ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447882

SSTA: SALIENT SPATIALLY TRANSFORMED ATTACK

Renyang Liu[1] *Wei Zhou*[1,*] *Sixing Wu*[1] *Jun Zhao*[2] *Kwok-Yan Lam*[2]

[1] Yunnan University
[2] Nanyang Technological University

ABSTRACT

Extensive studies have demonstrated that deep neural networks (DNNs) are vulnerable to adversarial examples (AEs), which brings a huge security risk to the application of DNNs, especially for the AI models developed in the real world. To impede the process of fully exploiting the vulnerabilities of existing DNNs and further improving their robustness in the face of such malicious inputs, many attack methods have been proposed to build AEs. Despite the significant progress that has been made recently, existing attack methods still suffer from the unsatisfactory performance of escaping from being detected by naked human eyes due to the formulation of AE heavily relying on a noise-adding manner. Such mentioned challenges will significantly increase the risk of exposure and result in an attack to be failed. Therefore, in this paper, we propose the Salient Spatially Transformed Attack (SSTA), a novel framework to craft imperceptible AEs, which enhance the stealthiness of AEs by estimating a smooth spatial transform metric on a most critical area to generate AEs instead of adding external noise to the whole image. Compared to SOTA baselines, extensive experiments indicated that SSTA could effectively improve the imperceptibility of the AEs while maintaining a 100% attack success rate.

Index Terms— Adversarial Attack, Imperceptible Adversarial Examples, Spatial Transformed Attack.

1. INTRODUCTION

Deep neural networks (DNNs) are susceptible to AEs, which are crafted by subtly perturbing a clean input [1, 2], especially for computer vision (CV) tasks, like image recognition. The critical point to carry out adversarial attacks on CV models is how to generate AEs with attack success rate and high imperceptibility. Various methods have been proposed to build AEs; among them, most such attacks are crafting AEs in optimizing noise and adding noise manner.

Although most existing attacks can obtain a high success rate, they are not ideal in terms of imperceptibility and similarity since the added perturbations are not harmonious with the clean image [3, 4]. To address these issues, some methods try to generate AEs in a non-noise addition way, such as the spatial transform-based attack, which crafts AEs by changing the specific pixel's position [5, 6]. Even though these methods ensure the adversarial perturbations are more harmonious with the clean counterparts, the imperceptibility is still weak because they disturb the entire image. In most cases, people can easily distinguish the AEs generated by these methods through the naked eyes.

To improve the concealment of AEs, we formulate the issue of synthesizing AEs beyond additive perturbations and propose a novel non-addition attack method called **SSTA**. More specifically, SSTA uses spatial transformation techniques [7] based on the salient region of the image to generate AEs, rather than directly adding well-designed noise to the benign image. The spatial transform technique can calculate a smooth flow field $\boldsymbol{f}$ for each pixel's new locations to formulate an eligible AE. To further improve the concealment and image quality, we constraint the obtained flow field $\boldsymbol{f}$ by limiting it with a small dynamic flow budget ξ.

Extensive experiments on ImageNet datasets indicate that the proposed SSTA can make AEs more inconspicuous while maintaining high attack performance. Besides, evaluation results on many metrics involve similarity and image quality showing that our AEs are more similar to their benign counterparts and preserved the vivid details. The main contributions could be summarized as follows:

- We formulate the imperceptible AE by applying spatial transform operations in the salient region, which are extracted by object detection method, rather than in a noise-adding manner.
- To balance the attack performance and the concealment of the generated AEs, we propose a dynamic strategy to update the extracted critical region and flow budget ξ associated with the number of optimizations increases.
- Comparing with the state-of-the-art imperceptible attacks, experimental results on various victim models show our method's superiority in synthesizing AEs with the attack ability, invisibility, and image quality and guarantee the AEs' similarity to the original image.

The rest of this paper is organized as follows. In Sec. 2, we provide the details of the proposed SSTA framework. The experiments are presented in Sec. 3, with the conclusion drawn in Sec. 4.

*Corresponding author. Email: zwei@ynu.edu.cn

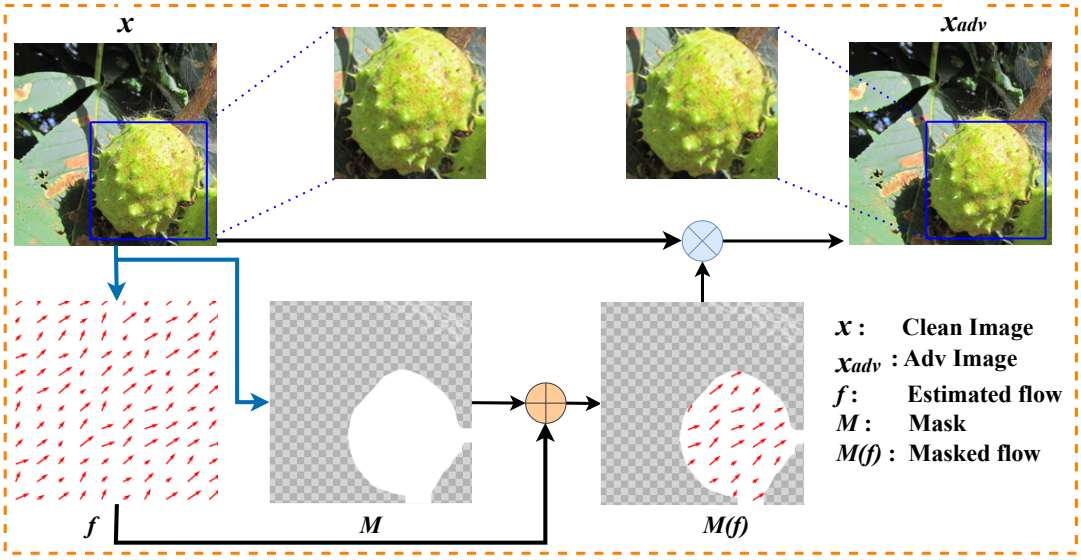

Fig. 1. Overview of SSTA, where $\oplus$ represents applying Mask M, and $\otimes$ represents the spatial transformation operation.

2. METHODOLOGY

2.1. Problem Definition

Given a well-trained DNN classifier $\boldsymbol{C}$ and an input $\boldsymbol{x}$ with its corresponding label y, we have $\boldsymbol{C}(\boldsymbol{x}) = y$. The AE $\boldsymbol{x}_{adv}$ is a neighbor of $\boldsymbol{x}$ and satisfies that $\boldsymbol{C}(\boldsymbol{x}_{adv}) \neq y$ and $\|\boldsymbol{x}_{adv} - \boldsymbol{x}\|_p \leq \epsilon$, where the $\boldsymbol{L}_p$-norm is used as the metric function and ϵ is usually a small noise budget. With this definition, the problem of finding an AE becomes a constrained optimization problem:

$$\boldsymbol{x}_{adv} = \underset{\|\boldsymbol{x}_{adv}-\boldsymbol{x}\|_p \leq \epsilon}{argmax}\, \mathcal{L}\, (\boldsymbol{C}(\boldsymbol{x}_{adv}) \neq y), \tag{1}$$

where $\mathcal{L}$ stands for a loss function that measures the confidence of the model outputs.

Previous works craft an AE $\boldsymbol{x}_{adv}$ by adding $\boldsymbol{L}_p$-norm constrained noise δ to the clean image $\boldsymbol{x}$ as

$$\boldsymbol{x}_{adv} = \boldsymbol{x} + \delta,\ s.t.\ \|\delta\|_p \leq \epsilon. \tag{2}$$

Unlike this, in this paper, we combine the salient region extraction and the spatial transform to build the imperceptible AE $\boldsymbol{x}_{adv}$. As illustrated in Fig. 1, the proposed salient spatially transformed attack framework can be divided into two stages: the first stage is to obtain a salient region mask $\boldsymbol{M}(\cdot)$; the other one is to calculate the flow field $\boldsymbol{f}$. Subsequently, we can formulate the AE $\boldsymbol{x}_{adv}$ by applying the calculated flow field $\boldsymbol{f}$ to the masked salient area $\boldsymbol{M}(\cdot)$ of clean image.

2.2. Salient Region Extraction

In this paper, we use the salient detection method TRACER [8], which can efficiently detect salient objects in images, to extract the critical area mask $\boldsymbol{M}(\cdot)$. In preliminary experiments, we also tried other area extraction methods like LC [9], FT [10], and Grad-CAM [11], but found TRACER [8] is more suitable because it can efficiently detect salient objects in an image and return their corresponding regions, the results are showing in Fig. 2.

Fig. 2. The extracted area by different methods.

Moreover, TRACER can return several regions $r_\tau(\tau = 0, ..., 255)$ with various scales depending on different thresholds τ when extracting salient areas, which will be helpful to the downstream tasks, such as image segmentation and background removal. In our work, we first take the region with a high threshold τ (i.e., $\tau = 250$) as the region mask $\boldsymbol{M}(\cdot)$. Then, in the process of generating AEs, the $\boldsymbol{M}(\cdot)$ will be updated by decreasing the threshold τ to get a larger region.

2.3. Adversarial Example Generation

After computing the mask $\boldsymbol{M}(\cdot)$, we subsequently build AEs by utilizing the spatial transform, which using a flow field matrix $\boldsymbol{f} = [2, h, w]$ to transform the original image x to x_{st} [5]. Specifically, assume the input is x and its transformed counterpart $\boldsymbol{x}_{st}$, for the i-th pixel in $\boldsymbol{x}_{st}$ at the pixel location (u^i_{st}, v^i_{st}), we need to calculate the flow field $\boldsymbol{f}_i = (\Delta u^i, \Delta v^i)$. So, the i-th pixel $\boldsymbol{x}^i$'s location in the transformed image can be indicated as:

$$(u^i, v^i) = (u^i_{st} + \Delta u^i, v^i_{st} + \Delta v^i). \tag{3}$$

To ensure the flow field $\boldsymbol{f}$ is differentiable, the bi-linear interpolation [12] is used to obtain the four neighboring pixels' value surrounding the location $(u_{st}^i + \Delta u^i, v_{st}^i + \Delta v^i)$ for the transformed image $\boldsymbol{x}_{st}$ as:

$$\boldsymbol{x}_{st}^i = \sum_{q \in \boldsymbol{N}(u^i, v^i)} \boldsymbol{x}^q (1 - |u^i - u^q|)(1 - |v^i - v^q|), \quad (4)$$

where $\boldsymbol{N}(u^i, v^i)$ is the neighborhood, that is, the four positions (top-left, top-right, bottom-left, bottom-right) tightly surrounding the target pixel (u^i, v^i). In adversarial attacks, the calculated $\boldsymbol{x}_{st}$ is the final AE $\boldsymbol{x}_{adv}$. Once the f has been computed, we can obtain the $\boldsymbol{x}_{adv}$ by combining $\boldsymbol{M}(\cdot)$ and flow field $\boldsymbol{f}$, which is given by:

$$\boldsymbol{x}_{adv} = \boldsymbol{M}(\sum_{q \in \boldsymbol{N}(u^i, v^i)} \boldsymbol{x}^q (1 - |u^i - u^q|)(1 - |v^i - v^q|)) + (\boldsymbol{x} - \boldsymbol{M}(\boldsymbol{x})), \quad (5)$$

where $\boldsymbol{M}(\boldsymbol{x})$ represents the salient region while the $\boldsymbol{x} - \boldsymbol{M}(\boldsymbol{x})$ indicates the area out of the salient region.

In practice, we regard the problem of calculating flow field $\boldsymbol{f}$ as an optimization task. In this paper, we use the AdamW to optimize flow $\boldsymbol{f}$.

2.4. Objective Functions

Taking the attack success rate and visual invisibility of the generated AEs into account, we divide the objective function into two parts, where one is the adversarial loss and the other is a constraint for the flow field. Unlike other flow field-based attack methods, which constrain the size of the flow field by the flow loss proposed in [5], in our method, we use a dynamically updated flow field budget ξ (a small number, like $1 * 10^{-2}$) to regularize the flow field f. For adversarial attacks, the goal is making the prediction $\boldsymbol{C}(\boldsymbol{x}_{adv}) \neq y$, so we give the objective function as:

$$\mathcal{L}_{adv}(\boldsymbol{x}, y, \boldsymbol{f}) = max[\boldsymbol{C}(\boldsymbol{x}_{adv})_y - \max_{k \neq y} \boldsymbol{C}(\boldsymbol{x}_{adv})_k, C], \quad s.t. \|\boldsymbol{f}\| \leq \xi. \quad (6)$$

where k is the predicted class and C is it's confidence.

3. EXPERIMENTS

3.1. Settings

Dataset: We verify the performance of our method on the development set of ImageNet-Compatible Dataset, a subset of ImageNet-K, which consists of 1,000 images, and we resized the image to 224x224x3 to adopt the victim models.

Models: We use the PyTorch pre-trained model as the victim models, including VGG-19 [13], ResNet-50 [14], DenseNet-121 [15], ViT-16 [16] and Swin_B [17].

Baselines: The baselines include the stAdv [5], Chroma-Shift [6] and AdvDrop [18].

Metrics: We compare our proposed method with baselines concerned with Attack Success Rate (ASR) for attack performance. For image quality, we use the following perceptual metrics referring to image quality, including LPIPS [19], DISTS [20], FID, MSE, UQI [21], SCC [22], PSNR [23], VIPF [24], SSIM and NIQE, to evaluate the difference between the generated AEs and their benign counterparts and the image quality of these AEs.

3.2. Attacking Performance

We investigate the ASR of the proposed method in attacking various image classifiers. The results are shown in Table. 1, we derive that SSTA can obtain the SOTA attack performance by only disturbing the minimal local area, i.e., the salient region, while other attacks need to distort the whole image. This demonstrates the superiority of our method.

Table 1. The ASR of baselines and SSTA.

Methods	VGG-19	ResNet-50	DenseNet-121	VIT-16	Swin_B
stAdv	**100**	**100**	**100**	**100**	**100**
Chroma-Shift	93.69	94.67	95.1	95.09	96.66
AdvDrop	**100**	99.07	**100**	95.97	99.79
SSTA	**100**	99.86	**100**	**100**	**100**

3.3. Image Quality and Similarity

The results of image quality and similarity are shown in Table. 2, which indicated that the proposed method has the lowest LPIPS, DISTS, FID, and MSE (the lower is better) are 0.0038, 0.0091, 16.3876 and 2.1210, respectively, and has the highest UQI, SCC, PSNR, VIPF, SSIM, and NIQE (the higher is better), achieving 0.9998, 0.9890, 49.2397, 0.9487, 0.9987 and 43.9611, respectively, in comparison to the baselines. The results point out that the proposed method is superior to the existing imperceptible attacks.

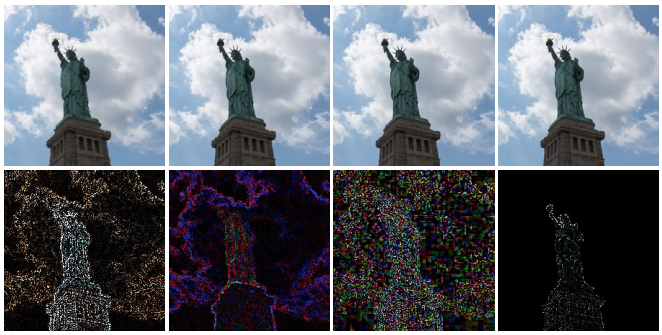

Fig. 3. AEs and their corresponding perturbations.

To visualize the difference between the AEs generated by our method and the baselines, we also draw the adversarial perturbation generated by stAdv, Chroma-Shift, AdvDrop and the proposed method in Fig. 3, the target model

Table 2. Perceptual distances were calculated on fooled examples by stAdv, Chroma-Shift and the proposed SSTA.

Metrics	stAdv	Chroma-Shift	AdvDrop	SSAT
LPIPS ↓	0.1595	0.0135	0.0956	**0.0038**
DISTS ↓	0.1524	0.0165	0.0678	**0.0091**
FID ↓	60.2464	88.8750	46.7813	**16.3876**
MSE ↓	95.7488	23.5399	17.0450	**2.1210**
UQI ↑	0.9925	0.9925	0.9952	**0.9998**
SCC ↑	0.6415	0.9623	0.6894	**0.9890**
PSNR ↑	29.8119	36.5651	36.1464	**49.2397**
VIFP ↑	0.5229	0.7644	0.6474	**0.9487**
SSIM ↑	0.9391	0.9771	0.9688	**0.9987**
NIQE ↑	33.3234	43.5860	39.8657	**43.9611**

is pre-trained ResNet-50. The first row is the AEs and their corresponding noise of stAdv, Chroma-Shift, AdvDrop and our method, respectively. Noted that, for better observation, we magnified the noise by a factor of 30. From Fig. 3, we can clearly observe that the baselines distort the whole image. In contrast, the noise in our generated AEs is milder and focused on the salient region, and more imperceptible to human eyes. These results indicate that the AEs generated by the proposed method have better concealment and can not easily be detected.

3.4. Further Human Perceptual Study

This experiment is for subjective evaluation, i.e., in most cases, whether AEs generated based on SSTA are indistinguishable from their original samples. We argue that AEs generated by SSTA not only satisfy imperceptibility but are also inconspicuous to the human eye. To validate this claim, we compare AEs generated by SSTA with those generated by baselines. In our human perception study, we display the original image and the AEs on the computer screen and give each participant 100 seconds to judge every image. Empirically, 100 seconds is enough to decide and point out any visible distortion for the participants. We used the randomly sampled 50 images for this experiment. The participants are shown 2-5 images, the left is always the clean image and its right side shows adversarial images generated by various methods or the same clean image. Participants will be asked "Are the images on the right the same as the left (the clean one)?" and each participant will provide more than 50 annotations. For each image to be checked, participants can zoom it as large as possible to provide convenience for participants to observe.

A total of 20+ participants were involved in assessing AEs. For the sake of fairness, we put the clean images and adversarial images generated by different methods into the dataset to be checked together. These participants provided more than 1,000 useful annotations. As shown in Fig. 4, the AEs generated by SSTA are generally considered to be the same as the original images. 88.98% of the annotations were considered unmodified, meaning that most participants could not distinguish the AEs generated by SSTA. Conversely, for AEs generated by baseline methods, participants were able to spot distortions more easily, more than 90%, 55% and 30% of the total annotations have been picked out for stAdv, AdvDrop and Chroma-shift, respectively, indicating that the AEs generated by these baseline methods did not affect humans to identify objects in images correctly but very easy to find that they had tampered.

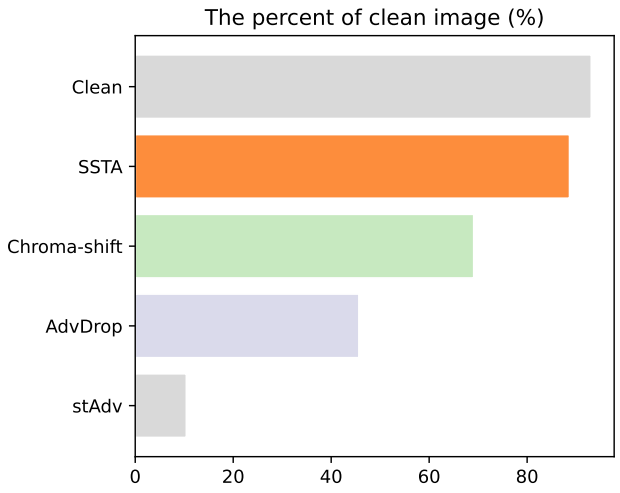

Fig. 4. Human perceptual study results.

4. CONCLUSIONS

In this paper, we present a novel non-noise additional method, called SSTA, which combines performing the spatial transformation in salient regions with the optimal flow field to synthesize AEs. Extensive experiments show that the proposed method is superior to the state-of-the-art methods in terms of prominent concealment and high image quality, and the generated AEs are indistinguishable by the human eyes. Benefitting from generating AEs without noise-adding, the proposed SSTA provides a new efficient way to evaluate the robustness of classifiers and enhance their performance using techniques like fine-tuning or adversarial training. Furthermore, the proposed approach can be used as a reliable tool to build more robust models.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62162067 and 62101480, in part by the Yunnan Province expert workstations under Grant 202305AF150078 and the Applied Basic Research Foundation of Yunnan Province under Grant Nos. 202201AT070156 and 202301AT070194.

6. REFERENCES

[1] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.

[2] Renyang Liu, Jinhong Zhang, Kwok-Yan Lam, Jun Zhao, and Wei Zhou, "SCME: A self-contrastive method for data-free and query-limited model extraction attack," in *ICONIP*, 2023, vol. 14451, pp. 370–382.

[3] Nicholas Carlini and David A. Wagner, "Towards evaluating the robustness of neural networks," in *S&P*, 2017.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[5] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song, "Spatially transformed adversarial examples," in *ICLR*, 2018.

[6] Ayberk Aydin, Deniz Sen, Berat Tuna Karli, Oguz Hanoglu, and Alptekin Temizel, "Imperceptible adversarial examples by spatial chroma-shift," in *ADVM*, 2021, pp. 8–14.

[7] Renyang Liu, Jinhong Zhang, Haoran Li, Jin Zhang, Yuanyu Wang, and Wei Zhou, "AFLOW: developing adversarial examples under extremely noise-limited settings," in *ICICS*, 2023, vol. 14252, pp. 502–518.

[8] Min Seok Lee, WooSeok Shin, and Sung Won Han, "TRACER: extreme attention guided salient object tracing network (student abstract)," in *AAAI*, 2022.

[9] Yun Zhai and Mubarak Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *ACM MM*, 2006, pp. 815–824.

[10] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009, pp. 1597–1604.

[11] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," in *NeurIPS*, 2015, pp. 2017–2025.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 2261–2269.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 9992–10002.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[18] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A. Kai Qin, and Yuan He, "Advdrop: Adversarial attack to dnns by dropping information," in *ICCV*, 2021, pp. 7486–7495.

[19] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[20] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[21] Zhou Wang and Alan C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[22] Jun Li, "Spatial quality evaluation of fusion of different resolution images," *International Archives of Photogrammetry and Remote Sensing*, vol. 33, 09 2000.

[23] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

[24] Hamid R. Sheikh and Alan C. Bovik, "Image information and visual quality," in *ICASSP*, 2004, pp. 709–712.