networks are presented on the top row and the results from our method are included in the bottom row. The corresponding left image and ground-truth are included in column (a). Our method can significantly improve the stereo matching performance even in scenario with poor lighting condition. Best viewed in color and zoom in for details.

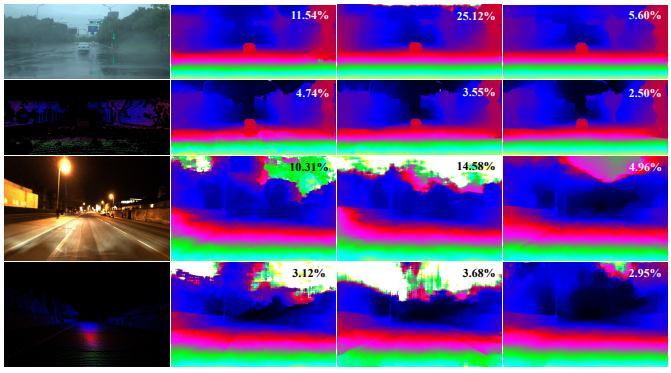
4.1.3 Shortcut Analysis

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 the left and right stereo images and (2) over-reliance on local chromaticity features of the reference stereo viewpoint.

To empirically verify the above conjectures, 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 test set using the following strategies: (1) Chromatic Augmentation (e.g. asymmetrical color jittering (ACJ) [62] and gray scaling) and (2) the shortcut-perturbation (SCP, ex-

TABLE 5: Robustness evaluation on anomalous scenarios. Our method (ITSA) consistently enhances the robustness of selected stereo matching networks and outperform the KITTI fine-tuned (KITTI-FT) models in the real-world anomalous scenarios including rainy and foggy weather and night-time. The performance was evaluated using the D1 metric (lower value indicates better performance).

Models	KITTI-FT	ITSA	Sun	Cloud	Rain	Fog	Night	Avg
PSMNet [24]	✓	✗	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 [48]	✓	✗	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 [5]	✓	✗	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



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

Fig. 6: Qualitative comparison on out-of-distribution data (e.g. rain and night) provided by the DrivingStereo [57] and the Oxford Robotcar [58]. The estimated disparity maps are generated using the PSMNet [24], GwcNet [48] and CFNet [5]. 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 KITTI-2015 [54] fine-tuned networks are included on the top row and the results of our method (ITSA) are included in the bottom row. The corresponding D1 error rate is also included on the predicted disparity map. Our method can significantly improve the performance of these stereo matching networks in challenging unseen domains.

plained in Section 3.5). If a network has learnt to utilize the transient attributes (related to a shortcut), distorting those in the input space will negatively impact its performance. Experimental results, given in Table 3, showed that using the augmented images as inputs has substantially worsened the performance of the stereo matching networks.

Interestingly, the SCP images also deteriorate the performance of the best performing robust stereo matching networks such as CFNet [5]. In Section 4.1.4 and 4.1.5, we show that our method can enhance the robustness of CFNet and significantly improve its performance in unseen realistic environments and anomalous scenarios.

The qualitative results, shown in Figure 4, demonstrate that the performance of the baseline networks (third row) deteriorated significantly when the color features consistency between stereo viewpoints is violated. Moreover, as shown in the fourth column of Figure 4, removing the chromaticity features from the reference image will causes substantial performance reduction in the baseline networks. In contrast, our proposed method reduces the exploitation of shortcut features and shows better robustness to adverse data augmentation scenarios, without using shortcut-related knowledge (see last row of Figure 4).

4.1.4 *Synthetic-to-Realistic Domain Generalization*

Table 4 shows a comparison of the synthetic-to-realistic domain generalization performance of our method with the state-of-the-art stereo matching networks [4], [5], [24], [28], [48], [49], [50] 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 [24] and GwcNet [48]), outperforming the state-of-the-art stereo matching networks in the realistic datasets. The improved networks also outperform DSMNet [28] on the KITTI 2012 [53] and KITTI 2015 [54] datasets, and achieve comparable performance as the CFNet on the Middlebury [55] datasets. In addition, we show that ITSA is even capable of further enhancing the robustness and cross-domain performance of CFNet [5], 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 Figure 5.

4.1.5 *Network Robustness Analysis*

Here, we analyze the robustness to anomalous conditions, including night-time, foggy and rainy weather situations, of a network trained on synthetic data with the proposed ITSA. To compare, 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 $\mathcal{L}_{\text{FI}}$ (ITSA). We also included the pre-trained counterpart of CFNet [5] to illustrate the efficacy of our method in further enhancing the network robustness.

TABLE 6: Analysis of the effect of data augmentation on the performance of optical flow networks, using PwcNet [27] and RAFT [7]. All networks are trained on the synthetic FlyingChairs and FlyingThings train set, and the EPE metric is employed for evaluation on the FlyingChairs test set (lower value indicates better performance). The results show that removing shortcut related artefacts (by data augmentation) negatively impact the performance of these networks. Asterisk ($*$) indicates networks trained with asymmetrical chromatic augmentation included.

Inputs	PwcNet	PwcNet*	RAFT*
No Aug (X)	2.42	2.35	1.12
ACJ	11.1	2.49	1.25
GrayScale (X_t)	12.6	2.51	1.32
GrayScale (X_{t+1})	10.0	2.47	1.20
SCP ($\epsilon = 0.1$)	3.73	3.16	2.59

Table 5 shows 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) can substantially improve the robustness and overall performance of the PSMNet [24] and GwcNet [48], without using the real-world data. The overall performance of fine-tuned CFNet is slightly better than its ITSA counterpart. However, as mentioned earlier, the proposed ITSA improves CFNet performance when only using synthetic data for training. The results demonstrate that our method has effectively improved the robustness and performance of existing stereo matching networks, and extends the use of these networks to real-world applications, without using the real data for fine-tuning.

When the annotated real data is available, 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 [24] and GwcNet [48] 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 [5] in the challenging real-world scenarios (1.75% overall improvement).

4.2 Optical Flow Estimation

4.2.1 Datasets Description

Synthetic Dataset: *FlyingChair* [66] dataset contains 22k image pairs with synthetically generated chair objects superimposed on random background images collected from Flickr website. Random translation and rotation transformations are applied to the chair objects and background to generate the second image frame and ground-truth flow fields. *MPI Sintel* [67] consists of 23 video sequences of an action movie, with simulated image degradation effects such as motion blur, defocus blur, and atmospheric effects. This dataset provides a total of 1109 training image pairs with the corresponding optical flow ground-truth labels. Following [64], a subset of the *FlyingThings3D* [52] synthetic dataset was also utilized for training the optical flow networks. Different from the dataset version included for the stereo matching task (section 4.1.1), extremely challenging samples are omitted in this subset.

TABLE 7: Cross-domain generalization evaluation of optical flow networks, using KITTI-2015 (realistic) and Sintel (synthetic) training sets (lower value indicates better performance). All methods are trained on FlyingChairs and FlyingThings3D datasets only. Our method (ITSA) substantially improve the RAFT networks and outperform existing methods in synthetic-to-real generalization. The best results are in **bold** and the second best are underlined.

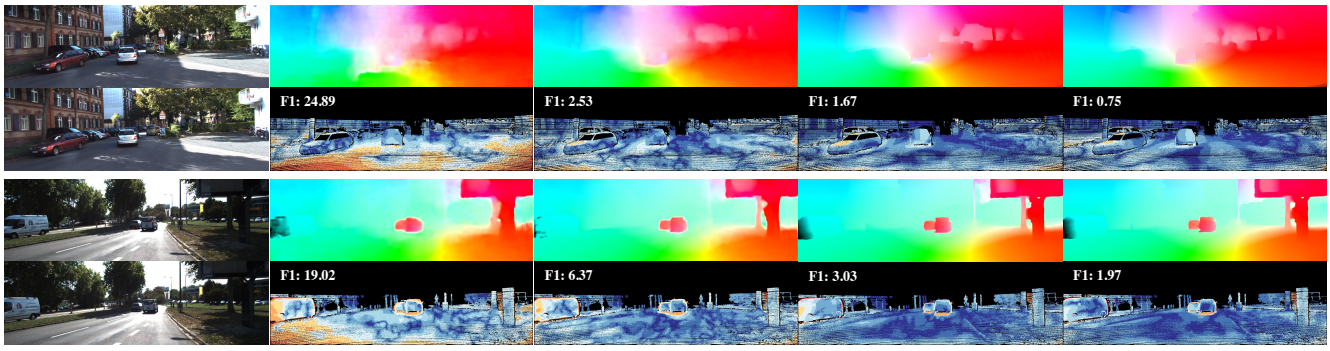
Method	Sintel (train)		KITTI-15 (train)	
	Clean	Final	F1-epe	F1-all
HD3 [49]	3.84	8.77	13.17	24.0
PWCNet [27]	2.55	3.93	10.35	33.7
LiteFlowNet2 [63]	2.24	3.78	8.97	25.9
VCN [8]	2.21	3.68	8.36	25.1
MaskFlowNet [9]	2.25	3.61	-	23.1
FlowNet2 [64]	2.02	3.54	10.08	30.0
DICL-Flow [65]	1.94	3.77	8.70	23.6
RAFT [7]	1.43	2.71	5.04	17.4
SeparableFlow [10]	**1.30**	**2.59**	4.60	15.9
ITSA-PWCNet	2.23	3.76	9.99	30.3
ITSA-RAFT	1.38	2.69	**4.48**	**15.7**

Realistic Dataset: Similar to the stereo matching task, the realistic dataset included in our experiments is the *KITTI 2015* [54] dataset, which consists of 200 image pairs and the sparse ground-truth flow fields. However, to our best knowledge, there are no challenging data samples of anomalous scenes with optical flow ground-truth labels available publicly. Thus, we omit the network robustness analysis for the optical flow estimation task that requires anomalous samples with ground-truth labels.

4.2.2 Implementation Details

For our experiments, we have selected the PwcNet [27] and the recently proposed and top performing RAFT [7] as our baseline networks for optical flow estimation. While there exists many other network architectures for optical flow [9], [64], [68], we have selected these two networks because they are designed specifically for optical flow without including additional auxiliary tasks (e.g. occlusion mask estimation [9], [69]). Also, these networks are well-studied by the optical flow research community, and are commonly used as a baseline in many previous works [70], [71], [72].

As dataset scheduling is utterly important to ensure the convergence of optical flow networks [64], we follow the conventional dataset scheduling procedure to optimize these networks. Both networks were pre-trained from scratch, using the FlyingChairs dataset. Then, the pre-trained networks are optimized on the FlyingThings3D dataset. In our experiments, all training setups and hyper-parameters for the RAFT network are kept the same as the original work. Meanwhile, in PwcNet, we have included batch normalization to the feature extraction sub-network as we find it helps to speed up the convergence of the network to good optimum. PwcNet was pre-trained for 216 epochs (approx. 400k iterations) on the FlyingChairs dataset,



(a) Input Image (b) PwcNet (c) PwcNet-ITSA (d) RAFT (e) RAFT-ITSA

Fig. 7: Qualitative results on KITTI 2015 train data for optical flow task. The colorized optical flow predictions (top row) and the error maps with F1 score superimposed (bottom row) are included for each method (lower value indicates better performance). Despite training with synthetic data only, the proposed ITSA can substantially improve the synthetic-to-realistic domain generalization performance of PwcNet [27] and RAFT [7] networks.

and further trained for 100 epochs (approx. 330k iterations) on the FlyingThings3D dataset. The hyper-parameters for training in PwcNet: such as training loss weights, learning rate, learning rate decay rate and trade-off weight (weight decay regularization); are kept the same as the original implementation. All networks were trained using one NVIDIA RTX 6000 Quadro GPUs.

4.2.3 Shortcut Analysis

We found that optical flow networks learn to exploit the same artefacts as stereo matching networks (see Section 4.1.3). Thus, as shown in Table 6, augmenting the in-domain synthetic images (e.g. FlyingChairs dataset), using asymmetrical color jitter (ACJ) and grayscaling, can significantly deteriorate the performance of optical flow networks. Intuitively, including these chromatic augmentations during training would alleviate shortcut learning in optical flow networks. However, we found that networks trained with these augmentations included (PwcNet* and RAFT* in Table 6) remain fragile to our proposed SCP, despite being robust to the included augmentations. This insight indicates that SCP is able to break the unknown shortcut connections exploited by the network, and worsen its performance.

Furthermore, SCP also managed to deteriorate the performance of the state-of-the-art RAFT [7] network. This observation suggests that certain shortcuts, apart from the identified ones, are also exploited by the the RAFT network. In the next section, we show that our proposed method can removethe shortcut connections in the RAFT network, and further improve its domain generalization performance.

4.2.4 Synthetic-to-Realistic Domain Generalization

In this section, we compare the synthetic-to-realistic domain generalization performance of our proposed ITSA method with the state-of-the-art optical flow networks, using the Sintel [67] (synthetic) and KITTI-2015 [54] (real) datasets. *All networks are trained on the synthetic FlyingChairs and FlyingThings3D datasets only.* As shown in Table 7, our proposed ITSA can consistently improved the synthetic-to-realistic domain generalization performance of the included optical flow networks by a substantial margin. Specifically, our method managed to improve the PwcNet and RAFT by 3.48% and 11.1% of EPE, and 5.4% and 1.7% of F1-all, respectively, when evaluated on the KITTI-2015 dataset. Results of qualitative comparisons are included in Figure 7.

Moreover, our results also show that the proposed method can effectively generalize the performance of optical flow networks between synthetic domains (e.g. FlyingThings3D→Sintel). For example, the performance improvement achieved by ITSA-PWCNet is between $4.33\% - 12.55\%$, and ITSA-RAFT is between $0.74\% - 3.50\%$, when tested on Sintel synthetic dataset.

More impressively, our ITSA-RAFT network performs competitively as compared to the top performing SeparableFlow [10] method in cross-domain generalization, without network modification or increase in model parameters and inference time.

4.3 Semantic Segmentation

4.3.1 Datasets Description

Synthetic Dataset: *GTAV* [73] is a large-scale driving scene semantic segmentation dataset generated using the Grand Theft Auto V game engine. This dataset consists of $24,966$ training samples with dense semantic ground-truth labels, and it has 19 objects categories that are compatible with the realistic Cityscapes dataset.

Realistic Dataset: *Cityscapes* [74] is a large-scale dataset containing high-resolution (e.g., 2048×1024) urban scene images collected from 50 different cities primarily in Germany. *BDD-100K* [75] is another real-world dataset that provides high-resolution (1280×720) and diverse urban driving scene images collected from various locations in the US. This dataset consists of $7,000$ training and $1,000$ validation images with densely annotated ground-truth. *Mapillary* [76] is a diverse street-view dataset consisting of $25,000$ high-resolution (the minimum resolution is 1920×1080) images collected from all around the world. In our experiments, we employed the finely annotated set of *Cityscapes* and the validation set of *BDD-100K* and *Mapillary* for synthetic-to-realistic domain generalization evaluation.

TABLE 8: Analysis of the effect of photometric transformations and SCP on the in-domain and out-of-domain performance of semantic segmentation networks. All networks are only trained on the synthetic GTAV training set, and the mIoU metric is employed for evaluation (higher value indicates better performance). The results show that semantic segmentation networks trained on synthetic data are susceptible to photometric bias. The proposed ITSA can mitigate shortcut learning beyond photometric bias, and significantly improves the synthetic-to-realistic domain generalization performance.

Methods	GTAV					CityScapes
	Baseline	Brightness	Saturation	Contrast	SCP ($\epsilon = 1.0$)	
w/o CJ	66.3	65.2 ($\downarrow$ 1.1)	63.3 ($\downarrow$ 3.0)	65.6 ($\downarrow$ 0.7)	8.9 ($\downarrow$ 57.4)	17.1
w/ CJ	68.5	68.3 ($\downarrow$ 0.2)	68.0 ($\downarrow$ 0.5)	68.3 ($\downarrow$ 0.2)	33.8 ($\downarrow$ 34.7)	30.4
w/ ITSA	67.1	66.4 ($\downarrow$ 0.7)	66.7 ($\downarrow$ 0.4)	66.7 ($\downarrow$ 0.4)	64.5 ($\downarrow$ 2.6)	40.9

4.3.2 Implementation Details

In our semantic segmentation experiments, we adopted the Fully Convolutional Networks (FCN) [77] "backboned" with the ImageNet pre-trained ResNet-50 [78] as our model. The networks were trained using the GTAV [73] synthetic dataset only. The optimized models were directly evaluated on the Cityscapes, BDD-100K and Mapillary realistic datasets, using the mean intersection over union (mIoU) metric. The mIoU is the average of all IoU values over all classes. Training was conducted for 40K iterations, using the SGD optimizer with momentum of 0.9. The learning rate was initialized as 1e-3 for the classifier (final fully connected layer) and 1e-4 for the feature extractor. We also adopted the polynomial learning rate scheduling with the power of 0.9. The batch size was set to 4, and the hyper-parameter ϵ and λ were set to 1.0 and 0.005, respectively. Furthermore, data augmentation techniques such as color jittering, Gaussian blur, random grey-scaling and random cropping were also implemented to prevent the model from overfitting [31].

4.3.3 Shortcut Analysis

In contrast to motion correspondence estimation tasks (e.g. stereo matching, optical flow) that find matching pixels between two viewpoints, semantic segmentation involves pixel-wise classification for a given image. Thus, information such as image context [79] and objectness [80] are assumed to be utilized by the semantic segmentation networks (SSNets) to achieve impressive performance. Instead, it was found that SSNets that are trained on synthetic images learn to extract features that are biased towards photometric characteristics of the input images [31]. As shown in Table 8, applying photometric transformations (e.g. brightness, saturation and contrast augmentation) to the in-domain synthetic test samples can deteriorate the performance of SSNets trained on the GTAV synthetic dataset. Consequently, the photometric-biased networks fail to generalize to unseen realistic domain, and perform poorly when tested on the Cityscapes dataset.

In our experiments, we found that including the aforementioned photometric transformations during training can effectively alleviate the photometric bias, and generalize the SSNets from synthetic to realistic domains (shown in Table 8 mIoU improved by 13.3). Despite the promising results, the SSNets trained with photometric transformation remain fragile to our proposed SCP augmentation (mIoU: 68.5 $\rightarrow$ 33.8). This insight suggests that photometric artefacts are not the only bias exploited by the SSNets. Furthermore, we show that regularizing the training using the proposed ITSA can alleviate the identified photometric bias and arguably other unidentified biases. As a result, the SSNets trained using the proposed ITSA achieve substantial performance improvement in synthetic-to-realistic domain generalization performance (mIoU improved by 23.8).

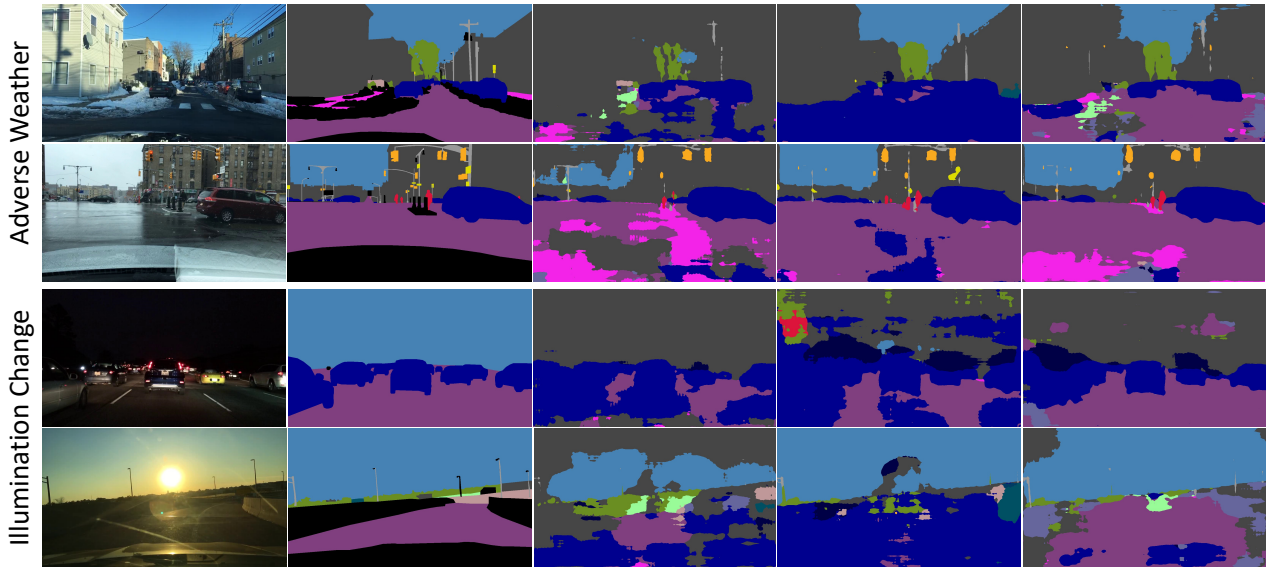
TABLE 9: Synthetic-to-realistic domain generalization evaluation of semantic segmentation, using Cityscapes, BDD-100K and Mapillary validation sets. All methods were trained with ResNet-50 as backbone, using the GTAV synthetic dataset. Mean intersection over union (mIoU) is employed as evaluation metric (higher value indicates better performance). For each method, the base results are included in the top row and the improved results are included in the bottom row.

Methods	Cityscapes		BDD-100K		Mapillary	
	mIoU	mIoU ↑	mIoU	mIoU ↑	mIoU	mIoU ↑
IBN-Net [30]	22.17	7.47	-	-	-	-
	29.64		-		-	
DRPC [33]	32.45	4.97	26.73	5.41	25.66	8.46
	37.42		32.14		34.12	
ISW [31]	28.95	7.63	25.14	**10.06**	28.18	12.15
	36.58		35.20		40.33	
GLTR [32]	31.70	6.90	-	-	-	-
	38.60		-		-	
CSG [11]	25.88	9.39	-	-	-	-
	35.27		-		-	
Ours (ITSA)	30.13	**10.86**	30.08	6.37	27.58	**14.76**
	40.99		**36.45**		**42.34**	

4.3.4 Synthetic-to-Realistic Domain Generalization

In this section, we compare the synthetic-to-realistic domain generalization performance of the proposed ITSA and the state-of-the-art domain generalization methods for semantic segmentation. *All methods included in this section were trained using the GTAV [73] synthetic dataset, and evaluated on three realistic datasets: Cityscapes [74], Mapillary [76] and BDD-100K [75].* As shown in Table 9, our proposed ITSA achieved the best synthetic-to-realistic domain performance improvement on Cityscapes (ResNet50: 10.86) and Mapillary (ResNet50: 14.76) datasets, outperforming the existing state-of-the-art domain generalization methods for semantic segmentation.

In contrast to the previously proposed methods, ITSA does not rely on additional auxiliary datasets (e.g. ImageNet) to perform image style (DRPC [33]) or texture (GLTR [32]) augmentations. Furthermore, ITSA also



(a) Input Image (b) ground-truth (c) Baseline (d) CS-fine-tuned (e) ITSA

Fig. 8: Qualitative results on challenging out-of-domain data provided by BDD-100K [75]. The proposed ITSA can significantly improve the performance of semantic segmentation networks in challenging unseen domains. Despite training on synthetic data only, our method has comparable performance as compared to the network fine-tuned on the realistic Cityscapes [74] dataset (CS-fine-tuned).

TABLE 10: Robust evaluation on anomalous scenarios provided by BDD-100K dataset [75]. The proposed method (ITSA) consistently improves the robustness of synthetically trained semantic segmentation networks and achieve comparable performance as the Cityscapes fine-tuned (CS-FT) models in the real-world anomalous scenarios. The performance was evaluated using the mIoU metric (higher value indicates better performance).

Method	Blur	Night	Weather	Exposure
Baseline	23.57	13.66	27.89	24.99
ITSA	32.29	18.07	33.11	33.90
CS-FT	32.70	15.82	35.94	39.03

does not involve network architecture modification (e.g. additional instance normalization are carefully inserted between residual layers in IBN-Net [30] and ISW [31]). Impressively, to our best knowledge, our method still outperforms all existing published synthetic-to-realistic domain generalization methods for semantic segmentation, on all three realistic datasets.

4.3.5 Network Robustness Analysis

In this section, we analyze the robustness of semantic segmentation networks to the challenging scenarios that are anomalous to the commonly observed data (e.g. Cityscapes [74]). These anomalous scenarios are night-time, adverse weathers, poor lighting and blur corruptions. To this end, we have selected the BDD-100K dataset [75] to evaluate the networks' robustness to the mentioned challenging scenarios. As the raw dataset was not previously divided according to the data attributes, we first split the dataset into the desired subsets, one for each anomalous condition. The selection criteria for each subset are included in the supplementary document (see Section 3). These subsets are then employed for evaluating the robustness of the semantic segmentation networks.

In this comparison, we trained the same semantic segmentation network using three different settings: (1) train only on GTAV synthetic data, (2) train on GTAV synthetic data with the proposed ITSA method, and (3) pre-train on GTAV synthetic data followed by fine-tuning on realistic Cityscapes dataset (without ITSA). To help with the following discussion, we refer to the networks trained on setting (1), (2) and (3) as the baseline, ITSA and fine-tuned networks, respectively. As shown in Table 10, the ITSA network consistently outperforms its baseline counterpart in all the anomalous scenarios (mIoU improvement: $4.41\% - 8.91\%$). This observation again strongly highlights the fact that our proposed ITSA method is not limited to promoting synthetic-to-realistic domain generalization, but also capable of improving the network robustness to challenging realistic anomalous scenarios. As a result, despite training on the synthetic data only, the ITSA network can achieve comparable performance as the fine-tuned network, when tested with the anomalous samples (see Table 10). Results of qualitative comparisons are also included in Figure 8.

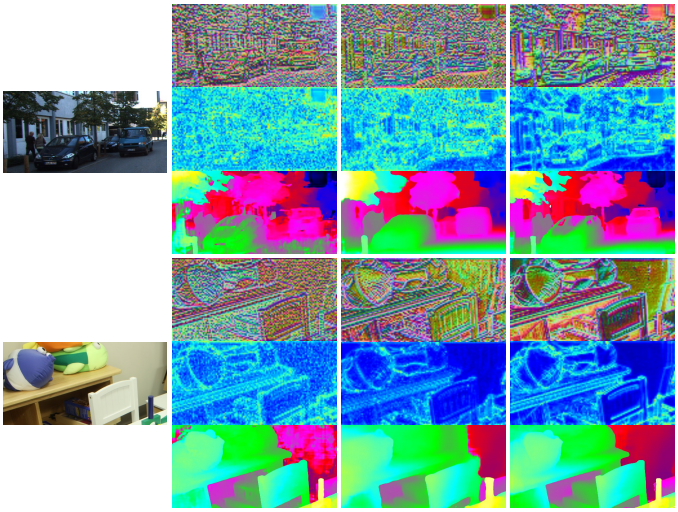
5 Discussions

5.1 Ablation Study

This section presents results of our study on efficacy of each component of the proposed method, for all included dense prediction tasks: stereo matching, optical flow and semantic segmentation. For each task, we first trained the baseline model with the proposed shortcut-perturbation augmentation (SCP) *only*. Next, we trained the baseline model with

TABLE 11: Ablation results of different dense prediction tasks: stereo matching, optical flow estimation and semantic segmentation. The included networks are: PSMNet [24] and GwcNet [48] for stereo matching, PwcNet [27] and RAFT [7] for optical flow and ResNet50 + FCN [77] for semantic segmentation. SCP is the proposed shortcut perturbations and $\mathcal{L}_{\text{FI}}$ is the proposed loss function in Eq. (6). The employed evaluation metrics are: D1 metric for stereo matching and mIoU for semantic segmentation. To evaluate the performance on optical flow, we employed endpoint-error (EPE) for Sintel dataset and F1 metric for KITTI 2015 dataset.

SCP	$\mathcal{L}_{FI}$	Stereo Matching				Optical Flow				Semantic Segmentation		
		KITTI 2012		KITTI 2015		Sintel (Clean)		KITTI 2015		City	BDD	Map
		PSMNet	GwcNet	PSMNet	GwcNet	PwcNet	RAFT	PwcNet	RAFT	ResNet50-FCN		
✗	✗	27.4	11.7	29.3	12.8	2.55	1.43	33.7	17.4	30.1	30.0	27.6
✓	✗	8.1	5.3	8.6	5.9	2.38	1.41	32.1	16.0	38.4	32.9	38.6
✓	✓	5.2	4.9	5.8	5.4	2.23	1.38	30.3	15.7	41.0	36.5	42.3



(a) Input (b) Baseline (c) KITTI-Ft (d) ITSA

Fig. 9: Qualitative illustration of feature maps extracted by stereo matching network (PSMNet) optimized via different training configurations. The normalized feature vectors (unit vectors) and feature vector magnitude are included in the top and middle rows for each example. The bottom row consists of the estimated disparity maps. The baseline model extracted features with substantial random patterns and failed to generalize to realistic domains. The KITTI fine-tuned (KITTI-Ft) model learns to extract features specific to the training domain. In contrast, our ITSA model consistently extracts features with rich structural details across different realistic domains.

both the shortcut-perturbed stereo-images and the proposed loss function $\mathcal{L}_{\text{FI}}$ in Eq. (6). All networks included in this study were trained using synthetic data *only*.

As shown in Tab. 11, for all included dense prediction tasks, the baseline networks perform poorly when tested on out-of-domain datasets. The performance improved when shortcut-perturbations (SCP) were used in the training stage for input image augmentations. Further improvement in the networks can be seen when using the proposed method i.e. SCP with the proposed loss function. For the stereo matching task, we omitted CFNet in the ablation study as it is specifically designed for synthetic to real domain generalization.

5.2 Feature Analysis for Dense Prediction Tasks

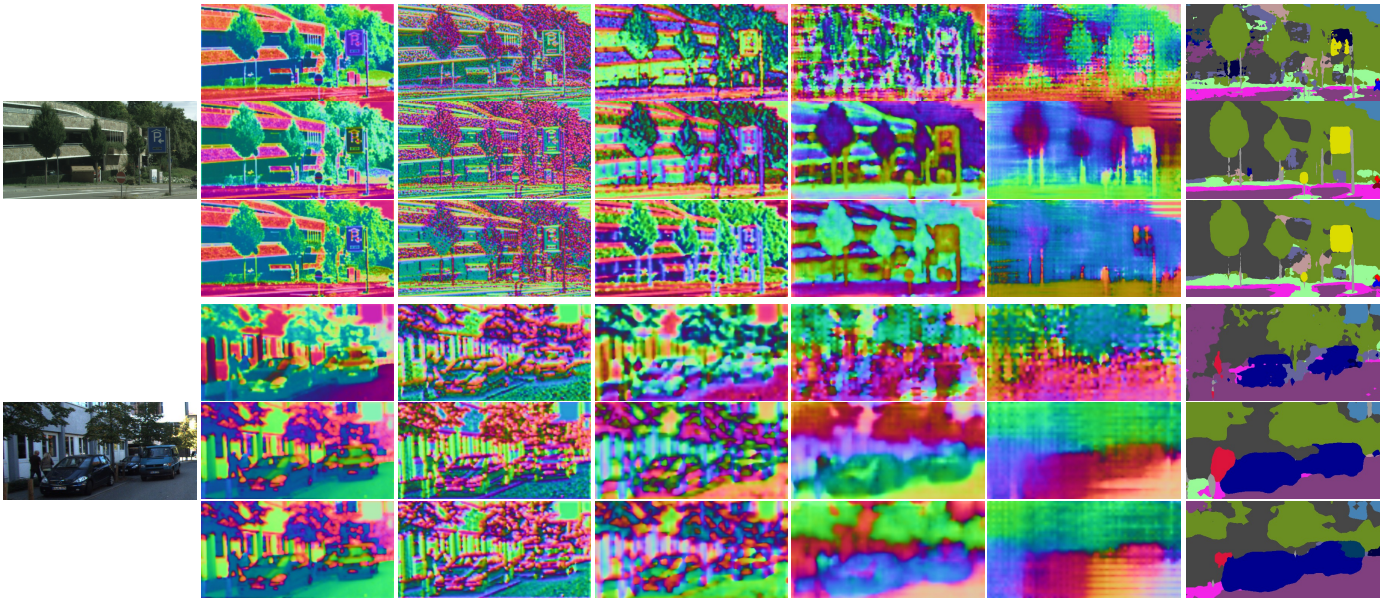
In this section, we analyze the feature maps extracted by dense prediction networks that are optimized via different training configurations: (1) Baseline: trained on synthetic dataset *only*, (2) fine-tuned: pre-trained on synthetic dataset and fine-tuned on realistic datasets (KITTI for stereo matching networks and Cityscapes for semantic segmentation networks) and (3) ITSA: trained on synthetic data only, using our proposed method. The channel dimension of all feature maps is reduced to three, using the Principle Component Analysis (PCA) approach. The pixel-wise normalized feature vector (unit vector) and vector magnitude are then computed using the dimensionally reduced feature maps.

5.2.1 *Stereo Matching Networks*

As depicted in Figure 9, the baseline and fine-tuned models learn to exploit features with substantial random patterns for disparity estimation. Moreover, feature maps with significantly lesser information (activation) are extracted by the fine-tuned model when tested on a different domain (e.g. Middlebury). We argue that these models have learned to exploit domain-specific spurious features and therefore fail to generalize to unseen domains. In contrast, our ITSA model can extract features with rich structural details (e.g. edges and corners), which is crucial for stereo matching. As a result, the ITSA model can generate accurate and detailed disparity measurements in multiple realistic domains, despite the model being trained on synthetic data only.

5.2.2 *Semantic Segmentation Networks*

For the semantic segmentation task, the models are expected to extract the same features for pixels of the same object (e.g. cars, person). Interestingly, we found that all three models (baseline, fine-tuned and ITSA) learn to extract similar feature representations at the early layers (e.g. L1-L3 shown in Figure 10). However, the feature maps extracted by the baseline model differ significantly from the other models at the last two layers (L4 and L5). As illustrated in Figure 10, the baseline model extracts feature maps with random patterns in the penultimate layer (L4), which suggests the exploitation of spurious features. From this insight, we conjecture that learning spurious features is more likely to occur at deeper layers (the penultimate and final layers). Dissimilar to stereo matching networks, the ITSA semantic segmentation network behaves similarly to the fine-tuned



(a) Input (b) L1 (c) L2 (d) L3 (e) L4 (f) L5 (g) Prediction

Fig. 10: Qualitative illustration of feature maps (unit vector) extracted by semantic segmentation network (FCN) optimized via different training configurations. For each example, the feature maps extracted using the baseline model, Cityscapes fine-tuned model and ITSA model are included in the top to bottom rows, respectively. Similar feature representations are learned by all models at early layers (L1-L3). However, feature maps with random patterns are exploited by the baseline model in penultimate (L4) and final (L5) layers for prediction.

one even at later layers, where feature representations with rich semantic details are extracted (objects such as trees, cars, roads and signs are clearly visible in L4 and L5). This observation suggests that both ITSA and fine-tuned models do not include spurious features for prediction, and both models should achieve comparable performance. The latter is supported by results included in Table 10, where the ITSA model can achieve comparable performance as the Cityscapes fine-tuned (CS-FT) model on challenging anomalous scenes.

6 Conclusion

In this work, we have introduced a general yet effective algorithm to mitigate shortcut learning, and improve the performance of synthetic-to-realistic domain generalization in dense prediction networks. Specifically, we proposed the ITSA: a novel information theory-based approach that minimizes 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 in dense prediction networks such as motion correspondence estimation networks (stereo matching and optical flow) and semantic segmentation networks. Extensive experimental results illustrated that the proposed method consistently promotes the learning of robust and shortcut-invariant features, and substantially enhances the performance of the dense prediction networks in cross-domain generalization. Despite training on synthetic data only, our proposed method can remarkably enhance the robustness of the dense prediction networks and performs favourably as compared to their fine-tuned counterparts in realistic anomalous scenarios.

References

[1] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272. 1

[2] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612. 1

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818. 1

[4] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 185–194. 1, 3, 7, 8

[5] Z. Shen, Y. Dai, and Z. 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)*, June 2021, pp. 13 906–13 915. 1, 3, 6, 7, 8, 9

[6] V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 362–14 372. 1

[7] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419. 1, 3, 9, 10, 13

[8] G. Yang and D. Ramanan, "Volumetric correspondence networks for optical flow," *Advances in neural information processing systems*, vol. 32, 2019. 1, 3, 9

[9] S. Zhao, Y. Sheng, Y. Dong, E. I. Chang, Y. Xu *et al.*, "Maskflownet: Asymmetric feature matching with learnable occlusion mask," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6278–6287. 1, 3, 9

[10] F. Zhang, O. J. Woodford, V. A. Prisacariu, and P. H. Torr, "Separable flow: Learning motion cost volumes for optical flow estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 807–10 817. 1, 9, 10

[11] W. Chen, Z. Yu, S. Mello, S. Liu, J. M. Alvarez, Z. Wang, and A. Anandkumar, "Contrastive syn-to-real generalization," *International Conference on Learning Representations*, 2021. 1, 11

[12] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 4

[13] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018. 1, 4

[14] B. Shi, D. Zhang, Q. Dai, Z. Zhu, Y. Mu, and J. Wang, "Informative dropout for robust representation learning: A shape-bias perspective," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8828–8839. 1, 4

[15] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *arXiv preprint arXiv:2004.07780*, 2020. 1, 2, 4, 5, 6

[16] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238. 1, 4

[17] Y. Tang, Z. Jiang, Z. Xie, Y. Cao, Z. Zhang, P. H. Torr, and H. Hu, "Breaking shortcut: Exploring fully convolutional cycle-consistency for video correspondence learning," *arXiv preprint arXiv:2105.05838*, 2021. 1

[18] M. Minderer, O. Bachem, N. Houlsby, and M. Tschannen, "Automatic shortcut removal for self-supervised representation learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6927–6937. 1, 4

[19] N. Dagaev, B. D. Roads, X. Luo, D. N. Barry, K. R. Patil, and B. C. Love, "A too-good-to-be-true prior to reduce shortcut reliance," *arXiv preprint arXiv:2102.06406*, 2021. 1, 4

[20] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 2015, pp. 1–5. 1, 4

[21] A. Pensia, V. Jog, and P.-L. Loh, "Extracting robust and accurate features via a robust information bottleneck," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 131–144, 2020. 2, 4, 5

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014. 2

[23] W. Chuah, R. Tennakoon, R. Hoseinnezhad, A. Bab-Hadiashar, and D. Suter, "Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[24] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418. 3, 6, 7, 8, 9, 13

[25] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75. 3

[26] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," *arXiv preprint arXiv:2004.09548*, 2020. 3

[27] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943. 3, 9, 10, 13

[28] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, and P. Torr, "Domain-invariant stereo matching networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 420–439. 3, 7, 8

[29] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020. [Online]. Available: https://doi.org/10.1038/s42256-020-00257-z 3

[30] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 464–479. 3, 11, 12

[31] S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. 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*, 2021, pp. 11 580–11 590. 3, 4, 11, 12

[32] D. Peng, Y. Lei, L. Liu, P. Zhang, and J. Liu, "Global and local texture randomization for synthetic-to-real semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 6594–6608, 2021. 3, 4, 11

[33] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. 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*, 2019, pp. 2100–2110. 3, 4, 11

[34] X. Pan, X. Zhan, J. Shi, X. Tang, and P. Luo, "Switchable whitening for deep representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1863–1871. 3

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456. 3

[36] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016. 3

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 4

[38] D. Duckworth, E. S. Dyer, J. Sohl-dickstein, N. Wadia, and S. S. Schoenholz, Eds., *Whitening and second order optimization both destroy information about the dataset, and can make generalization impossible*, 2021. [Online]. Available: http://proceedings.mlr.press/v139/wadia21a/wadia21a.pdf 4

[39] H. Wang, Z. He, Z. L. Lipton, and E. P. Xing, "Learning robust representations by projecting superficial statistics out," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rJEjjoR9K7 4

[40] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 5389–5400. 4

[41] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *Proceedings of the International Conference on Learning Representations*, 2019. 4

[42] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep variational information bottleneck," in *ICLR*, 2017. [Online]. Available: https://arxiv.org/abs/1612.00410 4

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 4

[44] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189. 4

[45] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. [Online]. Available: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf 4

[46] J. Denker, W. Gardner, H. Graf, D. Henderson, R. Howard, W. Hubbard, L. D. Jackel, H. Baird, and I. Guyon, "Neural network recognizer for hand-written zip code digits," *Advances in neural information processing systems*, vol. 1, 1988. 4

[47] Y. Shi, J. Seely, P. H. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," *arXiv preprint arXiv:2104.09937*, 2021. 5

[48] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282. 6, 7, 8, 9, 13

[49] Z. Yin, T. Darrell, and F. Yu, "Hierarchical discrete distribution decomposition for match density estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6044–6053. 7, 8, 9

[50] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. 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*, 2020, pp. 2495–2504. 7, 8

[51] C. Cai, M. Poggi, S. Mattoccia, and P. Mordohai, "Matching-space stereo networks for cross-domain generalization," in *2020 International Conference on 3D Vision (3DV)*, 2020, pp. 364–373. 7

[52] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. 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*, 2016, pp. 4040–4048. 7, 9

[53] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 7, 8

[54] M. Menze, C. Heipke, and A. Geiger, "Object scene flow," *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018. 7, 8, 9, 10

[55] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*. Springer, 2014, pp. 31–42. 7, 8

[56] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7

[57] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. 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*, 2019, pp. 899–908. 7, 8

[58] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017. [Online]. Available: http://dx.doi.org/10.1177/0278364916679498 7, 8

[59] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453. 7

[60] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving," *arXiv preprint arXiv:1906.06310*, 2019. 7

[61] Y. Zhang, Y. Chen, X. Bai, J. Zhou, K. Yu, Z. Li, and K. Yang, "Adaptive unimodal cost volume filtering for deep stereo matching," *arXiv preprint arXiv:1909.03751*, 2019. 7

[62] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5515–5524. 7

[63] T.-W. Hui, X. Tang, and C. C. Loy, "A lightweight optical flow cnn—revisiting data fidelity and regularization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2555–2569, 2020. 9

[64] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470. 9

[65] J. Wang, Y. Zhong, Y. Dai, K. Zhang, P. Ji, and H. Li, "Displacement-invariant matching cost learning for accurate optical flow estimation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 220–15 231, 2020. 9

[66] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/DFIB15 9

[67] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conf. on Computer Vision (ECCV)*, ser. Part IV, LNCS 7577, A. Fitzgibbon et al. (Eds.), Ed. Springer-Verlag, Oct. 2012, pp. 611–625. 9, 10

[68] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766. 9

[69] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5754–5763. 9

[70] F. Aleotti, M. Poggi, and S. Mattoccia, "Learning optical flow from still images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 201–15 211. 9

[71] S. Jiang, Y. Lu, H. Li, and R. Hartley, "Learning optical flow from a few matches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 592–16 600. 9

[72] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9772–9781. 9

[73] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European conference on computer vision*. Springer, 2016, pp. 102–118. 10, 11

[74] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 10, 11, 12

[75] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645. 10, 11, 12

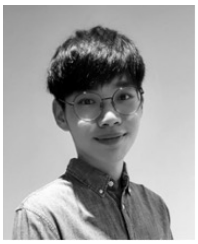
[76] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999. 10, 11

[77] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. 11, 13

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 11

[79] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic correlation promoted shape-variant context for segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8885–8894. 11

[80] R. Liu, Z. Wu, S. Yu, and S. Lin, "The emergence of objectness: Learning zero-shot segmentation from videos," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 11

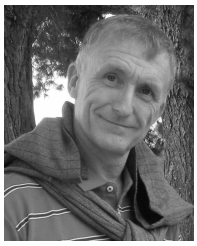


WeiQin Chuah received the B.Eng.(Hons) degree in Adv Manufacturing and Mechatronics in 2018, and he is currently pursuing the Ph.D. degree in engineering at the Royal Melbourne Institute of Technology (RMIT University). His research interests include computer vision, stereo matching and depth estimation systems, machine learning and deep learning, autonomous driving, and related applications.

Ruwan Tennakoon obtained his PhD degree in computer vision from Swinburne University of Technology, Australia (2015) and his BSc degrees (with fists class honours) in Electrical & Electronics Engineering from University of Peradeniya, Sri Lanka (2007). He is currently a Senior Lecturer in artificial intelligence at RMIT university, Australia. His main research interests include computer vision, machine learning and medical image analysis.

Reza Hoseinnezhad received his PhD in 2002 then held various positions at Swinburne University of Technology, The University of Melbourne, and RMIT University, where he is Associate Dean (Mechanical & Automotive Engineering). His main research interests include statistical information fusion, random finite sets, multi-object tracking, deep learning, and robust multi-structure data fitting in computer vision.

Alireza Bab-Hadiashar received the B.Sc. and M.Eng. degrees in mechanical engineering and the Ph.D. degree in robotics from Monash University. He has held various positions in Monash University, the Swinburne University of Technology, and RMIT University, where he is currently a Professor of mechatronics and leads the Intelligent Automation Research Group. His main research interests include intelligent automation in general, robust data fitting in machine vision, deep learning for detection and identification, and robust data segmentation.



David Suter received the BSc degree in applied mathematics and physics from the Flinders University of South Australia, in 1977, the graduate diploma degree in computing from the Royal Melbourne Institute of Technology, in 1984, and the PhD degree in computer science from La Trobe University, in 1991. He was a lecturer with La Trobe from 1988 to 1991; and a senior lecturer in 1992, associate professor in 2001, and profes- sor from 2006 to 2008 with Monash University, Melbourne, Australia. From 2008 to 2017, he was a professor with the University of Adelaide. Since 2018, he has been a professor with Edith Cowan University, Australia. He served on the ARC College of Experts from 2008 to 2010. He has previously served on the editorial boards of the Machine Vision and Applications, the International Journal of Computer Vision, and the Journal of Mathematical Imaging and Vision. He was general co-chair of ACCV 2002 and ICIP 2013.