Towards Robust Image Stitching: An Adaptive Resistance Learning against Compatible Attacks

Zhiying Jiang[1], **Xingyuan Li**[1], **Jinyuan Liu**[2], **Xin Fan**[1], **Risheng Liu**[1*]

[1] Shool of Software Engineering, Dalian University of Technology
[2] School of Mechanical Engineering, Dalian University of Technology
{zyjiang0630, icinesi.li}@gmail.com, atlan

Abstract

Image stitching seamlessly integrates images captured from varying perspectives into a single wide field-of-view image. Such integration not only broadens the captured scene but also augments holistic perception in computer vision applications. Given a pair of captured images, subtle perturbations and distortions which go unnoticed by the human visual system tend to attack the correspondence matching, impairing the performance of image stitching algorithms. In light of this challenge, this paper presents the first attempt to improve the robustness of image stitching against adversarial attacks. Specifically, we introduce a stitching-oriented attack (SoA), tailored to amplify the alignment loss within overlapping regions, thereby targeting the feature matching procedure. To establish an attack resistant model, we delve into the robustness of stitching architecture and develop an adaptive adversarial training (AAT) to balance attack resistance with stitching precision. In this way, we relieve the gap between the routine adversarial training and benign models, ensuring resilience without quality compromise. Comprehensive evaluation across real-world and synthetic datasets validate the deterioration of SoA on stitching performance. Furthermore, AAT emerges as a more robust solution against adversarial perturbations, delivering superior stitching results. Code is available at: https://github.com/Jzy2017/TRIS.

Introduction

Image stitching aims to relieve the limitations of camera field-of-view (FOV) by integrating images from different viewpoints to reconstruct wide FOV scenes. The primary challenge in this task is managing planar transformations in multi-view scenes, especially when aligning a target image with its reference counterpart on a shared plane, achieved through feature matching in overlapping regions.

Early approaches predominantly rely on feature detection, utilizing descriptors such as SIFT (Lowe 2004), SURF (Bay, Tuytelaars, and Van Gool 2006), and ORB (Rublee et al. 2011), and established correspondence matching grounded in neighborhood measurements. However, they fall short in cases of the complicated scenarios and often mismatch the detected points. In contrast, the advent of deep learning have ushered in notable advancements in the image stitching (Nie et al. 2021; Jiang et al. 2022b). These methods harness the rich feature representations offered by deep learning, enabling more precise multi-view alignments. As a result, they yield superior results characterized by minimal ghosting and more trustworthy reconstructions.

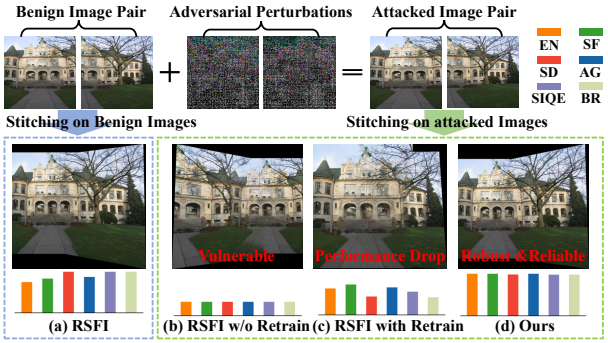

Figure 1: Illustration of our motivation. (a) presents the reference performance of prolific RSFI (Nie et al. 2021) on benign images. (b) reveals the vulnerability of RSFI under adversarial perturbations. Upon routine adversarial training, the robustness of RSFI improves, yet there remains a notable performance decline, as depicted in (c). In contrast, the proposed method in (d) not only exhibits resilience against perturbations but also delivers a performance surpassing that observed in the benign scenarios.

While deep learning based image stitching works have achieved significant advancements, their robustness against adversarial attacks remains a concern. Despite their imperceptibility to the human vision, subtle perturbations can drastically modify the predicted results (Liu et al. 2023d). Given the complexity and diversity of real-world scenes, the imperceptible perturbations can easily blend into the detailed content. However, at the feature level, these perturbations cause marked deviations, severely compromising the accuracy of feature matching in the stitching process. Thus, addressing the vulnerability of image stitching methods to adversarial attacks becomes a critical challenge in ensuring the reliability of the stitched results.

This paper presents the first endeavor to enhance the robustness of learning based image stitching against adversarial attacks. Specifically, we develop a stitching-oriented at-

*Corresponding Author.

tack perturbation (SoA) tailored for the vital alignment of overlapping content. This perturbation is grounded on an extensive investigation of existing attack strategies, which not only undermines the accuracy of stitching models but also demonstrates compatibility with various prominent attacks. To devise a robust stitching model, we conducted a comprehensive assessment of the adversarial resistance on conventional structures and developed an adaptive architecture search to strike a harmonious equilibrium between attack resistance and stitching accuracy. Through this, the optimal attack resistant architecture is discerned using the adaptive adversarial training (AAT), mitigating the performance disparities observed in routine adversarial training methods compared to benign models. With the above attack and adaptive training strategies, we construct a robust and flexible stitching framework for challenging and vulnerable applications. Main contributions can be summarized as follows:

- This paper advances the robustness of image stitching against adversarial challenges, providing a targeted attack and flexible adversarial training strategy, thereby paving the way for attack resistant image stitching across diverse domains.
- Given the deterioration of alignment accuracy, we developed a stitch-oriented attack perturbation compatible with conventional attacks, significantly impairing stitching performance.
- We explore the resistance of foundational structure and propose an adaptive adversarial training to determine the robust and effective model, alleviating the performance compromise against the benign models.
- Extensive experiments demonstrate that the proposed method achieves a remarkable promotion in both adversarial robustness and stitching performance, outperforming routine adversarial training and benign models.

Related Work

Image Stitching

Existing image stitching methods are mainly developed on feature detection, they employ feature descriptors (Lowe 2004; Bay, Tuytelaars, and Van Gool 2006) to extract feature points and utilize nearest-neighbor, such as RANSAC (Fischler and Bolles 1981) to match them. By finding the point pairs with similar features in different images, they establish the transformation required for alignment.

Nevertheless, the global consistent transformation often gives rise to the ghosting artifacts. (Gao, Kim, and Brown 2011) treated the foreground and background separately to estimate dual homography for the whole images. Smoothly varying affine (Lin et al. 2011) was proposed to address the variable parallax. As-projective-as-possible (Zaragoza et al. 2013) balanced local nonprojective deviations while adhering to a global projective constraint. Building upon the mesh framework, (Zhang et al. 2016) enhanced alignment and regularity constraints, enabling support for large baseline and non-planar structures. (Chen and Chuang 2016) incorporated a global similarity prior (GSP) to relieve the local distortion. Furthermore, superpixel based optimal homography was developed (Lee and Sim 2020) to conduct the stitching more adaptively.

More recently, deep learning based image stitching methods have gradually become mainstream. (Nie et al. 2020) introduced a content revision module to address the ghosting and seam artifacts. After that, they proposed ablation constraint to reconstruct the broad scene from feature to pixel (Nie et al. 2021). (Song et al. 2022) presented a weakly supervised learning method for fisheye panorama generation. Benefiting from the complementary of multi-modality data, (Jiang et al. 2022b, 2023) proposed infrared and visible image stitching. While the emergence of deep learning has brought about significant improvements in stitching, it also revealed vulnerabilities to adversarial attacks, compromising the accuracy and robustness of the stitched outcomes.

Adversarial Attacks

For deep neural networks, adversarial attacks refer to deceiving the models with perturbed input samples. These perturbations are usually imperceptible to the human vision but compromise the model inference significantly. Recently, various attack methods have been developed. Optimization-based attack method was designed in (Szegedy et al. 2013), which minimizes the perturbation magnitude while altering the classification results. (Goodfellow, Shlens, and Szegedy 2014) proposed a Fast Gradient Sign Method (FGSM), utilizing model gradients to generate attack examples. However, due to its linear approximation, the examples generated by FGSM are susceptible to detection. To address this, (Kurakin, Goodfellow, and Bengio 2018) presented a Basic Iterative Method (BIM) with multi-step scheme for attack generation. (Madry et al. 2017) also introduced an iterative scheme, essentially known as Projected Gradient Descent (PGD). In their approach, gradients of input samples with respect to the loss function are computed, and small perturbations are added in the direction of their gradients.

Some studies on the robustness of the deep learning algorithms have been conducted. In high-level vision tasks, (Xie et al. 2017) attempted to optimize the loss function over a set of targets to improve the adversarial robustness on segmentation and object detection. Based on generative adversarial networks, (Xiao et al. 2018) enabled the examples share the same distribution with original images, tracing high perceptual quality in defenses. (Joshi et al. 2019) developed an optimization based framework to generate semantically valid adversarial examples using parametric generative transformations to enhance the robustness of classifier. As for low-level tasks, (Yin et al. 2018) investigated the robust super resolution for different downstream tasks. (Choi et al. 2019) examined the robustness of super resolution methods against adversarial attacks. In image dehazing, (Gao et al. 2021) developed a potentially adversarial haze with high realisticity and misleading. (Yu et al. 2022) provides a comprehensive analysis on the robustness of existing deraining models. As for image stitching, subtle perturbations decrease the alignment of multi-viewpoints dramatically. However, there is limited works on the vulnerability of deep image stitching methods against adversarial attacks.

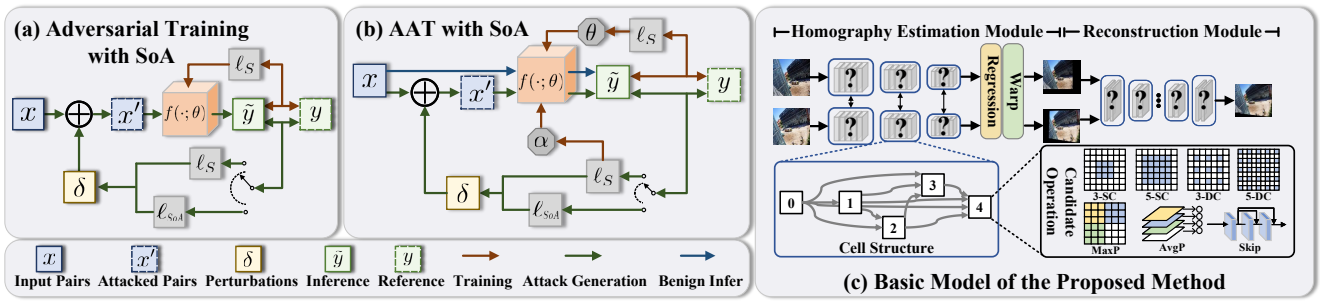

Figure 2: Illustration of the stitching-oriented attacks (SoA) based routine adversarial training (a) and the proposed adaptive adversarial training (AAT) (b).The basic architecture we employed is shown in (c).

The Proposed Method

Attacks on Image Stitching

Adversarial attacks aim to deteriorate the generated wide FOV image by adding subtle, imperceptible perturbations to the input image pairs. To thoroughly challenge network capabilities and to investigate robust and compatible stitching models, we develop a stitch-oriented attack method (SoA), which focuses on the degradation of alignment across different viewpoints and is instrumental in evaluating the attack resistance of stitching models.

We consider the stitching model $f(\cdot;\theta)$ parameterized by θ. Given the input image pair (x_1, x_2), perturbations δ and degradation metric M, the objective of adversarial attacks lies in maximizing the deviation of the generation from the attack-free counterpart, which can be expressed as:

$$\delta = \underset{\delta, ||\delta||_p \leq \epsilon}{\arg\max} M(f((x_1, x_2); \theta), f((x_1+\delta, x_2+\delta); \theta)). \quad (1)$$

In order to solve this maximization problem under the l_p-bound constraint, we adopt the Projected Gradient Descent (PGD) (Madry et al. 2017) and calculate the perturbations in an iterative manner, expressed as:

$$g = \nabla_{(x_1', x_2')} M(f((x_1, x_2); \theta), f((x_1', x_2'); \theta)), \quad (2)$$

$$(x_1', x_2') = \text{clip}_{[-\epsilon, \epsilon]}((x_1', x_2') + \beta \cdot sign(g)), \quad (3)$$

where $(x_1', x_2') \leftarrow (x_1 + \delta, x_2 + \delta)$, ∇denotes the gradient operation. β means the step length of each iteration. clip$(\cdot)$ guarantees the perturbations are within $[-\epsilon, \epsilon]$, where ϵ represents the maximum perturbation allowed for each pixel.

For image stitching task, the alignment across different viewpoints which primarily relies on image features, is paramount for achieving optimal stitching results. Although the introduced perturbations may remain imperceptible visually, their impact at the feature level is profound. Drawing from this analysis, we develop the corresponding metric M from the perspective of alignment, incorporating both a supervised loss based on homography and an unsupervised loss grounded in shared region consistency.

- *Homography based supervised loss*: homography estimation serves as a conventional procedure for aligning multi-view scenes. We leverage the ℓ_2-norm to quantify the disparity between the homography obtained from the adversarially attacked model and the ground truth, expressed as:

$$\ell_H = ||H - H'||_2, \quad (4)$$

where H is the ground truth homography and H' represents the inferred homography from the adversarially attacked model.

- *Shared region based unsupervised loss*: the shared content across multi-view scenes provides comprehensive information for assessing alignment accuracy. We only need to focus on the consistency of the shared content between the transformed multiple scenes, which can be formulated as:

$$\ell_S = ||\mathcal{H}(E) \odot x_1 - \mathcal{H}(x_2)||_2, \quad (5)$$

where $\mathcal{H}(\cdot)$ warps one image to align with the other using estimated homography, $\odot$ is the pixel-wise multiplication and E is and all-one matrix with identical size with x_1. In practice, for the corresponding measure between the aligned image pairs with/without attacks, we also employ the shared content and the corresponding metric can be represented as:

$$\ell_{AS} = ||\mathcal{W}(E, I) \odot \mathcal{W}(x_2, H) - \mathcal{W}(E, I) \odot \mathcal{W}(x_2, H')||_2, \quad (6)$$

where I and H are the identity matrix and the estimated homography matrix, respectively. And $\mathcal{W}(\cdot, \cdot)$ donates the operation of warping an image using a 3×3 transformation matrix with the stitching domain set to the latest large FOV.

Accordingly, the corresponding metric ℓ_{SoA} for the alignment performance between different viewpoints with/without attacks can be defined as:

$$\ell_{SoA} = \ell_H + \ell_{AS}. \quad (7)$$

The generation of the proposed stitch-oriented attack (SoA) perturbations can be summarized in Alg. 1, where we initialize $(x_1', x_2') \leftarrow (x_1, x_2)$ and the gradient is calculated on ℓ_S rather than ℓ_{SoA} for the first iteration.

Adaptive Adversarial Training

Extensive efforts are currently being devoted to developing specialized learning strategies. Recognizing the intricacies of perturbation generation, many studies have embraced adversarial training to bolster robustness against such attacks.

Algorithm 1: SoA based Perturbation Generation

Require: image pair (x_1, x_2), perturbation bound ϵ, step size β, network weights θ

$(x_1', x_2') \leftarrow (x_1, x_2)$
for $iter$ = 1 to m **do**
  if $iter$ = 1 **then**
    $g \leftarrow \nabla_{(x_1', x_2')} \ell_S(\theta, (x_1', x_2'))$
  else
    $g \leftarrow \nabla_{(x_1', x_2')} \ell_{SoA}(\theta, (x_1', x_2'), (x_1, x_2))$
  end if
  $(x_1', x_2') \leftarrow (x_1', x_2') + \beta \cdot sign(g)$
  $(x_1', x_2') \leftarrow \{(x_1', x_2') \mid \|(x_1', x_2') - (x_1, x_2)\|_\infty \leq \epsilon\}$
end for

Although integrating attacked data into the training process strengthens resistance, it often compromises task-specific performance (Liu et al. 2022b; Jiang et al. 2022a). In order to mitigate the performance degradation exhibited by the models following adversarial training, and to realize an image stitching model with powerful attack-resistance and effective stitching performance, we develop an Adaptive Adversarial Training (AAT) strategy from architecture perspective.

Specifically, the proposed strategy is developed on differentiable architecture search (DARTS) (Liu, Simonyan, and Yang 2018; Liu et al. 2022a, 2021, 2023a). The differentiable search strategy relaxes the discrete search space into a continuous one by introducing the continuous relaxation α, and the whole optimization objective for search can be formulated as:

$$\begin{aligned} \min_{\alpha} \quad & \mathcal{L}_{\text{val}}(\alpha; \theta^*) + \lambda \mathcal{L}_{\text{val}}^{atk}(\alpha; \theta^*), \\ \text{s.t.} \quad & \theta^* = \arg\min_{\theta} \mathcal{L}_{\text{train}}(\theta; \alpha), \end{aligned} \tag{8}$$

where $\mathcal{L}_{\text{train}}$, $\mathcal{L}_{\text{val}}$, and $\mathcal{L}_{\text{val}}^{atk}$ denote the training loss, normal validation and attacked validation loss guided with SoA perturbations. The optimization of the aforementioned objective can be decoupled in an iterative manner, focusing separately on the robust architecture training for α and the standard optimal parameter learning for θ. Initially, we populate the attacked data with its original counterpart. For the optimization of α, we employ mixed data comprising both normal and attacked samples for standard adversarial training, thereby facilitating robust architecture construction. To balance performance with robustness and prevent search oscillation, we exclusively use the normal data for the weight parameter optimization in the lower objective. The detailed procedure of AAT is delineated in Alg. 2.

Robust Stitching Model

Our base network is built upon PWCnet (Sun et al. 2018). This network features a three-scale pyramid designed for effective feature encoding and uses an iterative regression mechanism to achieve correspondence matching in a coarse-to-fine manner. In order to enhance the attack-resistance in image stitching, we investigate the robust and flexible structure for feature representation. We predefine a communal cell based on a five-node acyclic graph. Within this

Algorithm 2: SoA based Adaptive Adversarial Training

Require: dataset D, training epoch T, learning rate γ_1, γ_2, architecture parameters α, network weights θ

for epoch = 1, ..., T **do**
  for minibatch B $\sim$ D **do**
    % Adversarial Examples Generation with SoA
    for $iter$ = 1 to m **do**
      Compute project gradient descent g
      $g \leftarrow \mathbb{E}_{(x_1, x_2) \in B}[\nabla_{(x_1', x_2')}(\ell_S(\theta, (x_1', x_2')) \coprod \ell_{SoA}(\theta, (x_1', x_2'), (x_1, x_2)))]$
      Update adversarial examples (x_1', x_2') with g
      Project $(x_1', x_2') - (x_1, x_2)$ to ℓ_p-ball with ϵ
    end for
    % Architecture Search
    Compute $\ell_S(\cdot)$ on (x_1', x_2') with θ
    $\alpha \leftarrow \alpha - \gamma_1 \mathbb{E}_{(x_1, x_2) \in B}[\nabla_\alpha \ell_S(\theta, (x_1', x_2'))]$
    % Weights Learning
    Compute $\ell_S(\cdot)$ on (x_1, x_2) with α
    $\theta \leftarrow \theta - \gamma_2 \mathbb{E}_{(x_1, x_2) \in B}[\nabla_\theta \ell_S(\theta, (x_1, x_2))]$
  end for
end for

graph, each node is associated with a set of mixed operations. Specifically, these operations represent a weighted average of our defined set of candidate operations (Huang et al. 2022; Liu et al. 2020). Furthermore, every subsequent node connects with all its predecessors. In this paper, we introduce seven candidate operators, including Skip, Average Pooling (AvgP), Max Pooling (MaxP), 3×3 SepConv (3-SC), 5×5 SepConv (5-SC), 3×3 Dilated Conv (3-DC), 5×5 Dilated Conv (5-DC). The employed three-level pyramid is established on three communal cells and the homography can be obtained using the global correlation regression.

In reconstructing the wide FOV, we employ a U-Net (Ronneberger, Fischer, and Brox 2015; Liu et al. 2023b,c) inspired architecture, which incorporates six communal cells. In this setup, the three conventional down-sampling and three up-sampling encoding stages in the traditional U-Net are supplanted by the communal cells. To boost the robustness of our stitching framework, the communal cell is adapted with the similar candidate operator to the homography estimation and the inter connection is determined through the search learning (Li et al. 2022, 2023). A detailed depiction of the structure can be seen in Fig. 2 (c).

Experiments

Implementation Details

There are two benchmarks available for image stitching, including a synthetic dataset based on MS-COCO (Nie et al. 2020) and a real-world UDIS-D (Nie et al. 2021) collected from various moving videos. For the training of homography estimation module, we employed the synthesized MS-COCO dataset for initial 120 epochs and fine-tuned on the training set of UDIS-D for 20 epochs. The optimizer is Adam (Kingma and Ba 2014) with an initial learning rate of $1e^{-4}$ and the decay rate is 0.96. The reconstruction module is trained on UDIS-D for 30 epoch, adhering to the same

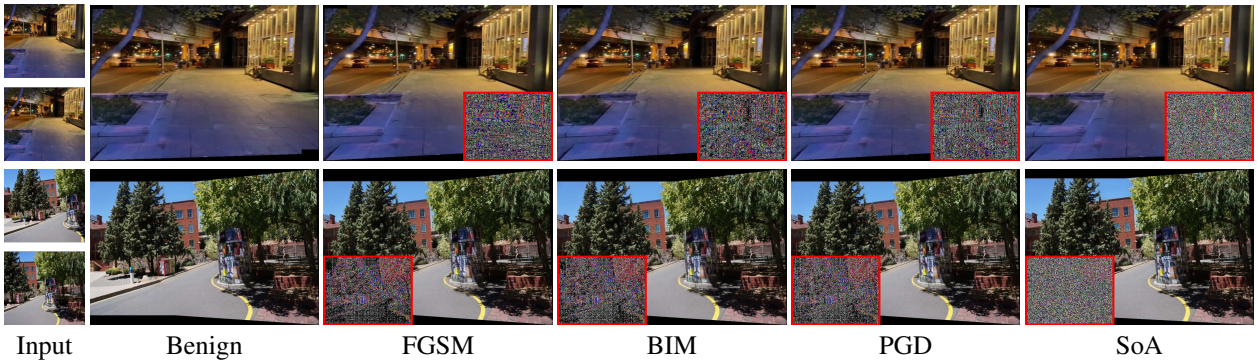

Figure 3: Results of our method under different attacks (i.e., FGSM, BIM, PGD and SoA) and in attack-free (benign) scenarios.

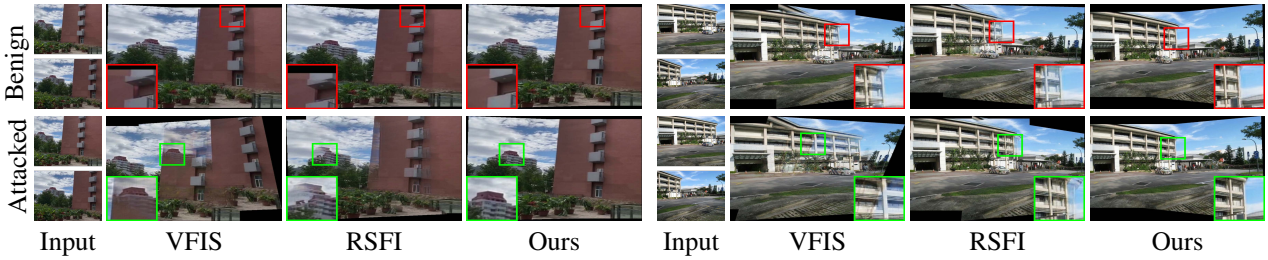

Figure 4: Visual comparisons of different adversarially trained stitching models on benign and attacked images.

Data	Method	Type	EN ↑	SF ↑	SD ↑	AG ↑	SIQE ↑	BR ↑	NIQE ↓	PI ↓
UDIS-D	VFIS	Benign	7.280	17.091	57.228	6.891	41.101	34.621	4.043	3.041
		Attacked	6.928	16.300	53.403	6.192	30.186	19.773	4.731	3.816
	RSFI	Benign	7.298	17.130	56.684	6.879	40.383	34.331	4.028	2.974
		Attacked	7.011	16.511	54.525	6.204	31.788	20.158	4.653	3.692
	Ours	Benign	**7.353**	**17.320**	56.731	7.258	47.675	**34.502**	**3.904**	2.945
		Attacked	7.129	16.847	**56.778**	**7.631**	**47.963**	31.426	4.281	**2.832**
RWCC	VFIS	Benign	7.253	13.929	63.158	5.925	39.108	27.836	3.678	3.743
		Attacked	6.728	12.339	56.945	4.879	21.741	23.129	4.388	4.448
	RSFI	Benign	7.274	13.856	63.027	5.979	39.032	27.365	3.689	3.960
		Attacked	6.817	12.447	56.794	5.001	23.328	23.466	4.253	4.325
	Ours	Benign	**8.327**	**14.276**	64.280	**6.217**	39.490	**28.396**	3.611	3.699
		Attacked	7.923	13.438	**65.932**	5.732	**40.117**	27.357	**3.526**	**3.454**

Table 1: Quantitative comparison between adversarially trained models and the proposed method on benign and attacked data.

hyperparameter configuration. For evaluation, the test set of UDIS-D is adopted, which contains 1106 image pairs. Moreover, we additionally obtained 62 pairs of real-world challenging cases (RWCC) from (Zhang et al. 2020; Lin et al. 2015; Chang, Sato, and Chuang 2014; Gao, Kim, and Brown 2011; Chen and Chuang 2016; Li et al. 2017) as comprehensive validation. For adversarial attack, perturbation intensity ϵ is set as $8/255$, the iteration count is 3, and the step size is $5/255$. Both the training and testing are implemented on Pytorch with an NVIDIA Tesla A40 GPU.

Performance on Image Stitching

Evaluation on the compatibility against different attacks. We evaluated the stitching performance of our method against attacks from FGSM, BIM, PGD, and SoA strategies, as well as in benign (attack-free) scenarios. As illustrated in Fig. 3, the first sample is drawn from the UDIS-D dataset, while the second originates from the RWCC dataset. Notably, the bottom corners of the attacked scenarios provide visual representations of various attack perturbations. Our method clearly exhibits robust performance against SoA attacks and also showcases strong compatibility with FGSM, BIM, and PGD attacks. Moreover, in benign scenarios, our method maintains consistent performance, mitigating the degradation observed in routine adversarial training.

Evaluation of AAT against routine adversarial training. Visual comparisons between the proposed adaptive adversarial training (AAT) with the routine adversarial training are presented in Fig. 4. For these comparisons, we adopt the prolific deep learning based stitching methods VFIS (Nie

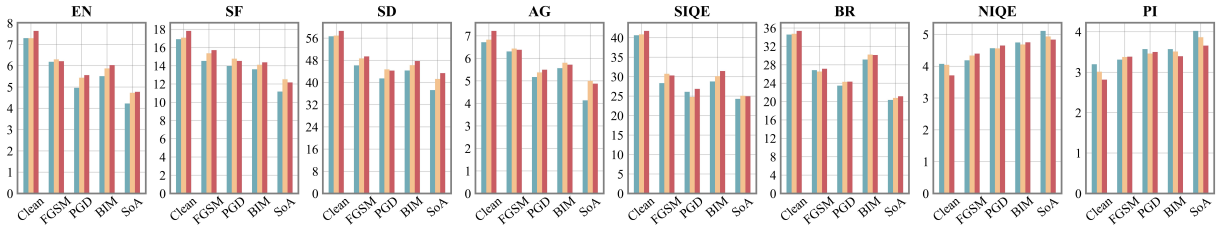

Figure 5: Performance deterioration from different attacks on deep learning models. Blue, yellow and red denote VFIS, RSFI and our baseline model trained with clean data.

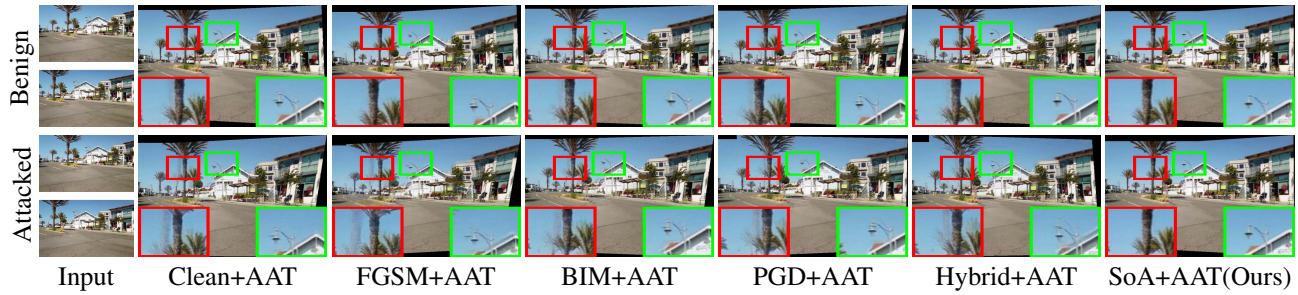

Figure 6: Performance comparison of the proposed adaptive adversarial training (AAT) mechanism incorporated with variant perturbations. Obviously, SoA + AAT strategy exhibits superior performance on both benign and attacked data.

et al. 2020) and RSFI (Nie et al. 2021). Both methods were retrained using the projected gradient descent (Madry et al. 2017) to derive their respective robust models. In the first sample, the results of VFIS and RSFI on benign images present the ghost effect within the red frame, while they illustrate misalignment on the attacked images within the green frame. In the second sample, VFIS exhibits misalignment in both benign and attacked images, and simultaneously introduces ghosting artifacts. In contrast, the proposed method consistently delivers reliable stitching results under both attacked and benign conditions.

Quantitative results on UDIS-D and RWCC are shown in Table. 1, where Entropy (EN) (Zhao et al. 2023), Spatial Frequency (SF) (Eskicioglu and Fisher 1995), Standard Deviation (SD) (Zhao et al. 2022), Average Gradient (AG) (Cui et al. 2015), Stitched Image Quality Evaluator (SIQE) (Madhusudana and Soundararajan 2019), Blind/Referenceless image spatial quality evaluator (BR) (Mittal, Moorthy, and Bovik 2012), Naturalness Image Quality Evaluator (Mittal, Moorthy, and Bovik 2012) (NIQE) and Perceptual Index (PI) are employed as metrics. For the first six metrics, higher values mean better image quality; for the last two, lower values are preferable. After adversarial training, both VFIS and RSFI exhibit decreased performance compared to their results on benign data. Although adversarial training bolsters the robustness of these models, there remains a discernible performance gap compared to the benign counterparts. Our method consistently exhibits minimal disparity in performance between the attacked and benign versions.

Ablation Study

Analysis on the performance degradation. To assess the impact of the proposed SoA perturbations on stitching performance, we illustrate the performance degradation resulting from various attack strategies, including FGSM, PGD, BIM, and SoA, on deep learning based stitching models (i.e., VFIS, RSFI, and our baseline model trained on benign data). For comparison, we also present the performance of these three models on clean images. In Fig. 5, blue, yellow, and red correspond to the VFIS, RSFI, and our baseline models, respectively. Clearly, the SoA perturbations result in a more significant performance degradation across all models when compared to other attack strategies. Based on these observations, we adopted SoA perturbations for adversarial training to bolster the robustness of stitching algorithms.

Analysis on the proposed SoA. The proposed method incorporates SoA perturbations into the adaptive adversarial training (AAT) to facilitate a robust stitching model. To further investigate the AAT mechanism, we combine it with different attack perturbations. A visual comparison is presented in Fig. 6, with 'Clean' denoting a training scenario without any attack perturbations, and 'Hybrid' perturbation emerges from amalgamating FGSM, BIM, and PGD perturbations. Notably, the model trained in Clean + AAT strategy reveals a significant disparity in performance between stitching benign images and their attacked counterparts. Models subjected to adversarial training using FGSM, BIM, or PGD perturbations demonstrate moderate robustness in the face of potent attacks, but this robustness is somewhat elevated when trained with Hybrid perturbations. Remarkably, our model stands out with unparalleled robustness, maintaining high efficiency even with clean images.

Analysis on the proposed AAT. Our AAT employs differentiable architecture search to determine the optimal model, where the updates for architecture parameters are governed

Method	Type	EN ↑	SF ↑	SD ↑	AG ↑	SIQE ↑	BR ↑	NIQE ↓	PI ↓
Bayesian	Benign	7.230	17.094	57.594	7.083	45.023	32.442	4.293	3.090
	Attacked	6.818	16.593	54.451	6.316	29.960	21.421	4.705	3.657
Quasi-Newton	Benign	7.250	17.160	57.061	6.910	44.071	33.636	4.124	3.149
	Attacked	6.996	16.633	54.174	6.353	28.073	22.198	4.648	3.722
Ours	Benign	**7.353**	**17.320**	56.731	7.258	47.675	**34.502**	**3.904**	2.945
	Attacked	7.129	16.847	**56.778**	**7.631**	**47.963**	31.426	4.281	**2.832**

Table 2: Robustness comparison of the proposed AAT mechanism with different optimization solvers.

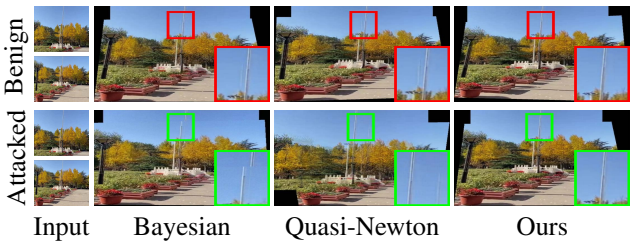

Figure 7: Analysis on the optimization solver of AAT.

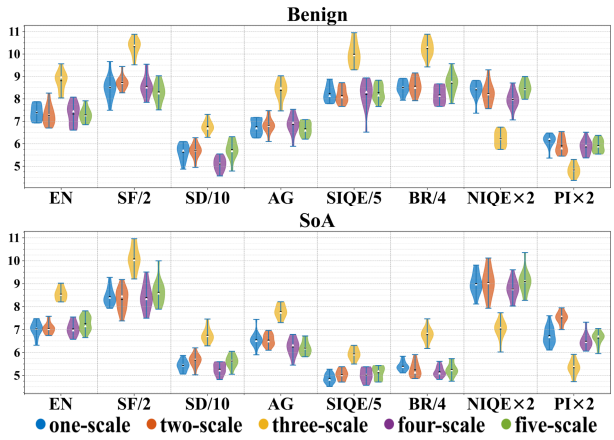

Figure 8: Analysis on the robust baseline model in terms of the constrcution of multi-scale features.

by optimization. To achieve optimal performance, we explore various solver strategies within DARTS. We conduct a comparative analysis of model performance derived from different strategies, including Bayesian optimization, quasi-Newton descent, and first-order derivatives (we used). Visual results are given in Fig. 7. The model trained via first-order derivatives (denoted as Ours) exhibits superior performance against adversarial attacks. This superiority remains consistent across both benign and adversarial data. The quantitative results presented in Table. 2 reinforce this observation, with the model derived from the first-order derivatives outperforming others across all evaluation criteria.

Analysis on the basic multi-scale architecture. Benefiting from the hierarchical representation, multi-scale features play a pivotal role in robust perception. Fig. 8 illustrates the comparison of networks with different scales, in which the first figure illustrates the performance on benign data and

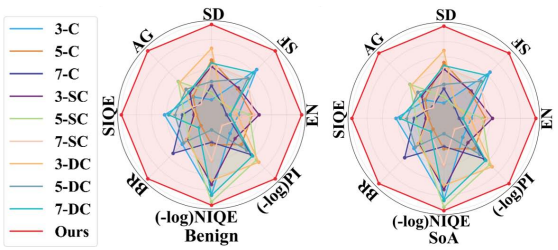

Figure 9: Analysis on the candidate operation setting.

the second depicts the corresponding performance with SoA attacks. It is evident that the pyramidal structure based on three scales offers not only enhanced stitching performance but also increased resistance to attacks. Consequently, we adopt the three-scale setting for our baseline network.

Analysis on candidate operation. The performance of the final model in neural architecture search is significantly influenced by the choice of candidate operations. We carried out an extensive experimental comparison of candidate operations, including substituting candidate operations within each cell with 3×3 Conv (3-C), 5×5 Conv (5-C), 7×7 Conv (7-C), 3×3 SepConv (3-SC), 5×5 SepConv (5-SC), 7×7 SepConv (7-SC), 3×3 Dilated Conv (3-DC), 5×5 Dilated Conv (5-DC), and 7×7 Dilated Conv (7-DC). Using our proposed SoA based AAT mechanism, we trained different architectures and identified the one with the best balance of performance and robustness. The evaluation of these models is shown in Fig. 9. Our model demonstrated superior performance on both benign and adversarial data, validating the appropriateness of our proposed candidate operations.

Conclusion

This paper addressed the vulnerability of deep learning based stitching models against imperceptible perturbations and proposed a robust image stitching method. We introduced a stitching-oriented attack (SoA) tailored for the alignment of shared regions. Furthermore, we developed an adaptive adversarial training (AAT) to facilitate the attack-resistant model. During the adversarial training, robust architectures and efficient parameters are automatically determined in an alternative manner. Extensive comparisons with existing learning based image stitching methods and attack strategies demonstrate that the proposed method possesses a strong resilience to perturbations, alleviating the performance disparity between the benign and attacked scenarios.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (Nos. 2020YFB1313500 and 2022YFA1004101), the National Natural Science Foundation of China (Nos. U22B2052 and 62302078), and China Postdoctoral Science Foundation (No. 2023M730741).

References

Bay, H.; Tuytelaars, T.; and Van Gool, L. 2006. Surf: Speeded up robust features. In *ECCV*, 404–417. Springer.

Chang, C.-H.; Sato, Y.; and Chuang, Y.-Y. 2014. Shape-preserving half-projective warps for image stitching. In *CVPR*, 3254–3261.

Chen, Y.-S.; and Chuang, Y.-Y. 2016. Natural image stitching with the global similarity prior. In *ECCV*, 186–201. Springer.

Choi, J.-H.; Zhang, H.; Kim, J.-H.; Hsieh, C.-J.; and Lee, J.-S. 2019. Evaluating robustness of deep image super-resolution against adversarial attacks. In *ICCV*, 303–311.

Cui, G.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2015. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Opt. Commun.*, 341: 199–209.

Eskicioglu, A. M.; and Fisher, P. S. 1995. Image quality measures and their performance. *IEEE Trans. Commun.*, 43(12): 2959–2965.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6): 381–395.

Gao, J.; Kim, S. J.; and Brown, M. S. 2011. Constructing image panoramas using dual-homography warping. In *CVPR*, 49–56. IEEE.

Gao, R.; Guo, Q.; Juefei-Xu, F.; Yu, H.; and Feng, W. 2021. Advhaze: Adversarial haze attack. *arXiv preprint arXiv:2104.13673*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Huang, Z.; Liu, J.; Fan, X.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *ECCV*, 539–555. Springer.

Jiang, Z.; Li, Z.; Yang, S.; Fan, X.; and Liu, R. 2022a. Target Oriented Perceptual Adversarial Fusion Network for Underwater Image Enhancement. *IEEE Trans. Circuits Syst. Video Technol.*, 32(10): 6584–6598.

Jiang, Z.; Zhang, Z.; Fan, X.; and Liu, R. 2022b. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *ACM MM*, 3783–3791.

Jiang, Z.; Zhang, Z.; Liu, J.; Fan, X.; and Liu, R. 2023. Multi-Spectral Image Stitching via Spatial Graph Reasoning. In *ACM MM*, 472–480.

Joshi, A.; Mukherjee, A.; Sarkar, S.; and Hegde, C. 2019. Semantic Adversarial Attacks: Parametric Transformations That Fool Deep Classifiers. In *ICCV*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Lee, K.-Y.; and Sim, J.-Y. 2020. Warping residual based image stitching for large parallax. In *CVPR*, 8198–8206.

Li, J.; Liu, J.; Zhou, S.; Zhang, Q.; and Kasabov, N. K. 2022. Learning a coordinated network for detail-refinement multi-exposure image fusion. *IEEE Trans. Circuits Syst. Video Technol.*, 33(2): 713–727.

Li, J.; Liu, J.; Zhou, S.; Zhang, Q.; and Kasabov, N. K. 2023. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Trans. Neural Netw. Learn. Syst.*

Li, J.; Wang, Z.; Lai, S.; Zhai, Y.; and Zhang, M. 2017. Parallax-tolerant image stitching based on robust elastic warping. *IEEE Trans. multimedia*, 20(7): 1672–1687.

Lin, C.-C.; Pankanti, S. U.; Natesan Ramamurthy, K.; and Aravkin, A. Y. 2015. Adaptive as-natural-as-possible image stitching. In *CVPR*, 1155–1163.

Lin, W.-Y.; Liu, S.; Matsushita, Y.; Ng, T.-T.; and Cheong, L.-F. 2011. Smoothly varying affine stitching. In *CVPR*, 345–352. IEEE.

Liu, H.; Simonyan, K.; and Yang, Y. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

Liu, J.; Jiang, Z.; Wu, G.; Liu, R.; and Fan, X. 2023a. A unified image fusion framework with flexible bilevel paradigm integration. *The Visual Comput.*, 39(10): 4869–4886.

Liu, J.; Lin, R.; Wu, G.; Liu, R.; Luo, Z.; and Fan, X. 2023b. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *Int. J. Comput. Vis.*, 1–28.

Liu, J.; Wu, G.; Luan, J.; Jiang, Z.; Liu, R.; and Fan, X. 2023c. HoLoCo: Holistic and local contrastive learning network for multi-exposure image fusion. *Inf. Fusion*, 95: 237–249.

Liu, J.; Wu, Y.; Huang, Z.; Liu, R.; and Fan, X. 2021. Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. *IEEE Signal Process. Lett.*, 28: 1818–1822.

Liu, J.; Wu, Y.; Wu, G.; Liu, R.; and Fan, X. 2022a. Learn to search a lightweight architecture for target-aware infrared and visible image fusion. *IEEE Signal Process. Lett.*, 29: 1614–1618.

Liu, R.; Jiang, Z.; Yang, S.; and Fan, X. 2022b. Twin Adversarial Contrastive Learning for Underwater Image Enhancement and Beyond. *IEEE Trans. Image Process.*, 31: 4922–4936.

Liu, R.; Liu, J.; Jiang, Z.; Fan, X.; and Luo, Z. 2020. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Trans. Image Process.*, 30: 1261–1274.

Liu, Z.; Liu, J.; Zhang, B.; Ma, L.; Fan, X.; and Liu, R. 2023d. PAIF: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *ACM MM*, 3706–3714.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60: 91–110.

Madhusudana, P. C.; and Soundararajan, R. 2019. Subjective and objective quality assessment of stitched images for virtual reality. *IEEE Trans. Image Process.*, 28(11): 5620–5635.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12): 4695–4708.

Nie, L.; Lin, C.; Liao, K.; Liu, M.; and Zhao, Y. 2020. A view-free image stitching network based on global homography. *J. Vis. Commun. Image Represent.*, 73: 102950.

Nie, L.; Lin, C.; Liao, K.; Liu, S.; and Zhao, Y. 2021. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Trans. Image Process.*, 30: 6184–6197.

Rao, Y.-J. 1997. In-fibre Bragg grating sensors. *Meas. Sci. Technol.*, 8(4): 355.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, 2564–2571. Ieee.

Song, D.-Y.; Lee, G.; Lee, H.; Um, G.-M.; and Cho, D. 2022. Weakly-Supervised Stitching Network for Real-World Panoramic Image Generation. In *ECCV*, 54–71. Springer.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 8934–8943.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Xiao, C.; Li, B.; Zhu, J.-Y.; He, W.; Liu, M.; and Song, D. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*.

Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; and Yuille, A. 2017. Adversarial Examples for Semantic Segmentation and Object Detection. In *ICCV*.

Yin, M.; Zhang, Y.; Li, X.; and Wang, S. 2018. When deep fool meets deep prior: Adversarial attack on super-resolution network. In *ACM MM*, 1930–1938.

Yu, Y.; Yang, W.; Tan, Y.-P.; and Kot, A. C. 2022. Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond. In *CVPR*, 6013–6022.

Zaragoza, J.; Chin, T.-J.; Brown, M. S.; and Suter, D. 2013. As-projective-as-possible image stitching with moving DLT. In *CVPR*, 2339–2346.

Zhang, G.; He, Y.; Chen, W.; Jia, J.; and Bao, H. 2016. Multi-viewpoint panorama construction with wide-baseline images. *IEEE Trans. Image Process.*, 25(7): 3099–3111.

Zhang, J.; Wang, C.; Liu, S.; Jia, L.; Ye, N.; Wang, J.; Zhou, J.; and Sun, J. 2020. Content-aware unsupervised deep homography estimation. In *ECCV*, 653–669. Springer.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *CVPR*, 5906–5916.

Zhao, Z.; Zhang, J.; Xu, S.; Lin, Z.; and Pfister, H. 2022. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, 5697–5707.