

Towards Adversarially Robust and Domain Generalizable Stereo Matching by Rethinking DNN Feature Backbones

Kelvin Cheng, Christopher Healey, and Tianfu Wu

North Carolina State University



Figure 1. Examples of attacking stereo matching in the KITTI2015 [29] dataset. GANet-Deep [48] results on the top row, our results on the bottom row. The attack is based on the proposed stereo-constrained projected gradient descent (PGD) attack within a patch, which by design preserves the photometric consistency of non-occluded regions. One of the state-of-the-art methods, GANet-Deep shows a significant drop in performance (the last column), while the proposed method shows much stronger resistance to the attack.

Abstract

Stereo matching has recently witnessed remarkable progress using Deep Neural Networks (DNNs). But, how robust are they? Although it has been well-known that DNNs often suffer from adversarial vulnerability with a catastrophic drop in performance, the situation is even worse in stereo matching. This paper first shows that a type of weak white-box attacks can overwhelm state-of-the-art methods. The attack is learned by a proposed stereo-constrained projected gradient descent (PGD) method in stereo matching. This observation raises serious concerns for the deployment of DNN-based stereo matching. Parallel to the adversarial vulnerability, DNN-based stereo matching is typically trained under the so-called simulation to reality pipeline, and thus domain generalizability is an important problem. This paper proposes to rethink the learnable DNN-based feature backbone towards adversarially-robust and domain generalizable stereo matching by completely removing it for matching. In experiments, the proposed method is tested in the SceneFlow dataset and the KITTI2015 benchmark, with promising results. We compute the matching cost volume using the classic multi-scale census transform (i.e., local bi-

nary pattern) of the raw input stereo images, followed by a stacked Hourglass head sub-network solving the matching problem. It significantly improves the adversarial robustness, while retaining accuracy performance comparable to state-of-the-art methods. It also shows better generalizability from simulation (SceneFlow) to real (KITTI) datasets when no fine-tuning is used.

1. Introduction

Stereo matching remains a long-standing problem in computer vision that has been studied for several decades. It has great potential in a wide range of applications such as autonomous driving and robot autonomy.

As in many other computer vision problems, Deep neural networks (DNNs) have made tremendous progress in stereo matching. The growing ubiquity of DNNs in computer vision dramatically increases their capabilities, but also increases the potential for new vulnerabilities to attacks [43, 24, 37, 14]. This situation has become critical as many powerful approaches have been developed where imperceptible perturbations to DNN inputs could

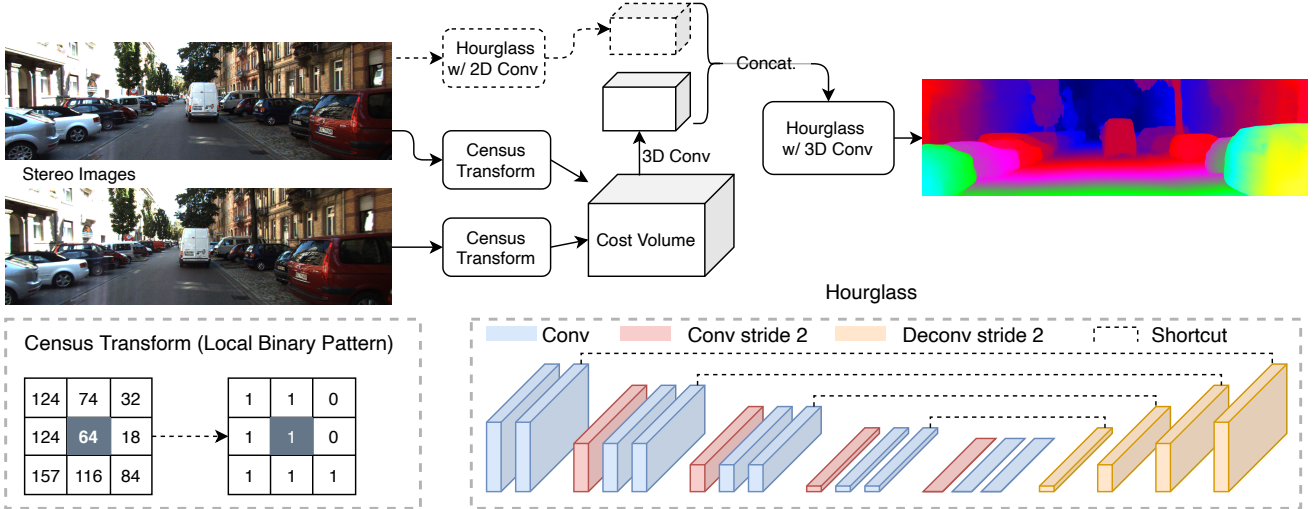


Figure 2. Illustration of the proposed minimally-simple workflow for stereo matching. The key difference between the proposed method and prior art lies in the way of computing the cost volume. The proposed method harnesses the classic multi-scale census transform (left-bottom) of raw intensity of an input stereo image pair, while prior art utilize features computed by a ConvNet feature backbone on an input stereo image pair. The proposed method also exploits ConvNet features computed only using the left reference image, as contextual information to the cost volume. Note that we also test the workflow without using the ConvNet feature context branch, that is to completely remove the ConvNet feature backbone. For the cost aggregation component, the proposed method utilizes a stacked Hourglass sub-network equipped with 3D convolution. Please see the text for detail.

deceive a well-trained DNN, significantly altering its prediction. Such results have initiated a rapidly proliferating field of research characterized by ever more complex attacks [18, 27, 31, 26, 45, 12, 49] that prove increasingly strong against defensive countermeasures [22, 44, 32]. For the trade-off between accuracy and adversarial vulnerability, DNNs seem to have become the Gordian Knot in state-of-the-art computer vision systems.

Since stereo matching methods are widely used in autonomous driving, adversarial vulnerabilities in these models can lead to catastrophic consequences. [42] test attacks on stereo images independently, resulting in perturbations that will alter the colors of the corresponding projections of the same physical point and thus may not be realizable and threatening in practice. To find out whether stereo matching methods are vulnerable in a physically realizable setting, we propose the stereo-constrained projected gradient descent (PGD) attack and show that state-of-the-art methods are indeed vulnerable even when the color differences between corresponding pixels are preserved.

To defend against adversarial attacks, most methods rely on adversarial training [27], which may suffer from decreasing performance, long training time, and over-fitting to specific attacks and datasets. In contrast, we propose to utilize domain-specific knowledge to facilitate the built-in robustness of the neural networks. Because of the strong photometric consistency between stereo images, stereo matching provides an ideal case to defend against adversarial attacks

through the design of the neural network. For non-occluded regions in stereo images, the corresponding pixels of the same physical point have similar colors. We suspect that by using DNN features for matching, attacks will increase the matching costs for features that belong to the same physical point. Therefore, we propose to remove DNN features for matching and use hand-crafted features that will preserve the low color differences for true pairs. To make the cost as hard to alter as possible, we use local binary patterns that compare each pixel intensity to their neighbors (i.e., Census Transform [21, 3]) as the feature descriptor. On the other hand, since DNN features are useful for high-level semantic information that will facilitate the estimation of occluded and textureless regions, we use a feature backbone for the reference image only to contextualize the input. The non-parametric cost volume and the contextual information will be fed through a head sub-network playing the role of a learnable optimizer that seeks the best matching result (Fig. 2). In essence, we cast stereo matching as a cost aggregation/optimization problem over a non-parametric cost volume. In experiments, we show that this more transparent approach improves adversarial robustness significantly while maintaining high accuracy.

Parallel to the adversarial vulnerability, *cross-domain generalizability* also is an important problem in stereo matching: DNN-based stereo matching is typically pre-trained under the so-called simulation to real (Sim2Real) pipeline due to the high cost of collecting ground-truth

matching results in practice and the data-hungry aspect of DNNs. It has been shown that DNNs may learn shortcut solutions that are strongly biased by the training dataset [17]. Removing the DNN feature backbone for matching will induce the DNN to be a more general cost volume optimizer, thus alleviating the opportunity of shortcut learning in the feature space and resulting in better performance in cross-domain deployment, especially when no fine-tuning is used. These are verified in our experiments from the SceneFlow dataset [28] to the KITTI benchmark [29] when no fine-tuning is used.

2. Related Work and Our Contributions

Deep Stereo Matching. After [47] developed the first deep learning approach for stereo matching, [28] built the first end-to-end trainable DNN-based method DispNet and constructed SceneFlow, a large-scale synthetic dataset containing around 40,000 images. In GCNet, [23] further extend the end-to-end approach by concatenating features in the cost volume stage, using 3D convolutional layers for cost aggregation, and introducing the soft arg min operator to compute the expected disparity. Most subsequent approaches followed these design choices and use SceneFlow for pretraining [29].

[8], [13], and [19] further improve the cost aggregation stage. Chang *et al.* proposed to use a Spatial Pyramid Pooling (SPP) Module for feature extraction and to use the stacked Hourglass structures [30] for the cost aggregation. [46] speed up stereo matching by using highly optimized hand-crafted features (e.g. Census Transform and Sum of Absolute Differences). Hourglass and Census Transform are also used in our approach.

[48, 9] proposed to propagate cost spatially to reduce the number of 3D convolutional layers. [10] applies Neural Architecture Search (NAS) techniques to automatically find optimal architectures for each stage and further improve the performance. These approaches are the current state-of-the-art in the KITTI 2015 benchmark [29].

Adversarial Attacks and Defense. Assuming full access to DNNs pretrained with clean images, white-box targeted attacks are powerful ways of investigating the brittleness of DNNs. Many white-box attack methods focus on norm-ball constrained objective functions [38, 25, 6, 11, 34]. By introducing momentum in the MIFGSM method [11] and the ℓ_p gradient projection in the PGD method [27], they usually achieve better performance in generating adversarial examples.

In autonomous driving, physically realizable attacks are investigated in many tasks [15, 33, 5, 41, 40], except for stereo matching. Although [42] show that DNN-based stereo matching methods are vulnerable against unconstrained adversarial attacks on both images separately, without enforcing photometric consistency, these attacks will

violate the underlying physical properties of binocular vision and thus are not realizable in practice. As a result, unconstrained attacks cannot compute adversarial patches to fool stereo systems. Therefore we intentionally design the stereo-constrained PGD attack to further investigate the adversarial robustness in more realistic settings with the presence of photometric consistency. [33] studied adversarial attacks in optical flow, which is inherently easier to attack than stereo matching due to the problem difficulty. By leveraging the insights from Ranjan’s work, our work may shed light on studying more robust optical flow networks.

Towards defense, adversarial training is the most widely used method to improve adversarial robustness [27, 2]. However, it also suffers from the disadvantages of dropping accuracy, long training time, and over-fitting to specific attacks and datasets. While adversarial training is universal to all kinds of DNNs, our method increases the built-in robustness by utilizing the photometric consistency of stereo matching, thus avoiding the mentioned disadvantages. It can also be combined with adversarial training to further improve robustness.

Our Contributions. This paper makes three main contributions to the field of stereo matching:

- It proposes a novel design for stereo matching, which shows significantly better adversarial robustness and cross-domain (Sim2Real) generalizability when no fine tuning is used.
- It presents the stereo-constrained projected gradient descent (PGD) attack method, which by design preserves photometric consistency to show the more serious vulnerabilities of state-of-the-art DNN-based stereo matching methods.
- It showcases a deep integrative learning paradigm by rethinking the end-to-end DNN feature backbones in stereo matching, which sheds light on potentially mitigating shortcut learning in DNNs via leveraging classic hand-crafted features if a problem-specific sweet spot can be identified (such as the cost volume in stereo matching).

3. Approach

In this section, we present the proposed method and the stereo-constrained PGD attack method to evaluate the brittleness of DNN-based stereo matching methods.

3.1. The Proposed Method

As illustrated in Fig. 2, the proposed workflow consists of three main components as follows.

i) Computing the Cost Volume Using Multi-Scale Census Transform. Most current stereo matching methods use DNN-based features to form the 4D cost volume.

In terms of matching, DNNs can increase the uniqueness of the feature for each pixel, but they also suffer from the inherent adversarial vulnerability. In contrast, traditional methods often use simple window-based similarity functions to initialize the costs, then rely on the optimization or cost aggregation stage to integrate all local cost information [39]. Following the same philosophy, we propose to use hand-crafted feature descriptors and similarity functions that are less sensitive to adversarial perturbations to initialize the costs, then rely on DNNs to integrate the local cost information. Specifically, we want the feature descriptor to change as little as possible when local intensities are perturbed. This specific requirement lead us to the Census Transform, a traditional feature descriptor that is developed to eliminate the issue of radiometric differences caused by different exposure timing or non-Lambertian surfaces. Previous studies find that Census Transform is the most robust and well-rounded cost function with global or semi-global methods [21, 3].

We use grey-scale raw intensity values in computing the census transform. Given a local window patch W centered at a pixel $u \in \Lambda$, the census transform computes the local binary pattern (the left-bottom in Fig. 2) by comparing each neighboring pixel $v \in W$ with u such that it equals 1 if $I(v) \geq I(u)$ and 0 otherwise. Hamming Distance (i.e. the number of different values in two bit strings) is used to compute the cost between two patches.

Unlike in traditional semi-global or global methods in which the cost of each pair can only be a scalar, we take advantage of the flexibility of DNNs and design the multi-scale census transform to incorporate the context at different scales. Specifically, We use local windows with sizes from k_1 to k_2 (e.g. $k_1 = 3, k_2 = 11$ in our experiments) so there are $K = k_2 - k_1 + 1$ (e.g., 9) costs associated with each matching candidate pairs. To normalize the cost at each scale, we divide the Hamming Distance by the number of pixels of each local window. For an input stereo image pair, I^L and I^R with the spatial dimensions $H \times W$, assume the maximum disparity level denoted by ℓ , the initial cost volume is a 4-D tensor of the size $H \times W \times \ell \times K$. To reduce the computational cost, we use 3D convolutions to down-scale the cost volume to be $1/3H \times 1/3W \times 1/3\ell \times C$, where $C = 32$ is the number of channels, as typically done by prior art.

ii) Contextualizing the Cost Volume and Aggregating the Cost. Although being robust to adversarial attacks, the census transform based cost volume alone is not sufficiently powerful to handle occlusion and more challenging semantic information, such as transparent objects and specular reflections. We introduce a 2-stack Hourglass module with 2D convolutions to extract context information from the left reference image, resulting in a $1/3H \times 1/3W \times C$ feature map which is unsqueezed along the second dimension (i.e.,

copying the feature map $1/3\ell$ times) to form a same size tensor as the down-scaled cost volume. The two are then concatenated along the second last dimension.

The contextualized cost volume will be fed into a 3-stack Hourglass module with 3D convolutions for the cost aggregation stage.

iii) Disparity Map Prediction and the Loss Function.

To predict the final disparity map $D(u), \forall u \in \Lambda$, the output of each stack in the Hourglass module of the cost aggregation is first up-sampled to the original size $H \times W \times \ell$, denoted as $D_s(x, y, d)$ where s is the stack index in the stacked Hourglass module. Then, similar to the method used in [23], the predicted disparity map $D_s(x, y)$ is computed by,

$$D_s(x, y) = \sum_{d=1}^{\ell} d \times \text{Softmax}(D_s(x, y, d)), \quad (1)$$

where *Softmax* is applied along the last dimension in $D_s(x, y, d)$.

In training, we use the smooth L_1 loss due to its robustness at disparity discontinuities and low sensitivity to outliers [8, 48]. Given the ground-truth disparity map $D^*(u)$, the loss is defined by,

$$\text{Loss}(\Theta; D^*) = \sum_{s=1}^S \beta_s \cdot \frac{1}{|\Lambda|} \sum_{u \in \Lambda} \text{Smooth}_{L_1}(D_s(u) - D^*(u)), \quad (2)$$

where Θ collects all parameters in our model, β_s represents the weight for the output from a stack s (e.g., 0.5, 0.7, and 1 are used for the 3-stack Hourglass module in our experiments), $u = (x, y) \in \Lambda$, and the smooth L_1 function is defined by,

$$\text{Smooth}_{L_1}(z) = \begin{cases} \frac{z^2}{2}, & \text{if } z < 1 \\ |z| - 0.5, & \text{otherwise.} \end{cases} \quad (3)$$

4. Stereo Constrained PGD Attacks

To study the brittleness of DNN based stereo matching models, we intentionally develop a realizable attacking method based on the PGD method [27], which retains the underlying photometric consistency in stereo matching by changing the intensities of the same physical point in both images. More specifically, in learning attacks, the same amounts of perturbations are added to each pair of correspondence pixels in the left and right images simultaneously while occluded areas will not be modified. Since the left image is the reference image for computing the disparity loss, we disallow to attack and evaluate occluded regions of the reference image, which prevents the perturbation to attack the regions where the estimation does not rely on matching.

Given a perturbation map $P(x, y), \forall (x, y) \in \Lambda$, the distorted intensities for each pixel location (x, y) are computed

as:

$$\begin{aligned} I_{adv}^L(x, y) &= I^L(x, y) + P(x - D(x, y), y), \\ I_{adv}^R(x, y) &= I^R(x, y) + P(x, y), \end{aligned} \quad (4)$$

where $D(x, y)$ is the ground-truth disparity map.

Consider two corresponding patches on the left and the right images containing the same physical points, the absolute sum of difference between these two patches will remain the same after the attack.

We use the L_∞ norm to measure similarities between images. Two images will appear visually identical under a certain threshold. To learn a L_∞ bounded adversarial perturbation P^{adv} , the iterative PGD method is used,

$$P_{t+1}^{adv} = clip_P^\epsilon \{ P_t^{adv} + \alpha \cdot \text{sign}(\nabla_P L(P_t^{adv})) \}, \quad (5)$$

where $t = 1, 2, \dots, T$ and P_0^{adv} starts with all zeros. L denotes the mean absolute error between the predicted disparity map for the perturbed images and the ground-truth disparity map. And, $clip_P^\epsilon$ clips the perturbation to be within the ϵ -ball of the corresponding zero-plane and the maximum color range. Throughout our experiments, we set $\epsilon = 0.06$ or 0.03 , $\alpha = 0.01$, and $T = 20$.

Attack Census Transform. Since Census Transform contains the non-differentiable comparison operator, the gradients from the constructed cost volume cannot be back-propagated directly to the input images thus leading to an illusion of safety, *i.e.* the obfuscated gradient problem [1]. For fair comparisons with differentiable methods, we combine subtraction and the sigmoid function as a differentiable approximation of the comparison operator.

$$a > b \approx \text{sigmoid}(a - b) \cdot C \quad (6)$$

We use a large constant (*i.e.* $C = 10^5$) such that the output of the sigmoid function is close to either zero or one. Without using this differentiable approximation, our method without the contextual feature backbone will be **unattackable** since the gradient flows are completely blocked.

Robustness of Census Transform. From the perspective of attacks, the binary patterns generated by Census Transform is more difficult to alter due to the comparison operator. Given a threshold of maximum pixel difference in perturbation, **neighbors will not be altered if their difference with the center is larger than twice the threshold.** If the attack does not violate photometric consistency, it will be even harder to alter the cost between binary patches of corresponding pairs. Specifically, if a neighboring pixel appears in both the left and the right binary patches, its relative magnitude relationship with the center pixel will be the same for both patches, no matter how its intensities change. It is our interest to test if this highly non-linear operator can defend the DNNs against attacks.

5. Experiments

In this section, we first present details of training and testing the proposed method. Then, we present the results on the Sim2Real cross-domain generalizability, followed by showing results on the adversarial robustness. **Our PyTorch source code is provided in the supplementary material.**

5.1. Settings and Implementation Details

Data. We evaluate our method on the SceneFlow [28] and KITTI2015 [29] datasets. The SceneFlow dataset is a large-scale synthetic dataset that contains 35,454 training images and 4,370 test images at the resolution of 540×960 . Since it provides dense ground-truth disparities, it is widely used for pretraining DNN-based stereo matching methods. The KITTI2015 dataset is a real-world dataset of driving scenes, which contains 200 training images and 200 test images at the resolution of 375×1242 . Since the depth of each scene is obtained through LiDAR, the ground truth is not dense. In addition, we also test pretrained models on the KITTI2012 [16] and the Middlebury [35] dataset at quarter resolution.

Implementation Details. Our method is implemented in PyTorch and trained end-to-end using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All images are preprocessed with color normalization. During training, we use a batch size of 8 on four GPUs (Tesla V100) using 240×576 random crops from the input images. The maximum disparity level is set to 192 and any values larger than this threshold will be ignored during training. For SceneFlow, we train our model from random initialization for 20 epochs with a constant learning rate of 0.001. For KITTI2015, we split the 200 training images into a training set of 140 images and a validation set of 60 images. We fine-tune our model pretrained on SceneFlow for another 600 epochs and use the validation set to select the best model. If no feature backbone is used to extract context information from the left image (Fig. 2), our model is denoted as **“ours w/o backbone”** or **“ours w/o b.”** in tables and figures.

To compare with adversarial training, we fine-tuned each method on *KITTI2015* training images perturbed by 3-step unconstrained PGD attacks for 20 epochs, denoted as **adv.** in tables.

Evaluation Metrics. We adopt the provided protocols in the two datasets. There are three metrics: **EPE [px]** which measures the end-point error in pixels, **Bad 1.0 [%]** and **Bad 3.0 [%]** which represents the error rate of errors larger than 1 pixel and 3 pixels respectively.

Baseline Methods. We compare with state-of-the-art deep stereo matching methods: the PSMNet [8], the GANet [48], and the LEAStereo [10]. We use their publicly released codes and trained model checkpoints in comparisons.

Models (trained on SceneFlow)	KITTI 2015			KITTI 2012			Middlebury		
	EPE	Bad 1.0	Bad 3.0	EPE	Bad 1.0	Bad 3.0	EPE	Bad 1.0	Bad 3.0
PSMNet	6.89	72.93	31.55	5.90	71.59	28.42	4.33	73.01	19.01
GANet	1.66	42.12	10.48	1.48	31.61	9.51	2.26	27.45	11.40
LEAStereo	2.00	51.29	13.90	1.91	44.26	14.28	3.47	32.67	14.81
Ours w/o backbone	1.25	25.95	6.12	1.23	19.66	6.80	1.71	18.72	9.16
Ours	1.26	27.92	6.31	1.28	20.62	7.16	1.96	20.09	10.05

Table 1. Comparisons for the Sim2Real cross-domain generalizability from the SceneFlow trained models to the KITTI 2015, KITTI 2012 and Middlebury datasets in testing without any fine-tuning.

	EPE				Bad 1.0				Bad 3.0			
	CL	CT	CT	UCT	CL	CT	CT	UCT	CL	CT	CT	UCT
		0.03	0.06	0.03		0.03	0.06	0.03		0.03	0.06	0.03
PSMNet	0.28	29.05	84.04	91.08	2.00	84.75	90.41	92.75	0.16	54.80	83.68	89.91
GANet	0.25	3.93	9.75	23.75	1.42	70.64	84.68	89.48	0.10	29.94	68.70	79.11
LEAStereo	0.37	4.02	11.38	14.71	4.54	71.20	83.24	82.42	0.42	29.09	63.61	64.31
Ours w/o b.	0.38	1.13	1.43	2.36	4.14	24.64	30.69	41.34	0.32	2.46	8.05	16.30
Ours	0.36	0.88	1.16	1.81	3.61	21.20	29.19	36.42	0.27	3.75	6.17	11.29
PSMNet + adv.	0.46	0.70	1.02	1.06	8.04	17.78	33.54	36.50	0.66	1.40	3.08	4.14
GANet + adv.	0.42	0.65	0.98	1.05	6.47	14.99	28.56	31.22	0.63	1.40	3.76	4.36
LEAStereo + adv.	0.51	0.81	1.23	1.30	9.89	21.73	38.72	42.06	0.99	2.34	5.59	6.07
Ours w/o b. + adv.	0.42	0.78	0.90	1.26	5.95	16.83	21.42	32.27	0.73	2.88	3.83	7.51
Ours + adv.	0.41	0.61	0.69	0.88	5.77	13.46	16.29	22.93	0.52	1.39	2.00	3.99

Table 2. PGD Attack Results in the KITTI2015 training dataset [29]. For each metric, the four columns show that metric on *CL*ean image, stereo-constrained attacked image (*CT*, $\epsilon = 0.03$), stereo-constrained attacked image (*CT*, $\epsilon = 0.06$), and unconstrained attacked image (*UCT*, $\epsilon = 0.03$). Note that on clean images, the results are performance on all the training and validation data, which are affected by different training-validation splits.

5.2. Sim2Real Cross-Domain Generalizability

To verify the conjecture that cross-domain generalizability in stereo matching can be induced by removing the dependency between the cost volume computation and the dataset-dependent feature backbone, we evaluate all models pretrained on SceneFlow directly on the KITTI training datasets and the Middlebury training dataset [35]. As shown in **Table 1**, **our method outperforms prior art by a large margin**. This result shows that our proposed design of combining a non-parametric cost volume formed by the multi-scale census transform and a generalized cost aggregation/optimization DNN is indeed more cross-domain consistent. It also shows that the head sub-network DNN indeed learns to play the role of a domain-independent optimizer over a given cost volume.

5.3. Results in KITTI

Experiment I): Adversarial Robustness Comparisons. To evaluate the adversarial robustness in KITTI2015, we directly test the trained models on the entire training dataset (200 images). Due to the GPU memory limitation, we only use the 240×384 center part of each image. Because of cropping, we also ignore those pixels where their

correspondences are outside of the cropped images. We test both the stereo-constrained attack with $\epsilon = 0.03, 0.06$ and the unconstrained attack with $\epsilon = 0.03$. **Table 2** shows the comparison results.

From the results, we show that state-of-the-art stereo matching methods are indeed vulnerable against adversarial attacks, even when photometric-consistency is preserved. Such vulnerability may raise serious concerns for the deployment of DNNs in safety critical applications. In contrast, **our method shows significantly better robustness on both stereo-constrained and unconstrained attacks**.

Note that our approach without feature backbone is unattackable if our proposed differentiable approximation of the comparison operator is not applied. Although this approach is highly non-linear, adversarial attacks can still find ways to perturb the input images, which further demonstrate the vulnerability of the DNNs. Interestingly, our method with the context feature backbone is more robust than its counterpart, showing that **the majority of the vulnerability actually comes from the matching part rather than the contextual information**.

Experiment II): Adversarial Patch Attack. To test if the adversarial vulnerability can be intentionally exploited

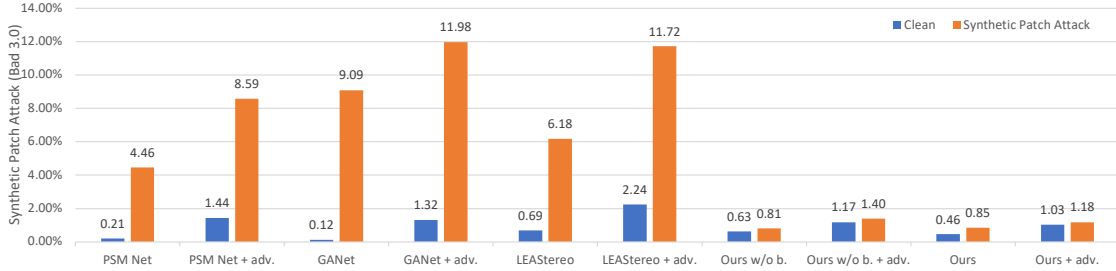


Figure 3. Adversarial Patch Attack Results in the KITTI2015 training dataset with photometric consistency retained in attack.

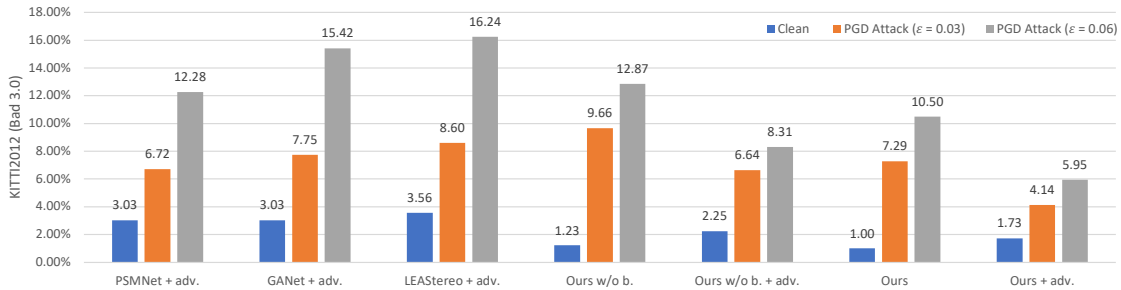


Figure 4. Transferrability of Adversarial Robustness: stereo-constrained 20-step PGD Attack Results in the KITTI2012 training dataset using adversarially trained neural networks on KITTI2015.

in a more realistic setting, such as autonomous driving, we constructed the patch attack experiment to demonstrate the possibility of such attempts. We select 10 scenarios where 40×40 adversarial patches can be put on more flat surfaces (e.g. Fig. 1). To preserve the depth of the scene, the ground truth disparities of the patches are the same as the corresponding part of the original image. For each image pair, we apply stereo-constrained PGD attacks with 100 iterations. The **Bad 3.0** results are shown in Fig. 3. Note that the errors are computed using the whole image, while only a small portion of the image is affected. **Our method is very robust against adversarial patches. In contrast, other methods perform poorly, even with adversarial training.** This experiment also demonstrates the over-fitting tendency of adversarial training towards certain types of attack. More illustrations are shown in the supplementary materials.

Experiment III): Comparisons with Adversarial Training. Our method increases the built-in adversarial robustness of stereo matching DNNs and thus it is orthogonal to existing defense methods such as adversarial training. From Table 2, our method without adversarial training shows comparable adversarial robustness with $\epsilon = 0.03, 0.06$, especially for EPE and Bad 1.0. For the patch attack experiment, our method is much more robust than others with adversarial training. **With adversarial training, our method has a stronger robustness than all other methods,** showing that our approach is indeed orthogonal

to adversarial training and they can be jointly used to further improved robustness.

Besides adversarial robustness of the trained dataset KITTI2015, we also test on KITTI2012 to see how the adversarial robustness generalize on unseen data. In Fig. 4, **our method shows a stronger cross-domain adversarial robustness than other adversarially trained methods.** Similarly, our method with adversarial training is still the most robust over all methods.

Experiment IV): Ablation Study. The census transform (CT) is chosen due to its non-differentiability and the fact that it is a well-rounded choice in the literature. We use multi-scale representations to respect the common recognition of its expressivity, and to alleviate choosing window size as a dataset-dependent hyper-parameter. We compare with traditional Sum of Absolute Difference (SAD) and show the results in Table 3. **We can see that CT is indeed much more robust than SAD due to its non-differentiability. CT with multiple scales has a stronger robustness than the single-scale version, while having a slightly better accuracy due to its flexibility.**

Experiment V): Leaderboard Comparisons. Table 4 shows the comparisons. Our method is slightly worse than state-of-the-art methods. As aforementioned, the detail of fine-tuning the pretrained model may play a significant role in the leaderboard comparisons. Our method is only fine-tuned by 140 training images in a vanilla manner without

Models	SceneFlow		KITTI15 (pretrained)		KITTI15 Attack ($\epsilon = 0.03$)	
	EPE [px]	Bad 3.0 [%]	EPE [px]	Bad 3.0	EPE [px]	Bad 3.0 [%]
multi-scale SAD	1.02	4.02	1.71	9.69	2.30	18.20
CT (w=11)	1.18	4.77	1.28	6.38	1.88	7.22
Ours w/o backbone	1.10	4.40	1.25	6.12	1.13	2.46
Ours	0.84	3.70	1.26	6.31	0.88	3.75

Table 3. Comparison with CT with window size 11 and multi-scale SAD

any bells and whistles. The gap may be bridged if more ablation studies are conducted to tune hyperparameters.

Models	Bad 3.0 [%]		Non-Occlusion		All Areas	
	FG	Avg All	FG	Avg All	FG	Avg All
GCNet [23]	5.58	2.61	6.16	2.87		
PSMNet [8]	4.31	2.14	4.62	2.32		
GANet-15 [48]	3.39	1.84	3.91	2.03		
GANet-Deep [48]	1.34	1.63	1.48	1.81		
LEAStereo [10]	2.65	1.51	2.91	1.65		
Ours	3.54	2.09	4.16	2.39		

Table 4. KITTI2015 leaderboard. FG: foreground regions.

6. Conclusions

This paper presents a novel workflow for stereo matching, which harnesses the best of classic features (multi-scale census transform) and end-to-end trainable DNNs for adversarially-robust and cross-domain generalizable stereo matching. The proposed method is motivated by the observation that DNN-based stereo matching methods can be deceived by a type of physically realizable attacks that entail stereo constraints in learning the perturbation. To address the adversarial vulnerability, this paper proposes to rethink the DNN feature backbone used in computing the cost volume by removing it for the matching stage. In experiments, the proposed method is tested in SceneFlow and KITTI2015 datasets with significantly better adversarial robustness and Sim2Real cross-domain generalizability (when no fine-tuning is used) achieved. It also obtains on-par performance on clean images.

References

- [1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283, Stockholm, Sweden, 10–15 Jul 2018. PMLR.
- [2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4312–4321. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Survey Track.
- [3] Michael Bleyer and Sylvie Chambon. Does color really help in dense stereo matching. Citeseer, 2010.
- [4] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.
- [5] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 3d adversarial object against msf-based perception in autonomous driving. In *Third Conference on Machine Learning and Systems (MLSys)*, March 2020.
- [6] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.
- [7] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, and John Duchi. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 2019.
- [8] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018.
- [9] Xinjing Chen, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *arXiv*, 2018.
- [10] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33, 2020.
- [11] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 9185–9193. IEEE Computer Society, 2018.
- [12] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4312–4321, 2019.
- [13] Xianzhi Du, Mostafa El-Khamy, and Jungwon Lee. Amnet: Deep atrous multiscale stereo disparity estimation networks. *arXiv*, 2019.
- [14] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36, 2018.
- [15] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

- [17] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- [18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [19] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [20] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814 vol. 2, 2005.
- [21] Heiko Hirschmuller and Daniel Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2008.
- [22] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *CoRR*, abs/1803.06373, 2018.
- [23] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 66–75, 2017.
- [24] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 36–42, 2018.
- [25] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR (Workshop)*. OpenReview.net, 2017.
- [26] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.
- [27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.
- [28] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.
- [29] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016.
- [31] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pages 506–519, 2017.
- [32] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pages 582–597. IEEE Computer Society, 2016.
- [33] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J. Black. Attacking optical flow. In *International Conference on Computer Vision (ICCV)*, 2019.
- [34] Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4322–4330, 2019.
- [35] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014.
- [36] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5019–5031, 2018.
- [37] Mahmood Sharif, Sruti Bhagavatula, Lujio Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1528–1540, 2016.
- [38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.
- [39] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.
- [40] James Tu, Huichen Li, Xinchun Yan, Mengye Ren, Yun Chen, Ming Liang, Eilyan Bitar, Ersin Yumer, and Raquel Urtasun. Exploring adversarial robustness of multi-sensor perception systems in self driving, 2021.
- [41] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13716–13725, 2020.
- [42] Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial pertur-

- bations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2021.
- [43] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1378–1387, 2017.
- [44] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 501–509, 2019.
- [45] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2730–2739, 2019.
- [46] Kyle Yee and Ayan Chakrabarti. Fast deep stereo with 2d convolutional processing of cost signatures. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [47] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 2016.
- [48] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [49] Zekun Zhang and Tianfu Wu. Learning ordered top-k adversarial attacks via adversarial distillation. In *Proceedings of CVPR 2020 Workshop on Adversarial Machine Learning in Computer Vision*, 2020.