Improving Domain Generalization in Self-Supervised Monocular Depth Estimation via Stabilized Adversarial Training

Yuanqi Yao[1], Gang Wu[1], Kui Jiang[1], Siao Liu[2], Jian Kuai[1], Xianming Liu[1], and Junjun Jiang[*1]

[1] Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China
{yuanqiyao,kuaijian}@stu.hit.edu.cn
{gwu,jiangkui,csxm,jiangjunjun}@hit.edu.cn
[2] Academy for Engineering & Technology, Fudan University
saliu20@fudan.edu.com

Abstract. Learning a self-supervised Monocular Depth Estimation (MDE) model with great generalization remains significantly challenging. Despite the success of adversarial augmentation in the supervised learning generalization, naively incorporating it into self-supervised MDE models potentially causes over-regularization, suffering from severe performance degradation. In this paper, we conduct qualitative analysis and illuminate the main causes: (i) inherent sensitivity in the UNet-alike depth network and (ii) dual optimization conflict caused by over-regularization. To tackle these issues, we propose a general adversarial training framework, named Stabilized Conflict-optimization Adversarial Training (SCAT), integrating adversarial data augmentation into self-supervised MDE methods to achieve a balance between stability and generalization. Specifically, we devise an effective scaling depth network that tunes the coefficients of long skip connection and effectively stabilizes the training process. Then, we propose a conflict gradient surgery strategy, which progressively integrates the adversarial gradient and optimizes the model toward a conflict-free direction. Extensive experiments on five benchmarks demonstrate that SCAT can achieve state-of-the-art performance and significantly improve the generalization capability of existing self-supervised MDE methods.

Keywords: Domain Generalization · Self-Supervised Monocular Depth Estimation · Stabilized Adversarial Training

1 Introduction

Monocular depth estimation (MDE) plays an important role in various 3D perceptional fields such as robotic navigation [17], autonomous driving [16] and 3D reconstruction [30]. However, due to the dynamic nature of the real world, even minor perturbations in the environment can result in significant domain shifts in
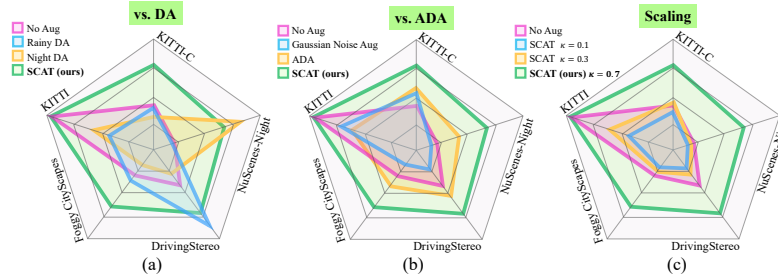
---

[*] Corresponding author

Fig. 1: Visualization of the domain generalization over various methods.(a) Comparisons with offline scenario-specific data augmentation methods.(b) Comparisons with vanilla Gaussian noise and vanilla adversarial data augmentation.(c) Comparisons of different LSCs scaling factors. The results show that our SCAT has excellent generalization performance under multiple unseen domains and retains the performance on the original training set.

the visual observations, which makes the trained model hard to generalize into unseen scenarios and restricts its application in the physical world.

To improve the generalization capability, several studies [19, 28, 66] utilize data augmentation methods to generate synthetic data and diversify the training environments, yielding considerable performance improvements. However, existing methods such as [19] mostly select some specific data augmentation schemes for target scenarios, yielding poor generalization performance in environments varying far from the augmented images. Compared with above mentioned offline data augmentation, adversarial data augmentation (ADA) does not make any target distribution assumption and synchronously optimizes the augmenter during the training phase, providing a promising pre-processing solution. Unfortunately, there is a dilemma in naively incorporating ADA into self-supervised MDE. Although adversarial data augmentation can effectively improve generalization capability in multiple supervised visual tasks [19], self-supervised MDE algorithms are quite sensitive to such excessive perturbation, resulting in significant performance degradation and training collapse. Therefore, it is necessary to rethink why self-supervised MDE cannot benefit from ADA as much as supervised learning.

In this work, we first conduct extensive quantitative analysis to investigate the causes of performance degradation when applying adversarial data augmentation to common self-supervised MDE models. There are two primary factors for this phenomenon: (i) inherent sensitivity of long skip connections (LSC) in UNet-alike depth estimation networks; (ii) dual optimization conflict caused by over-regularization. Specifically, as a core component in the MDE, LSC is widely adopted in the UNet-alike Depth Networks to combine multi-scale features and preserve low-level detail, yielding better prediction performance. However, the presence of these shortcut connections amplifies adversarial gradients, when coupled with pixel-level adversarial data augmentation, leading to severe training instability and collapse. Moreover, compared to offline data augmentation, ad-

versarial augmentation data commonly act as the worst-case training examples [36, 37] and often provides over-regularization for the model training, which results in optimization gradients that oppose those of the original data, leading to severe gradient conflicts and a subsequent decline in convergence performance.

To tackle these issues, we propose a general adversarial training framework, named **S**tabilized **C**onflict-optimization **A**dversarial **T**raining (**SCAT**), tailored for self-supervised monocular depth estimation. SCAT is designed to enhance both generalization capability and stability. In particular, we develop an innovative scaling depth network that adjusts the coefficients of long skip connections within UNet architecture, thereby ensuring a more stable training regime through a theoretically supported. Furthermore, we advance a Conflict Gradient Surgery method, progressively blending the adversarial gradient to guide the model optimization along a direction devoid of conflicts. It is worth noting that one of the merits of our method SCAT is model-agnostic, and the stabilized adversarial training can be applied to diverse SOTA methods, yielding consistent generalization enhancement. To validate the effectiveness of SCAT, we conduct extensive experiments on both KITTI and KITTI-C datasets, demonstrating its cross-domain generalization capability on Foggy CityScapes, DrivingStereo and NuScenes datasets. In summary, our contributions are summarized as follows:

- We point out the inherent sensitivity of UNet-based depth networks and the dual optimization conflict caused by over-regularization, providing a theoretical analysis of the self-supervised MDE model's instability.
- We propose a general adversarial training framework named Stabilized Conflict-optimization Adversarial Training (SCAT), which efficiently incorporates adversarial augmentation into self-supervised MDE, and significantly improves the model generalization capabilities across multiple unseen domains.
- We developed Conflict Gradient Surgery (CGS) to address the dual optimization conflicts induced by adversarial over-regularization, achieving a balance between model stability and generalization.

2   Related Works

2.1   Self-Supervised Monocular Depth Estimation

Self-supervised monocular depth estimation (MDE) has made remarkable progress in recent years, enabling depth learning from unlabeled data. The seminal work by Zhou et al. [23] laid the foundation by jointly optimizing depth and pose networks using an image reconstruction loss. Since then, two main approaches have emerged: stereo training and monocular training. Stereo training methods utilize synchronized stereo image pairs to predict disparity maps, with notable contributions including photometric consistency loss [4], left-right consistency [6], and continuous disparity prediction [45]. Monocular training, on the other hand, relies on the consistency between synthesized and actual scene views, with SfM-Learner [23] pioneering the joint training of a DepthNet and a PoseNet using a photometric loss. Researchers have also explored complementary techniques

to further boost performance, such as occlusion modeling [38–41], scale ambiguity [42] advanced network architectures [12, 43, 44], semantic information integration [46–50], and object size cues [48, 51]. The proposed use of orthogonal planes for improved depth representation in driving scenarios contributes to the ongoing advancements in this field.

2.2   Domain Generalization in Monocular Depth Estimation

A line of studies investigated how to utilize data augmentation or synthetic data generation to improve the domain generalization capability of MDE [19, 28, 29, 52–56]. ADDS-DepthNet [28] generates night image pairs from day images using GANs, training the network to be effective in both day and night domains. It uses day depth estimates as pseudo-supervision for night scenes, which limits the night depth estimation to the accuracy of the day estimates. Additionally, ADDS-DepthNet focuses on reconstructing GAN-generated night images, which generates reconstruction targets with high variance, resulting detrimental to the self-supervised MDE algorithm. ITDFA [56] employs a fixed depth decoder and adapts the encoders for each domain to enforce feature consistency across domains, which is limited to the performance. Robust-Depth [19] put forward an offline data augmentation method. By exploiting the correspondence between unaugmented and augmented data they introduce a pseudo-supervised loss for both depth and pose estimation. However, previous approaches limited the ability to handle environments with pre-defined variations and require substantial architectural modifications to obtain realistic depth estimates, leading to severe domain bias when applied to real-world scenes [19, 28].

2.3   Adversarial Training

Adversarial training has emerged as a popular technique to improve the generalization capability of deep neural networks. This approach involves incorporating both benign and adversarial examples during the training process, enabling models to learn more robust features [57]. Researchers have successfully applied adversarial training to various computer vision tasks, such as image classification [26, 57], object detection [58, 59], and segmentation [60, 61]. However, generating adversarial examples typically requires ground truth labels, which limits the applicability of adversarial training to supervised learning settings. To address this limitation, semi-supervised adversarial learning methods [62, 63] have been proposed, which leverage a small portion of labeled data to enhance robustness. Additionally, contrastive learning [64, 65] has been combined with adversarial examples to improve self-supervised learning and model robustness. In the context of MDE, adversarial training remains largely unexplored, particularly in self-supervised MDE which absent the ground truth depth information.

3   Preliminaries

3.1   Problem Formulation

Following [5, 7], we regard the self-supervised depth estimation as an image reconstruction task, where the objective is to reconstruct the target current frame image I_t from consecutive frames $I'_t \in \{I_{t-1}, I_{t+1}\}$. Specifically, we construct a dual-branch framework, involving a DepthNet and a PoseNet. The former is used for estimating depth map D_t while the latter is introduced to predict the relative poses $T_{t'\to t}$ through the inverse warping as below, the predicted depth map D_t in conjunction with the relative poses $T_{t'\to t}$ can synthesize the target image $I_{t'\to t}$, depicted as:

$$I_{t'\to t} \;=\; I_{t'}\Big\langle proj\big(D_t, T_{t\to t'}, K\big)\Big\rangle, \tag{1}$$

where $proj(\cdot)$ generates the resulting 2D coordinates of the projected depths Dt in It' and $\langle\rangle$ refers to the sampling operator. We minimize the photometric reprojection error $\mathcal{L}_p$ to supervise the DepthNet,

$$\mathcal{L}_p \;=\; \sum_{t'} pe(I_t, I_{t'\to t}), \tag{2}$$

where pe is the photometric reconstruction error. As suggestion in [6, 13], we adopt the weighted combination of L_1 loss and SSIM [24] to formulate the photometric error function pe, denoted as:

$$pe(I_a, I_b) = \frac{\alpha}{2}(1 - \text{SSIM}(I_a, I_b)) + (1-\alpha)||I_a - I_b||_1. \tag{3}$$

3.2   Analysis on Unstable Self-Supervised MDE

In this section, we aim to investigate the primary causes why naively applying adversarial data augmentation in self-supervised MDE leads to instability.
Pitfall1: Sensitive UNet-alike Depth Network. In previous self-supervised MDE methods [7–9], UNet [32] is the most popular depth network backbone, since its long skip connects (LSCs) to connect distant network blocks can fuse multi-granularity semantic information and spatial information:

$$\text{DepthNet}(x) = f_0(x),\;\; f_i(x) = b_{i+1} \circ [a_{i+1} \circ x + f_{i+1}\left(a_{i+1} \circ x\right)], \tag{4}$$

where $x \in R^m$ denotes the input, a_i and $b_i (i \ge 1)$ are the trainable parameter of the i-th block. For the vector operation $\circ$, it can be designed to implement different networks. Assume an extra noise $\epsilon_\delta \sim N(0,\; \sigma_\delta^2 I)$ is injected into I_t which yields a new input $\tilde{I}'_t$, self-supervised MDE with data augmentations aims to reprojection $I_{t'\to t}$ from the noisy input $\tilde{I}_{t'}$. Accordingly, if the variance σ_δ^2 of extra noise is large, it can hinder re-projecting the desired target image $I_{t'\to t}$. In adversarial training, the variance of this extra noise varies along training iterations, further exacerbating the instability of self-supervised MDE training.

Recent work [35] also reveals a similar phenomenon, which demonstrates that the UNet-based segmentation architecture is sensitive to perturbed noise, suffering from catastrophic training collapse. It is noteworthy that, although our theoretical analysis is focused on UNet-based depth estimation architectures, the LSC component is extensively employed across various mainstream architectures [11], thereby lending a degree of extensibility to our empirical results.

Pitfall2: Dual Optimization Conflict by Over-Regularization. Compared to the offline data augmentation in Robust-Depth [19], Adversarial Data Augmentation (ADA) makes no assumptions about the target distribution and synchronously optimizes augmenters during the training phase, offering a promising pre-processing solution. Unfortunately, naively incorporating ADA into self-supervised Monocular Depth Estimation (MDE) presents a dilemma. While adversarial data augmentation can effectively enhance generalization in various supervised visual tasks, a dual optimization conflict arises due to the opposing gradient optimization directions between the adversarial-perturbed data and the original data. Similar to [31], we also define ϕ_{ij} as the angle between origin data g_i and adversarial-perturbed data gradient g_j, the gradients as conflicting when $cos\phi_{ij} < 0$. Therefore, for self-supervised MDE, over-regularization leads to ADA dominating the gradient updates, resulting in significant performance degradation and training collapse. We refer to this issue as a dual optimization conflict caused by over-regularization.

4   Method

4.1   Architectural Overview

In this section, we propose a general adversarial training framework named Stabilized Conflict-optimization Adversarial Training (SCAT), which is compatible with any self-supervised MDE method. An overview of the SCAT architecture is provided in Fig. 2. To circumvent erroneous bootstrapping from adversarial augmented data, augmentation was applied to improve generalization by reconstructing the unaugmented I_t, rather than the augmented image $\tilde{I}_t$.

$$\tilde{I}_{t'} = I_{t'} + \delta \tag{5}$$

$$\tilde{I}_{t' \to t} \; = \; I_{t'} \Big\langle proj\big(\tilde{D}_t, \tilde{T}_{t \to t'}, K\big) \Big\rangle, \tag{6}$$

$$\mathcal{L}_p \; = \; \sum_{t'} pe(I_t, I_{t' \to t}) + pe(I_t, \tilde{I}_{t' \to t}). \tag{7}$$

If δ is learned (i.e. SCAT is implemented with an adversarial generator $\delta = g_\phi(z)$, to introduce adversarial training into self-supervised MDE, we calculate the reprojection loss for the current frame image reconstructed from the perturbed images and use it as the self-supervised constraint loss $\mathcal{L}_{AD}$:

$$\mathcal{L}_{AD} \; = \; \sum_{t'} pe(I_t, \tilde{I}_{t' \to t}). \tag{8}$$
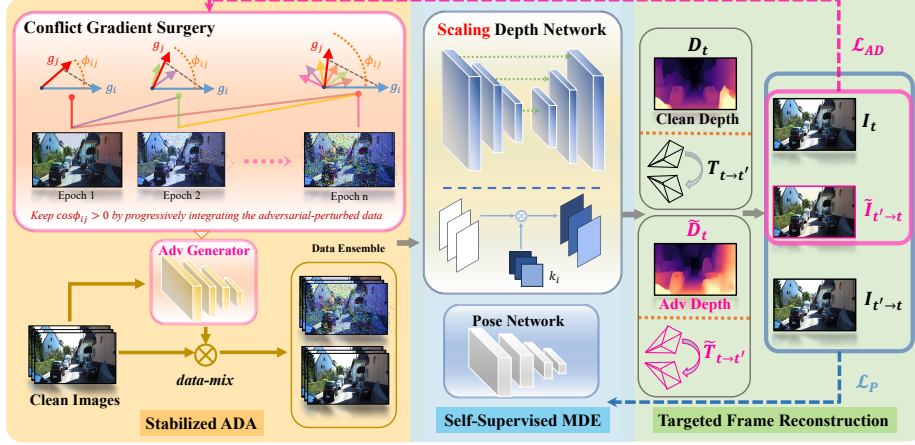
Fig. 2: Overview of our SCAT architecture. Our method introduces an adversarial noise generator, which is optimized through $\mathcal{L}_{AD}$ and acts as an adversarial constraint, while $\mathcal{L}_p$ imposes self-supervised constraints to optimize self-supervised MDE model. CGS incrementally applies adversarial augmenters from multiple iterations, enabling a balanced achievement in model generalization and stability. Meanwhile, we utilize a scaling depth network(SDN) to stabilize the training process.

Our goal is to find a noise distribution $p_\phi(\delta)$, $\delta \in R^N$ and $\|\delta\|_2{=}\epsilon$, such that noise samples added to input clean image I_t maximally confuse the depth network. More concisely, we optimize the following equation, depicted as

$$\min_\theta \max_\phi \mathbb{E}_{\tilde{I}_{t'},I_t} \mathbb{E}_{\delta\sim p_\phi(\delta)} \left[\mathcal{L}\left(f_\theta(\tilde{I}_{t'}), I_t \right) \right]. \tag{9}$$

Specifically, max_ϕ represents the adversarial noise generator's attempt to maximize the $\mathcal{L}_{AD}$, thereby generating adversarial samples that maximally perturb the depth network. Conversely, min_θ denotes the depth network's effort to minimize the $\mathcal{L}_p$ on these adversarial samples, aiming to achieve robustness in depth estimation across various adversarial inputs. Through training iterations provided in Algorithm 1 in our SCAT framework, self-supervised MDE models becomes proficient at accurately inferring various perturbed images.

However, as the variance of this extra adversarial noise varies along training, the standard UNet-based depth network is too sensitive to these perturbations. Hence, we propose a Scaling Depth Network (SDN), which appropriately scales the coefficients κ_i to lower reconstruction error. In addition, conflict arises by opposing gradient optimization directions between the adversarial-perturbed data and original data, further exacerbating the training instability. To ensure the model has enough representational power while preventing unstable parameter updates, we propose Conflict Gradient Surgery (CGS) to refine the adversarial training gradient update process, which we introduce in the following section.

4.2   Scaling Depth Network

To alleviate the impact of extra noise, we devise scaling depth network (SDN) to scale the coefficients κ_i of LSCs :

$$\text{DepthNet}(x) = f_0(x), \ f_i(x) = b_{i+1} \circ [k_{i+1} \cdot a_{i+1} \circ x + f_{i+1}(a_{i+1} \circ x)], \quad (10)$$

where $\kappa_i > 0, (i \geq 1)$ are the scaling coefficients and are set to 1 in standard UNet-based depth network. For scaling depth network in Eq.10, assume $M_0 = \max\{||b_i \circ a_i||_2, 1 \leq i \leq N\}$ and f_N is $L_0 - Lipschitz \ continuous$, c_0 is a constant related to M_0 and L_0. Suppose $I_t^{\epsilon_\delta}$ is an perturbed input of the vanilla input I_t with a small perturbation $\epsilon_\delta = ||I_t^{\epsilon_\delta} - I_t||_2$. Then we have

$$||f_\theta(\mathbf{I}_t^{\epsilon_\delta}) - f_\theta(\mathbf{I}_t)||_2 \ \leq \ \epsilon_\delta \left[\sum_{i=1}^{N} \kappa_i M_0^i + c_0\right], \quad (11)$$

where N is the number of the long skip connections. See the proof in Appendix.

Eq.11 shows that for a perturbation magnitude ϵ_δ, the reconstruction error bound of depth network is $\mathcal{O}(\epsilon_\delta(\sum_{k=1} K_i M_0^i)$. For standard UNet-based depth network ($\kappa_i = 1 \forall i$), this bound becomes a very large bound $\mathcal{O}(N M_0^N)$. This implies that a standard UNet-based depth network is sensitive to extra noise in the input, especially when LSC number N is large. Intuitively, setting appropriate scaling coefficients κ_i in our scaling depth network can enhance the robustness of the model to input perturbations, thereby improving better stability for self-supervised MDE. To derive the scaling coefficient κ, we adopt a heuristic algorithm to assign a uniform constant value of κ_* across all LSCs, with the default set to 0.7. In terms of implementation, a line of parameter optimization techniques [33,35] can be applied, e.g. incorporating an adaptive projection layer [35] for each LSC component. The empirical discussions in the Appendix demonstrate that a singular scaling parameter κ is sufficient to strike to balance performance and efficiency for the self-supervised MDE task.

4.3   Conflict Gradient Surgery

Here, we propose to modify the standard adversarial data augmentation (ADA) by incorporating conflict gradient surgery named CGS, aiming to change the distribution of $cos(\theta)$ tends towards positive values, where θ is the angle between the gradient of the original data and the gradient of the CGS adversarial data, so that the gradient update guided by the adversarial data points in the direction of the conflict-free direction, and finally improves the robustness of the model to perturbations:

$$\mathbb{E}\left[\cos(\theta)\right] > 0, \quad \text{where} \ \cos(\theta_i) = \frac{g_i \cdot g_{\text{mix}_i}}{|g_i||g_{\text{mix}_i}|}$$

where g_i is the gradient of the i-th original data, g_{mix_i} is the gradient of the i-th mixed adversarial data. We calculate the gradient cosine similarity between

the gradient of the adversarial-perturbed data and the original data. We define two gradients as conflicting if their cosine similarity is negative, indicating they are diverging from one another. If the distribution of cosine similarities tends to be negative, it signifies the presence of numerous adversely oriented gradients, which can mislead the gradient optimization process. As illustrated in Fig. 3, we extracted 1000 images from the KITTI dataset and recorded the gradient cosine similarity between pairs of adversarial perturbation data and original data. Without employing conflict gradient surgery, the gradient conflict becomes a prevalent issue as the adversarial training progresses iteratively, consequently leading slower convergence and performance degradation.
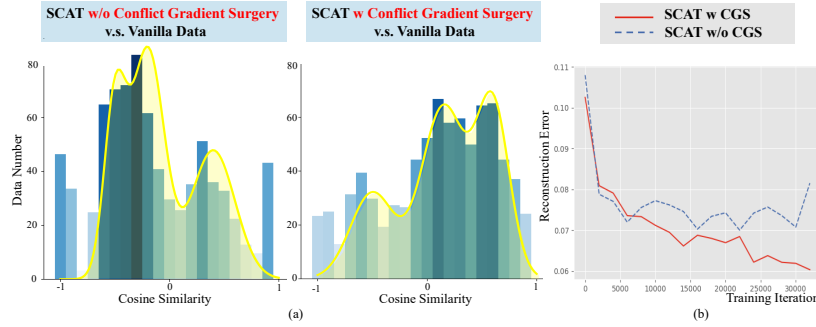

Fig. 3: (a) Statistics of gradient cosine similarity. Through CGS, we have shifted the distribution of cosine similarities from being negatively skewed to positively skewed, mitigating the previously prevalent issue of adversely oriented gradients. **(b) Illustration of training oscillation issue arising from Dual Optimization Conflict.**

Experimental observations provided in Fig. 3 demonstrates that our approach ensures the expectation of $cos(\theta)$ is positive, thereby skewing the distribution of $cos(\theta)$ tends towards positive values, so that the gradient update guided by the adversarial data points in the direction of the conflict-free direction, ultimately enhancing the model's robustness against perturbations. Specifically, given a set of gradient vectors (one for each adversarial data augmentation), we construct the adversarial perturbation data while employing multiple adversarial generators from previous iterations, restraining the over-regularization caused by the conflict component through our conflict gradient surgery. The update procedure of the CGS algorithm is provided in Algorithm 1.

5   Experiments

5.1   Experimental Settings

KITTI [1] KITTI dataset is collected for mobile robotics and autonomous driving, involving hours of traffic scenarios recorded with a variety of sensor modalities, including high-resolution RGB, grayscale stereo cameras, and a 3D laser scanner. Following the setting in [5], we use 39,810 images for training and 4,424 for validation. Subsequently, we rigorously evaluate the proposed method and other compared methods on the KITTI eigen test dataset [3].

Algorithm 1 Generic SCAT self-supervised MDE algorithm
($\triangleright$ vanilla self-supervised MDE, ▶ our modifications)

1: **for** Epoch $e = 1...n$ do **do**
2:     **Adversarial Data Augmentation via CGS:**
3:         $g_{1:j} \sim Sample(\mathcal{B}, j)$ ▶Randomly select generators from History Buffer.
4:         $\delta^{1:j} = g_{1:j}(I_t)$ ▶Generate adversarial noise from clean images.
5:         $\tilde{I}_{t'}^{1:j} = I_t + \delta^{1:j}$ ▶Obtain adversarial images.
6:     **Deterministic Reprojection with SDN:**
7:         $D_t = ScalingDepthNet(I_t), P_t = PoseNet(I_t)$
      $\triangleright$ Obtain clean depth map and camera poses.
8:         $\tilde{D}_t = ScalingDepthNet(\tilde{I}_t), \tilde{P}_t = PoseNet(\tilde{I}_t)$
      ▶ Obtain adversarial depth map and camera poses.
9:         $\tilde{I}_{t' \to t} \Leftarrow f_\theta(\tilde{I}_{t'}), I_{t' \to t} \Leftarrow f_\theta(I_{t'})$
10:        $I_t \simeq \tilde{I}_{t' \to t}, I_{t' \to t}$
      ▶ Reconstruct adversarial and clean images with the same target.
11:        $\mathcal{B} \leftarrow \mathcal{B} \cup g_e$ ▶Add current Generator into History Buffer $\mathcal{B}$.
12:     **Module Optimization via SGD:**
13:        $\theta \leftarrow \theta - \eta_\theta \nabla \theta \mathcal{L}_P(f, I_t; \theta)$ $\triangleright$Optimize MDE f_θ with $\mathcal{L}_P$ in Eq. 7.
14:        $\phi \leftarrow \phi + \eta_\phi \nabla \phi \mathcal{L}_{AD}(g_\phi, I_t; \phi)$ ▶Optimize Generator g_ϕ with $\mathcal{L}_{AD}$ in Eq. 8.
15: **end for**

Table 1: Quantitative Results for the KITTI-C Dataset. Pink cell denotes the best performance, while yellow one denotes suboptimal.

Method	W × H	mCE(%)↓	mRR(%)↑	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
MonoDepth2 [7]	640 × 192	101.04	84.08	0.248	1.764	6.852	0.291	0.698	0.874	0.944
+ DA	640 × 192	97.72	84.96	0.227	1.761	6.735	0.293	0.710	0.876	0.944
+ ADA	640 × 192	89.31	87.31	0.193	1.632	6.427	0.278	0.734	0.882	0.945
+ SCAT (ours)	640 × 192	86.32	90.13	0.165	1.573	6.157	0.269	0.762	0.893	0.951
CADepth [8]	640 × 192	108.10	80.12	0.271	1.902	6.900	0.302	0.674	0.860	0.938
+ DA	640 × 192	107.14	81.31	0.261	1.874	6.832	0.299	0.691	0.865	0.939
+ ADA	640 × 192	99.81	83.13	0.235	1.866	6.565	0.278	0.721	0.871	0.940
+ SCAT (ours)	640 × 192	95.54	87.93	0.197	1.794	6.258	0.264	0.747	0.883	0.948
HR-Depth [9]	640 × 192	103.67	82.88	0.255	1.960	6.909	0.309	0.670	0.854	0.933
+ DA	640 × 192	101.74	83.23	0.243	1.942	6.787	0.294	0.685	0.859	0.936
+ ADA	640 × 192	93.84	85.32	0.215	1.831	6.625	0.278	0.699	0.861	0.939
+ SCAT (ours)	640 × 192	87.06	89.47	0.189	1.732	6.204	0.258	0.741	0.871	0.945
MonoVit [11]	640 × 192	80.54	88.98	0.191	1.274	5.998	0.245	0.771	0.922	0.964
+ DA	640 × 192	78.32	90.47	0.187	1.256	5.873	0.228	0.787	0.931	0.965
+ ADA	640 × 192	72.41	92.34	0.166	1.134	5.626	0.211	0.819	0.935	0.969
+ SCAT (ours)	640 × 192	62.74	95.38	0.127	1.058	5.274	0.208	0.846	0.947	0.976
Robust-Depth* [19]	640 × 192	55.72	96.46	0.121	0.954	5.051	0.201	0.854	0.952	0.978
+ DA	640 × 192	60.12	96.34	0.124	0.963	5.012	0.213	0.852	0.951	0.977
+ ADA	640 × 192	54.23	97.59	0.119	0.943	4.985	0.199	0.853	0.951	0.977
+ SCAT (ours)	640 × 192	53.37	98.19	0.117	0.932	4.872	0.187	0.861	0.955	0.979

KITTI-C [67] To evaluate the robustness and safety of our method in out-of-distribution (OoD) scenarios, we utilize KITTI-C dataset as a benchmark. The diverse corruptions enable the reliability of simulation to the potential perturbation distribution in real-world scenarios.

DrivingStereo [14] It contains 500 images captured under different weather conditions, involving foggy, cloudy, rainy, and sunny. These images offer a realistic representation of various driving scenarios, enabling robust evaluation of depth estimation performance across diverse domains.

Foggy CityScapes [2] Foggy CityScapes is a synthetic fog dataset that simulates real-world foggy scenarios. Each foggy sample is generated using the corresponding clear image and depth map from CityScapes dataset.

$\epsilon_m = 20$ $\epsilon_m = 40$ $\epsilon_m = 80$ $\epsilon_m = 135$

Fig. 4: Example images with different ϵ_m of adversarial perturbation.

NuScenes [15] The NuScenes dataset comprises approximately 15,000 images captured in real nighttime urban street settings. To validate the model generalization, NuScenes is also selected as an evaluation benchmark in our study.

Table 2: Quantitative Results for the KITTI Eigen Test Dataset. The results indicate that SCAT largely maintains the performance of all baselines compared to vanilla ADA, ensuring the ability to infer accurate depth information on clean images.

Method	W × H	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [7]	640 × 192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
+ ADA	640 × 192	0.121	0.978	4.992	0.221	0.862	0.953	0.978
+ **SCAT (ours)**	640 × 192	0.116	0.942	4.877	0.193	0.877	0.958	0.981
CADepth [8]	640 × 192	0.105	0.769	4.535	0.181	0.892	0.964	0.983
+ ADA	640 × 192	0.108	0.812	4.585	0.190	0.887	0.962	0.982
+ **SCAT (ours)**	640 × 192	0.106	0.796	4.579	0.183	0.890	0.963	0.983
DIFFNet [10]	640 × 192	0.102	0.749	4.445	0.179	0.897	0.965	0.983
+ ADA	640 × 192	0.106	0.793	4.592	0.184	0.889	0.963	0.982
+ **SCAT (ours)**	640 × 192	0.103	0.764	4.453	0.180	0.897	0.965	0.983
MonoVit [11]	640 × 192	0.099	0.708	4.372	0.175	0.900	0.967	0.984
+ ADA	640 × 192	0.106	0.733	4.591	0.184	0.897	0.963	0.982
+ **SCAT (ours)**	640 × 192	0.100	0.716	4.389	0.175	0.899	0.967	0.984
Robust-Depth* [19]	640 × 192	0.100	0.747	4.455	0.177	0.895	0.966	0.984
+ ADA	640 × 192	0.105	0.783	4.698	0.182	0.889	0.953	0.981
+ **SCAT (ours)**	640 × 192	0.101	0.754	4.459	0.177	0.895	0.966	0.984

5.2   Effect of Perturbation Sizes

To evaluate the effect of adversarial perturbation magnitude on model generalization, here we employ ϵ_m to perturbation sizes:

$$\epsilon_m = \min\|\delta_{\min}\|_2. \tag{12}$$

We measure model generalization capability by calculating the median perturbation size ϵ_m and report the results in Tab. 3. To provide a better intuition for the noise level for particular ϵ_m, we display example images in Fig. 4.

5.3   Results on KITTI-C and KITTI

In this section, we compare the depth estimation performance of our SCAT on the KITTI-C and KITTI datasets. Quantitative results on the KITTI-C dataset are shown in Tab. 1. We conducted extensive experiments on multiple baseline models, comparing it with the generalization performance between offline data augmentation (DA) and vanilla adversarial data augmentation (ADA). As

Table 3: Qualitative Results for different perturbation sizes. We compare the results obtained by $\epsilon_m = 135.0$ adversarial noise generator with its counterparts of different perturbation sizes. Note that $\epsilon_m = 135.0$ was used in all experiments in this paper apart from this ablation study.

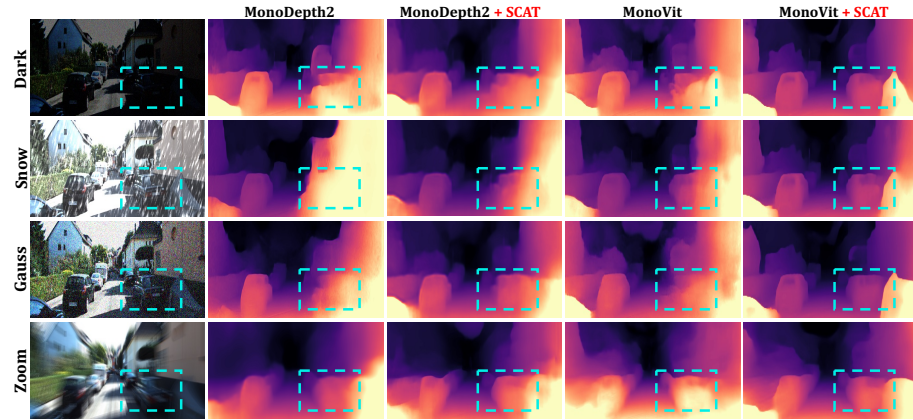| Model	MCE(%)↓	mRR(%)↑	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
MonoDepth2 [7]	101.04	84.08	0.248	1.764	6.852	0.291	0.698	0.874	0.944
$+ g_\phi \ \epsilon_m = 20$	93.96	86.17	0.196	1.619	6.385	0.285	0.738	0.883	0.947
$+ g_\phi \ \epsilon_m = 40$	90.15	86.93	0.187	1.613	6.351	0.282	0.741	0.885	0.949
$+ g_\phi \ \epsilon_m = 80$	88.95	87.21	0.182	1.602	6.288	0.279	0.749	0.890	0.949
$+ g_\phi \ \epsilon_m = 135$	**86.32**	**90.13**	**0.165**	**1.573**	**6.157**	**0.269**	**0.762**	**0.893**	**0.951**
$+ g_\phi \ \epsilon_m = 180$	88.17	86.83	0.177	1.594	6.195	0.277	0.758	0.891	0.950

Fig. 5: Qualitative Results for KITTI-C. As the SOTA self-supervised MDE methods, MonoDepth and MonoVit excel on the KITTI dataset, but struggle to accurately infer depth information from various types of damaged images in out-of-distribution (OoD) domains. With our SCAT framework, their depth estimation performance in challenging cross-domain scenarios can be substantially improved.

expected, our method demonstrates a significant advantage on the mean corruption error (mCE) and mean resilience rate (mRR) scores. Especially for 18 common out-of-distribution corruption types, with all baseline models, our proposed SCAT method outperforms the competitors in almost all of metrics, providing a general framework to enhance cross-domain generalization capability for existing self-supervised monocular depth estimation methods.

5.4   Results on Cross-Domain Generalization

To demonstrate the generalization capability of our SCAT over real cross-domain scenarios, like foggy and nighttime samples, we also conduct the experiments on Foggy CityScapes dataset, DrivingStereo dataset and NuScenes dataset, which are shown in Fig. 6 and Fig. 7. As expected, our SCAT generates depth estimation results with more reliable and clearer contours in all weather conditions. By contrast, vanilla MonoDepth2 and MonoVit suffer from significant performance
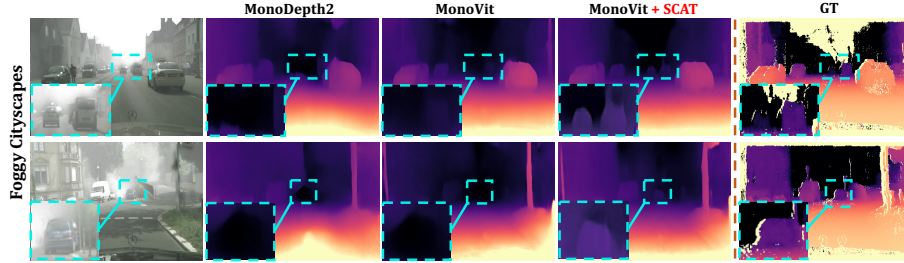
Fig. 6: Qualitative Results for Foggy CityScapes. We conduct testing to assess SCAT's generalization on a synthetically generated dataset with more severe fog.

Table 4: Quantitative Results for the Real-World NuScenes-Night Dataset.

Dataset	Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
NuScenes-Night [15]	Monodepth2 [7]	0.397	6.206	14.569	0.568	0.378	0.650	0.794
	+ **SCAT (ours)**	**0.355**	**5.112**	**12.436**	**0.499**	**0.392**	**0.671**	**0.863**
	HR-Depth [9]	0.461	6.633	15.028	0.622	0.301	0.571	0.749
	+ **SCAT (ours)**	**0.424**	**5.802**	**13.685**	**0.602**	**0.324**	**0.584**	**0.827**
	CADepth [8]	0.421	5.949	14.509	0.593	0.331	0.613	0.776
	+ **SCAT (ours)**	**0.387**	**5.524**	**11.962**	**0.577**	**0.365**	**0.655**	**0.834**
	MonoVit [11]	0.313	4.143	12.252	0.455	0.485	0.736	0.858
	+ **SCAT (ours)**	**0.284**	**4.126**	**10.139**	**0.401**	**0.497**	**0.752**	**0.879**
	Robust-Depth* [19]	0.276	4.075	10.470	0.380	0.607	0.819	0.912
	+ **SCAT (ours)**	**0.263**	**3.973**	**9.462**	**0.371**	**0.609**	**0.819**	**0.913**

decline when handling real and unknown scenarios, in particular in the rainy and nighttime scenes. Notably, it can be observed that our SCAT method can effectively recover challenging distant objects such as vehicles or streetlights when dealing with complex and unknown images in real-world cross-domain scenarios. Quantitative results are summarized in Tab. 4, showing that our proposed SCAT framework gains better performance on all metrics with multiple state-of-the-art methods. More results are shown in the Appendix.

5.5 Ablation Study

Effectiveness of Individual Components. To improve the stability and generalization capability of self-supervised MDE methods, SCAT presents a stabilized adversarial training framework with Conflict Gradient Surgery (CGS), progressively integrating the adversarial-perturbed data to prevent gradient conflict. In addition, we propose a scaling depth network named SDN, reducing the perturbation sensitivity of the UNet-based depth network. We perform an ablation study to investigate the effectiveness of individual components in SCAT and the results are shown in Tab. 5. Individually, each of these contributes significantly to the improvement of generalization performance across multiple datasets.

Impact of LSCs Coefficient κ. In our experiments, the value of the long skip connections (LSCs) coefficient κ was set to balance the self-supervised MDE model's stability and generalization capability, which defaults to 0.7 in the paper. To assess the perturbation sensitivity of UNet-based depth networks, we
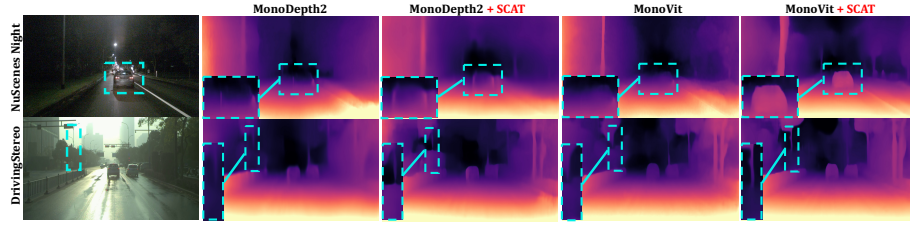
| | MonoDepth2 | MonoDepth2 + SCAT | MonoVit | MonoVit + SCAT |

Fig. 7: Qualitative results on Driving Stereo and Nuscenes. It can be observed that our SCAT method effectively recovers challenging distant objects such as vehicles or streetlights when dealing with complex and unknown images in real scenarios.

compare the model performance over constant values of $\kappa \in \{0.1, 0.3, 0.7, 1.0\}$. Fig. 1 shows that when κ is set to 0.7, our scaling depth network has extraordinary generalization capability to unseen scenarios in the KITTI-C dataset while keeping the depth estimation performance of the KITTI dataset.

CGS	SDN	KITTI Abs Rel	$\delta < 1.25$	KITTI-C Abs Rel	$\delta < 1.25$
		0.121	0.862	0.193	0.734
✓		0.117	0.869	0.174	0.752
	✓	0.118	0.865	0.179	0.748
✓	✓	0.116	0.877	0.165	0.762

Table 5: Effect of individual components.
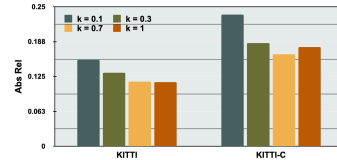

Fig. 8: Effect of the scaling factor κ.

6 Conclusion

In this work, we initially illuminate the main causes of training collapse under adversarial data augmentation: (i) inherent sensitivity in the UNet-alike depth network and (ii) dual optimization conflict caused by over-regularization. To address these issues, we propose SCAT, a simple yet effective model-agnostic framework that leverages stabilized adversarial training to bootstrap the generalization capability of self-supervised MDE models in unseen scenarios. Extensive experiments on a variety of scenarios from five benchmarks validate the merits of our adversarial data augmentation, which endows the model with universal generalization capability and better training efficiency. In addition, empirical results indicate that each component in our method and the utilization of targeted reconstruction are crucial for generalization performance gain. Our exploration may inspire more researchers to dig into the great potential of utilizing adversarial training in self-supervised MDE.

7 Limitation

In the future, we would like to explore iterative adversarial data augmentation [34], which is a more promising parameter-efficient solution.

References

1. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.
2. C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, 2018.
3. D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing System (NeurIPS)*, 2014.
4. R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 740–756.
5. T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1851–1858.
6. C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 270–279.
7. C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth prediction," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3828–3838.
8. J. Yan, H. Zhao, P. Bu, and Y. Jin, "Channel-wise attention-based network for self-supervised monocular depth estimation," in *IEEE International Conference on 3D Vision (3DV)*, 2021, pp. 464–473.
9. X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, and Y. Yuan, "Hr-depth: High resolution self-supervised monocular depth estimation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 2294–2301.
10. H. Zhou, D. Greenwood, and S. Taylor, "Self-supervised monocular depth estimation with internal feature fusion," in *British Machine Vision Conference (BMVC)*, 2021.
11. C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "MonoVit: Self-supervised monocular depth estimation with a vision transformer," in *IEEE International Conference on 3D Vision (3DV)*, 2022.
12. S. Pillai, R. Ambru, and A. Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9250–9256.
13. H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging (TCI)*, vol. 3, no. 1, pp. 47–57, 2016.
14. L. Wang, X. Sun, W. Jiang, and J. Yang, "Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5554–5561.
15. H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv preprint arXiv:1903.11027*, 2019.

16. 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

17. X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "Towards real-time monocular depth estimation for robotics: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 16 940–16 961, 2022.

18. Z. Cheng, J. Liang, G. Tao, D. Liu, and X. Zhang, "Adversarial training of self-supervised monocular depth estimation against physical-world attacks," *arXiv preprint arXiv:2301.13487*, 2023.

19. K. Saunders, G. Vogiatzis, and L. Manso, "Self-supervised monocular depth estimation: Let's talk about the weather," *arXiv preprint arXiv:2307.08357*, 2023.

20. J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *2019 IEEE winter conference on applications of computer vision (WACV)*.   IEEE, 2019, pp. 1043–1051.

21. F. Aleotti, F. Tosi, M. Poggi, and S. Mattoccia, "Generative adversarial networks for unsupervised monocular depth prediction," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

22. Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, "Single view stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 155–163.

23. T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

24. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

25. J. Bae, S. Moon, and S. Im, "Deep digging into the generalization of self-supervised monocular depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 187–196.

26. E. Rusak, L. Schott, R. S. Zimmermann, J. Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel, "A simple way to make neural networks robust against diverse image corruptions," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*.   Springer, 2020, pp. 53–69.

27. L. Kong, S. Xie, H. Hu, L. X. Ng, B. Cottereau, and W. T. Ooi, "Robodepth: Robust out-of-distribution depth estimation under corruptions," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

28. L. Liu, X. Song, M. Wang, Y. Liu, and L. Zhang, "Self-supervised monocular depth estimation for all day images using domain separation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 737–12 746.

29. A. Gurram, A. F. Tuna, F. Shen, O. Urfalioglu, and A. M. López, "Monocular depth estimation through virtual-world supervision and real-world sfm self-supervision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12 738–12 751, 2021.

30. J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. T. Freeman, and J. B. Tenenbaum, "Learning shape priors for single-view 3d completion and reconstruction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 646–662.

31. T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

32. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18.* Springer, 2015, pp. 234–241.

33. P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: a big comparison for nas," *arXiv preprint arXiv:1912.06059*, 2019.

34. S. Liu, Z. Chen, W. Li, J. Zhu, J. Wang, W. Zhang, and Z. Gan, "Efficient universal shuffle attack for visual object tracking," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 2739–2743.

35. Z. Huang, P. Zhou, S. Yan, and L. Lin, "Scalelong: Towards more stable training of diffusion model via scaling network long skip connection," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

36. L. Rice, A. Bair, H. Zhang, and J. Z. Kolter, "Robustness between the worst and average case," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 840–27 851, 2021.

37. Y. Wang, Y. Ba, H. C. Zhang, H. Zhang, A. Kadambi, S. Soatto, A. Wong, and C.-J. Hsieh, "Evaluating worst case adversarial weather perturbations robustness," in *NeurIPS ML Safety Workshop*, 2022.

38. C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

39. M. Poggi, F. Tosi, and S. Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," in *2018 International conference on 3d vision (3DV).* IEEE, 2018, pp. 324–333.

40. M. Ramamonjisoa, Y. Du, and V. Lepetit, "Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 648–14 657.

41. A. Wong and S. Soatto, "Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5644–5653.

42. S. Zhang, J. Zhang, and D. Tao, "Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating imu motion dynamics," in *European Conference on Computer Vision.* Springer, 2022, pp. 143–160.

43. W. Han, J. Yin, X. Jin, X. Dai, and J. Shen, "Brnet: Exploring comprehensive features for monocular depth estimation," in *European Conference on Computer Vision.* Springer, 2022, pp. 586–602.

44. Z. Zhou and Q. Dong, "Self-distilled feature aggregation for self-supervised monocular depth estimation," in *European Conference on Computer Vision.* Springer, 2022, pp. 709–726.

45. D. Garg, Y. Wang, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "Wasserstein distances for stereo disparity estimation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 517–22 529, 2020.

46. X. Chen, R. Zhang, J. Jiang, Y. Wang, G. Li, and T. H. Li, "Self-supervised monocular depth estimation: Solving the edge-fattening problem," in *Proceedings*

of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5776–5786.

47. J. Ma, X. Lei, N. Liu, X. Zhao, and S. Pu, "Towards comprehensive representation enhancement in semantics-guided self-supervised monocular depth estimation," in *European Conference on Computer Vision.*   Springer, 2022, pp. 304–321.

48. S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 116–13 125.

49. P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 2624–2632.

50. H. Jung, E. Park, and S. Yoo, "Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 642–12 652.

51. R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 560–15 569.

52. K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, and J. Yang, "Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 055–16 064.

53. M. Vankadari, S. Golodetz, S. Garg, S. Shin, A. Markham, and N. Trigoni, "When the sun goes down: Repairing photometric losses for all-day depth estimation," in *Conference on Robot Learning.*   PMLR, 2023, pp. 1992–2003.

54. J. Spencer, R. Bowden, and S. Hadfield, "Defeat-net: General monocular depth via simultaneous unsupervised representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 402–14 413.

55. A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2800–2810.

56. C. Zhao, Y. Tang, and Q. Sun, "Unsupervised monocular depth estimation in highly complex environments," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 5, pp. 1237–1246, 2022.

57. N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp).*   Ieee, 2017, pp. 39–57.

58. H. Zhang and J. Wang, "Towards adversarially robust object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 421–430.

59. X. Chen, C. Xie, M. Tan, L. Zhang, C.-J. Hsieh, and B. Gong, "Robust and accurate object detection via adversarial learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 622–16 631.

60. X. Xu, H. Zhao, and J. Jia, "Dynamic divide-and-conquer adversarial training for robust semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7486–7495.

61. W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.

62. Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," *Advances in neural information processing systems*, vol. 32, 2019.

63. J.-B. Alayrac, J. Uesato, P.-S. Huang, A. Fawzi, R. Stanforth, and P. Kohli, "Are labels required for improving adversarial robustness?" *Advances in Neural Information Processing Systems*, vol. 32, 2019.

64. C.-H. Ho and N. Nvasconcelos, "Contrastive learning with adversarial examples," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 081–17 093, 2020.

65. M. Kim, J. Tack, and S. J. Hwang, "Adversarial self-supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2983–2994, 2020.

66. S. Liu, Z. Chen, Y. Liu, Y. Wang, D. Yang, Z. Zhao, Z. Zhou, X. Yi, W. Li, W. Zhang *et al.*, "Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 436–23 446.

67. L. Kong, Y. Niu, S. Xie, H. Hu, L. X. Ng, B. R. Cottereau, D. Zhao, L. Zhang, H. Wang, W. T. Ooi, R. Zhu, Z. Song, L. Liu, T. Zhang, J. Yu, M. Jing, P. Li, X. Qi, C. Jin, Y. Chen, J. Hou, J. Zhang, Z. Kan, Q. Ling, L. Peng, M. Li, D. Xu, C. Yang, Y. Yao, G. Wu, J. Kuai, X. Liu, J. Jiang, J. Huang, B. Li, J. Chen, S. Zhang, S. Ao, Z. Li, R. Chen, H. Luo, F. Zhao, and J. Yu, "The robodepth challenge: Methods and advancements towards robust depth estimation," 2023.