Stealthiness Assessment of Adversarial Perturbation: From A Visual Perspective

Hangcheng Liu, Yuan Zhou, *Member, IEEE*, Ying Yang, Qingchuan Zhao, Tianwei Zhang, *Member, IEEE*, Tao Xiang, *Senior Member, IEEE*

***Abstract*—Assessing the stealthiness of adversarial perturbations is challenging due to the lack of appropriate evaluation metrics. Existing evaluation metrics, e.g., L_p norms or Image Quality Assessment (IQA), fall short of assessing the pixel-level stealthiness of subtle adversarial perturbations since these metrics are primarily designed for traditional distortions. To bridge this gap, we present the *first* comprehensive study on the subjective and objective assessment of the stealthiness of adversarial perturbations from a visual perspective at a pixel level. Specifically, we propose new subjective assessment criteria for human observers to score adversarial stealthiness in a fine-grained manner. Then, we create a large-scale adversarial example dataset comprising 10586 pairs of clean and adversarial samples encompassing twelve state-of-the-art adversarial attacks. To obtain the subjective scores according to the proposed criterion, we recruit 60 human observers, and each adversarial example is evaluated by at least 15 observers. The mean opinion score of each adversarial example is utilized for labeling. Finally, we develop a three-stage objective scoring model that mimics human scoring habits to predict adversarial perturbation's stealthiness. Experimental results demonstrate that our objective model exhibits superior consistency with the human visual system, surpassing commonly employed metrics like PSNR and SSIM.**

***Index Terms*—Adversarial stealthiness assessment, adversarial attack, classification**

I. Introduction

Adversarial attacks significantly threaten Deep Neural Network (DNN) models. An adversary can carefully craft adversarial perturbations on the target objects (i.e., adversarial examples) to deceive DNN models and induce wrong predictions [1]. A successful adversarial attack generally seeks to achieve two primary objectives: ❶ **high adversarial effectiveness**, ensuring that the adversarial examples exhibit a high likelihood of misleading the model; ❷ **high adversarial stealthiness**, where the perturbation remains imperceptible to humans. This is crucial to ensuring the feasibility of the attack. If an adversarial perturbation is effective but conspicuous, it risks early detection, compromising the attack. For instance, drivers could detect and report suspicious adversarial patches on traffic signs [2], preventing potential accidents. These two goals serve as key motivations for developing effective adversarial attacks. However, existing efforts primarily focus on improving the attack's effectiveness [3]–[6], with comparatively less emphasis on investigating the adversarial stealthiness.

Adversarial attack methods are generally categorized into restricted [4], [7] and unrestricted [8], [9]. Restricted attacks focus on *pixel-level* stealth, imposing L_p-norm constraints on adversarial perturbations (e.g., $\|r\|_\infty < 0.01$ where r denotes the perturbation) to maintain pixel-level similarity between clean and adversarial images. In contrast, unrestricted attacks prioritize semantic-level stealth, allowing significant changes (e.g., adding glasses to a face) as long as the result remains realistic. Considering that (1) distinct methodologies evaluate these two types of stealth, (2) high pixel-level similarity implies semantic plausibility, and (3) restricted attacks are well-studied, *we focus on assessing pixel-level stealthiness for restricted attacks*. In the subsequent sections, "adversarial attacks" and "stealth" refer to restricted attacks and pixel-level stealth, unless otherwise specified.

The absence of dedicated objective metrics for perceptual assessment hinders advancements in adversarial stealthiness. Most related works [4], [7], [10]–[14] use L_p norms, or Full-Reference Image Quality Assessment (FR-IQA) metrics like Peak-Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index (SSIM) [15]. However, as noted in [16], most FR-IQA metrics are tailored for specific applications, aiming to capture relevant distortions, such as blur and blocking for compression, fast fading for wireless transmission, and artifacts for image generation. Adversarial perturbations often have characteristics that differ significantly from distortions typically addressed by the IQA community. Moreover, the commonly used Five-grade Discrete Impairment Scale (FDIS) [17] focuses on perceptible noise, making objective metrics based on FDIS insufficiently sensitive to imperceptible noise (detailed explanation provided later). Consequently, *existing objective metrics easily lead to erroneous conclusions regarding stealthiness due to these factors.* For example, as shown in Fig. 1, Fig. 1(c) demonstrates better stealthiness than Fig. 1(b) as the perturbation in Fig. 1(b) is more annoying. However, PSNR contradicts this by assigning a higher score to Fig. 1(b).

To bridge this gap, introducing a suitable subjective criterion (for constructing a specialized adversarial example dataset) and a dedicated objective metric is essential. To the best

H. Liu and T. Zhang are with the College of Computing and Data Science, Nanyang Technologiwcal University, 639798, Singapore (email: {hangcheng.liu, tianwei.zhang}@ntu.edu.sg).

Y. Zhou is with the School of Computer Science and Technology, Zhejiang Sci-Tech University, Zhejiang 310018, China (email: yuanzhou@zstu.edu.cn).

Y. Yang is with the Institute of High Performance Computing (IHPC) and Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), 138632, Singapore (email: senereone@gmail.com).

Q. Zhao is with the Department of Computer Science, City University of Hong Kong, Kowloon Tong 999077, Hong Kong (email: qizhao@cityu.edu.hk)

T. Xiang is with the College of Computer Science, Chongqing University, Chongqing 400044, China (email: {hwang,txiang}@cqu.edu.cn).

This work is supported by Singapore Ministry of Education (MOE) AcRF Tier 2 MOE-T2EP20121-0006, National Natural Science Foundation of China under Grants U20A20176 and 62072062. (*Corresponding author: Y. Yang.*)

(a) (b) (c)

Fig. 1: The widely used PSNR metric is unreliable for evaluating pixel-level stealthiness across different attacks. (a) Clean example. (b) The adversarial example generated by FGSM ($\epsilon = 0.04$) exhibits annoying perturbation due to the large perturbation budget. (c) The adversarial example generated by GUAP (default weight, spatially transformed) is more visually closer to (a) than (b). MOS scores represent subjective stealthiness scores between the clean and adversarial examples obtained from human observers, while PSNR and 2ASM (**ours**) serve as objective metrics. A higher score of MOS, PSNR, and A2SM indicates better stealthiness. Notably, A2SM aligns with subjective perception by assigning a higher score to (c), while PSNR incorrectly favors (b), highlighting its limitations on assessing adversarial stealthiness.

of our knowledge, our work presents the *FIRST* attempt to investigate both subjective and objective assessments of adversarial stealthiness at the pixel level. We study the stealthiness from the *attacker's perspective*, assuming access to clean samples during evaluation. This access is vital for evaluating adversarial perturbation stealthiness, as attackers rely on clean images to craft adversarial examples. By comparing clean and adversarial images, attackers can emphasize subtle differences and reduce detectable distortions, improving attack effectiveness by avoiding detection.

To conduct this study, we must address two challenges. ❶ *Developing a subjective criterion to capture human perceptions of stealthiness.* A widely used criterion in [17] assesses image quality through full-reference pairwise comparisons (i.e., full-reference). However, it categorizes stealthiness into just two levels, weak and strong, offering limited guidance for improvement. ❷ *Creating a Human Visual System (HVS)-aligned objective method for assessing stealthiness.* Typically, this necessitates a specialized dataset containing clean samples, adversarial examples, and subjective scores. Unfortunately, no such datasets are publicly available. As discussed in Section III-C and Section V-D, existing FR-IQA datasets inadequately support adversarial stealth assessments. Thus, constructing a specialized dataset becomes essential. Moreover, modeling human scoring habits is crucial for developing this method.

We tackle these challenges through three significant contributions. First, we extend the existing subjective assessment criterion [17] for a more fine-grained assessment of the whole stealthiness spectrum. Specifically, we introduce a novel *two-step* subjective assessment. The *first* step involves a rough categorization of stealthiness. Human observers classify adversarial perturbations as **strong** or **weak** based on the detectability of differences between fixed-size image pairs viewed briefly (e.g., 5 seconds). Easily detectable differences indicate weak stealthiness; otherwise, it is strong. The *second* step refines these initial classifications. For weak stealthiness, observers assess how annoying the perturbation is. For strong stealthiness, they evaluate how imperceptible the perturbation appears. Further elucidation on our new subjective criteria is provided in Section III-B.

Second, we construct the first large-scale dataset for assessing adversarial stealthiness. This dataset consists of 10586 pairs of clean and adversarial examples, each labeled with a subjective stealthiness score. To ensure diversity, we generate adversarial examples using five attack types, as shown in Table I. Meanwhile, we recruited 60 human observers for subjective scoring. The final labeled score is the Mean Opinion Score (MOS).

Third, we propose an objective scoring model with three stages: feature extraction, feature fusion, and score prediction. In the first stage, we extract a comprehensive set of global and local low-level features to facilitate accurate scoring. In the second stage, we design two specialized modules to evaluate patch significance in images, considering two key factors: (1) the difference map between the clean and adversarial images determines how noticeable a patch is. Patches with severe perturbations are more conspicuous. (2) The importance of clean images. For instance, perturbations in patches with critical semantics are more annoying than those in non-critical patches. In contrast, previous works have either ignored these factors [18] or considered only one [19]. In the last stage, we predict the score for each patch and calculate their weighted sum as the final score.

We compare our objective method with state-of-the-art FR-IQA methods. The experimental results of Spearman's Rank Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC) show that our model's objective scores on our proposed dataset have the highest correlation with MOS (PLCC=0.984 and SROCC=0.978). Additionally, we evaluate our scoring model on other widely-used IQA datasets, including LIVE [20], CSIQ [21], TID2013 [22], and PIPAL [23], further validating its effectiveness.

Our contributions can be summarized as follows:

- We propose the first subjective criterion to assess the stealthiness of adversarial examples. A rating software is developed for subjective assessment.
- We create the first large-scale adversarial example dataset, including human opinion scores for adversarial stealthiness assessment.
- We design a three-stage model to objectively predict adversarial stealthiness by simulating human scoring habits.
- We evaluate the superiority of our objective scoring model on different datasets.

II. Background and Related Works

A. Adversarial Attacks

In their pioneering work [1], Szegedy et al. discovered that classification models could be fooled by applying "hardly

TABLE I: Summary of adversarial attacks that are used to construct APSD.

Category	Method	Parameter
Gradient-based	FGSM [24]	ϵ=0.005, 0.01, 0.02, 0.04, 0.08
	MIFGSM [3]	α=0.001, 0.002, 0.004, 0.008, 0.01
	PGD [4]	α=0.001, 0.002, 0.004, 0.008, 0.01
	NES [11]	α=0.008, 0.01, 0.02, 0.04, 0.05
Optimization-based	CW [7]	(Confidence, learning rate)=(0, 0.01)
Generative model-based	CDP [25]	Weight=Default weight
	AdvGAN [26]	(Weight, ϵ)=(Default weight, 0.008), (Default weight, 0.016), (Default weight, 0.06)
	GAP [10]	(Weight, Mode)=(Default weight, Universal), (Default weight, Image-dependent)
	GUAP [27]	(Weight, Mode)=(Default weight, Spatial transformed), (Default weight, Adversarial)
Pixel-based	SimBA [28]	(ϵ, Early stop)=(0.2, True), (0.8, False), (0.8, True)
	Pixel [29]	Max iterations=50
Block-based	Square [30]	(L_p, ϵ)=(L_∞, 0.01), (L_∞, 0.03), (L_∞, 0.06)

(a) (b) (c) (d) (e)

Fig. 2: Adversarial examples generated by different methods. From left to right: gradient-based, optimization-based, generative model-based, pixel-based, block-based.

perceptible" perturbations to the original inputs. This phenomenon quickly gained attention, leading to the fast development of adversarial attacks. Most subsequent attacks followed the settings of adversarial attack in [1], attacking target models using imperceptible perturbations constrained by L_p-norms. All these attacks are referred to as restricted attacks, and they have since expanded to tasks beyond classification [14], [31]–[33]. After analyzing representative restricted attacks' mechanisms and perturbation patterns, we categorize them into five types: gradient-based, optimization-based, generative model-based, pixel-based, and block-based (see Fig. 2).

- **Gradient-Based Attack.** This is the most common type of attack and constitutes a significant portion of the overall family of adversarial attacks. Gradient-based attacks produce adversarial examples guided by the gradient directions. For example, the famous FGSM [24] can be described as:

$$x' = x + \epsilon \times \text{sign}(\nabla_x J(\theta_f, x, y)), \tag{1}$$

  where J is an object function, e.g., cross-entropy, y is the ground truth, f is the target model, θ_f is the weight parameters of f, x and x' are clean and adversarial images respectively, and ϵ represents perturbation budgets, that is $\|x - x'\|_\infty \leq \epsilon$. Based on FGSM, several classical multi-step gradient-based attacks [3], [4], [11], [34] are proposed to enhance the adversarial strength. Commonly, gradient-based attacks use L_∞ to ensure stealthiness.
- **Optimization-Based Attack.** As the name suggests, this type of attack treats the generation of adversarial samples as an optimization problem. Szegedy et al. [1] first demonstrated this optimization problem as:

$$\min\ c\|r\|_2 + J(\theta_f, x + r, y)\ s.t.\ x + r \in [0,1]^m, \tag{2}$$

  where m denotes the dimension of x. To approximate the optimal solution, Szegedy et al. [1] first utilized box-constrained L-BFGS. Carlini et al. [7] introduced an additional variable to remove the explicit box constraint ($x + r \in [0,1]^m$) and solve the optimization problem using Adam. Optimization-based attacks generally exhibit superior stealthiness compared to other attack types.
- **Generative Model-Based Attack.** This type of attack employs generative models for end-to-end adversarial example generation. For example, Xiao et al. [26] trained a GAN to transform clean samples into adversarial perturbations. Poursaeed et al. [10] proposed a generator capable of producing universal perturbations from random noise, which can be applied to various clean samples to generate adversarial examples. While these attacks demonstrate strong adversarial strength, they often introduce noticeable artifacts (Fig. 2(c)) due to the nature of generative models themselves.
- **Pixel-Based Attack.** Pixel-based attack methods, commonly used in the black-box scenario [28], [29], [35], [36], change the minimum number of pixels necessary to induce misleading. Hence, the corresponding perturbations are sparse. For example, one-pixel attack [35] aimed to deceive the target model by altering only a single pixel. Despite manipulating much fewer pixels compared to other attack types, these attacks often introduce more noticeable perturbations to human observers due to the significant changes made to individual pixels (Fig. 2(d)).
- **Block-Based Attack.** The smallest processing unit in this type of attack is a local region of images rather than an individual pixel. For example, the Square attack [30] samples a perturbation for a randomly selected square region every time and determines whether to perform an update based on improving the objective function. Similar to pixel-based attacks, the perturbations are also easily detectable (Fig. 2(e)).

Beyond restricted adversarial attacks, another class known as unrestricted or semantic adversarial attacks [8], [9], [37]–[41] focuses on preserving semantic plausibility rather than

Fig. 3: Clean image samples in APSD.

pixel-level similarity. For example, Bhattad et al. [9] and Yuan et al. [38] altered object colors while maintaining semantic coherence, such as changing a blue umbrella to red. Jia et al. [39] attacked face recognition by manipulating facial attributes like glasses and smiles using StyleGAN. Chen et al. [40] utilized latent space perturbations in diffusion models to ensure natural-looking adversarial examples, akin to approaches in [8].

B. Full-Reference Image Quality Assessment (FR-IQA)

IQA methods can be categorized into three types based on the availability of a distortion-free reference image: Full-Reference (FR) [42], [43], Reduced-Reference (RR) [44], and No-Reference (NR) [45]. Our study focuses on FR-IQA since attackers possess clean images.

FR-IQA metrics are widely used to quantitatively estimate the level of image degradation relative to a reference image caused by various processes, such as formation, restoration, transformation, or enhancement algorithms. As the evaluation mechanism, IQA plays a critical role in guiding the development of image processing algorithms, including super-resolution [46], defogging [47], deraining [48], image generation [23], etc. The famous New Trends in Image Restoration and Enhancement workshop (NTIRE) [1][2][3] proposed challenges for IQA, encompassing multiple tracks to stimulate the development of various image tasks. However, the assessment of adversarial stealthiness as well as adversarial example visual quality is less explored, hindering the development of imperceptive adversarial attack techniques.

In FR-IQA, the standard research route involves creating a dataset consisting of reference images, distorted images, and MOS, followed by designing objective assessment methods based on the dataset. Several widely-used datasets like LIVE [20], CSIQ [21], TID2013 [22], and PIPAL [23] already exist for most image distortions, such as blurring, compression, and Gaussian noise. These datasets serve as solid foundations for most FR-IQA metrics, both learning-based [18], [19], [49], [50] and conventional [15], [51]–[54]. Regrettably, publicly accessible datasets for adversarial perturbations are currently unavailable. Existing objective assessment schemes cannot be directly employed to evaluate adversarial stealthiness. This is because existing IQA datasets mainly consist of visible distortions, while adversarial stealthiness assessment focuses more on imperceptible perturbations.

[1]https://data.vision.ee.ethz.ch/cvl/ntire22/
[2]https://cvlai.net/ntire/2023/
[3]https://cvlai.net/ntire/2024/

C. Perceptual Assessment Of Adversarial Stealthiness

Adversarial stealthiness can be assessed at two levels: pixel level for restricted attacks and semantic level for unrestricted attacks.

1) Pixel-Level Stealthiness Assessment

This FR assessment compares clean and adversarial images to measure visual similarity. In other words, it assesses the degree of visual degradation of the adversarial image relative to the clean image, where lower visual degradation indicates better stealthiness. Although crucial for developing stealthy restricted attacks [55]–[58], it has not gained widespread attention. The only two relevant works [16], [59] completely followed the subjective criteria [17] widely used in traditional FR-IQA research to collect subjective opinions for visual quality and evaluated the correlation between current objective metrics and human subjective scores. However, these studies did not propose objective methods for future research, and the evaluation of imperceptible perturbations remains unaddressed (see Section III-B). Pixel-level stealthiness assessment is still an open challenge, which this study aims to tackle.

2) Semantic-Level Stealthiness Assessment

Semantic-level assessment, applicable to unrestricted attacks [8], focuses on the realism of adversarial images rather than pixel similarity to specific reference images. The assessment of semantic plausibility in adversarial images is not FR but NR because such decisions are often made based on everyday experience rather than specific clean images [38], [39]. This is particularly true for assessing the stealthiness of the unrestricted attack proposed in [8], where no corresponding clean images exist. Yuan et al. [38] used NIMA [60], an NR metric, to evaluate color-change stealthiness, while Jia et al. [39] employed FID [61] to assess the overall stealthiness of attribute-modified face images. FID measures distribution similarity between generated and real images but lacks precision in evaluating individual images since its precision depends on dataset size.

III. Subjective Assessment of Adversarial Stealthiness

To the best of our knowledge, there are neither subjective assessment criteria nor publicly available datasets for adversarial stealthiness assessment, leading to a long-term lack of dedicated objective assessment metrics. To fill this gap, we build the first **A**dversarial **P**erturbation **S**tealthiness Assessment **D**ataset (APSD), involving human subjective assessments for adversarial perturbations. This dataset can serve as a benchmark, where we can devise objective assessment metrics that are more appropriate for this task. A summary of APSD is provided in Table II.

A. Clean Samples and Adversarial Examples

To construct APSD, we start by collecting many pairs of clean samples and adversarial examples. In this collection

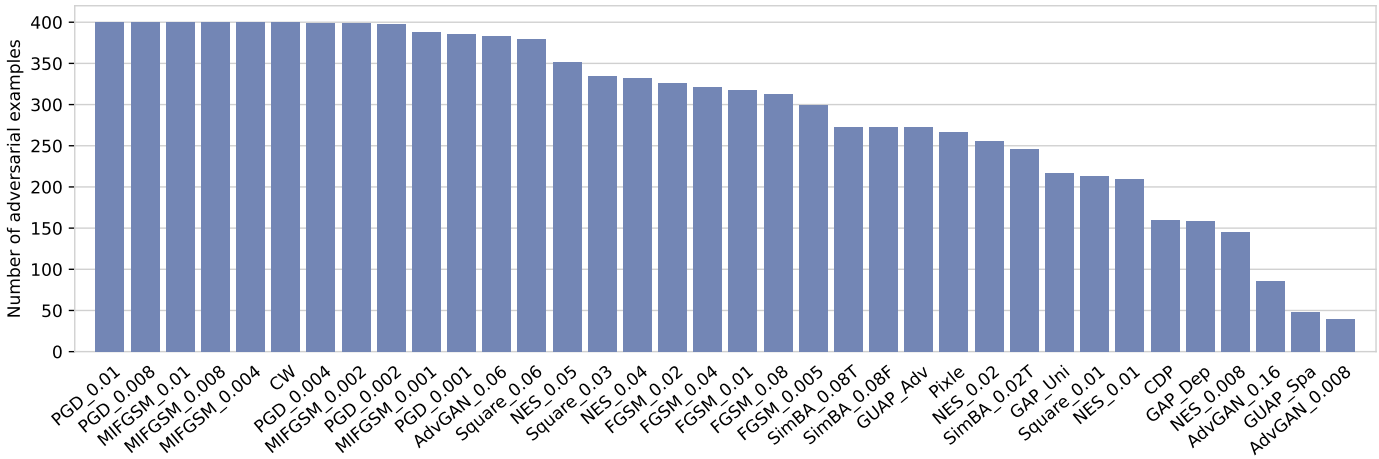


Fig. 4: Distribution of adversarial examples generated by different attacks in APSD. The suffix of each attack indicates the corresponding parameter.

process, we consider the diversity of clean samples and adversarial perturbations. Specifically, we use up to **5** types of adversarial attacks and **12** attack methods to construct the adversarial example dataset, as illustrated in Table I. It ensures that the subsequent objective metrics are *universally applicable* to various adversarial attacks. In contrast, previous studies [16], [59] focused on only one or two attack types.

1) Collection of Clean Samples

Specifically, we focus on attacks against image classifiers because they are widely studied [4], [7], [24], [26] and applied in the real world. To cover a broad range of images, we manually collect clean images from three well-known classification datasets: ImageNet [62], COCO [63], and VOC2012 [64]. The collection is guided by the following three criteria aimed at collecting representative samples from these datasets.

- **High quality.** In the initial filtering, we first exclude images with resolutions below 300×300, as low resolutions are not conducive to subjective observation. We then manually screened and removed images with noticeable artifacts, as these could introduce biases in the perception of adversarial stealthiness.
- **Varying texture richness.** We place a strong emphasis on selecting images with diverse texture richness to ensure our dataset effectively captures the varying impacts of different attack methods. Thus, we collect images ranging from highly detailed textures (e.g., grass, animal fur, textiles) to more uniform ones (e.g., sky, wall). To illustrate this, we analyze texture richness in APSD and an existing large-scale dataset PIPAL [23] (this dataset is proposed for quality assessment of images generated by models) using Local Binary Patterns (LBP) [65] and Gray-Level Co-occurrence Matrix (GLCM) [66]. LBP is a texture descriptor that compares each pixel to its neighbors, generating a binary code, which can capture local texture information robustly. In experiments, we compare each pixel with its 8 neighbor pixels when calculating LBP. GLCM measures the frequency of pixel pairs with specific values in defined spatial relationships (distance and orientation). We consider three distances (1, 2, and 3) and four angles ($0°$, $45°$, $90°$, and $135°$) for pixel pairs when calculating GLCM. From GLCM, we further derive three key texture features, energy, homogeneity, and entropy. Energy measures image regularity, homogeneity reflects the gray-level similarity between adjacent pixels, and entropy indicates texture complexity. In Fig. 5, we present histograms of energy, entropy, homogeneity, and LBP. The histograms of APSD closely resemble those of PIPAL, indicating comparable texture richness. In Fig. 5, we show the histograms of energy, entropy, homogeneity, and LBP. *The histograms of the two datasets exhibit consistent trends and cover similar ranges on the horizontal axis, indicating comparable texture richness.*
- **Diverse categories.** We also collect images from diverse categories to ensure a broad representation of visual content, enhancing the generalizability of our findings. Specifically, APSD includes over **200** categories of clean images, encompassing common types such as animals, plants, transportation, buildings, food, and furniture.

Finally, we gather 400 clean images featuring a variety of objects with diverse levels of texture richness, color richness, and brightness (Fig. 3 shows some examples). We follow the conventions of most IQA datasets [20], [22], [23] by resizing images to a uniform dimension (e.g., 512×512). The complete dataset is available at https://github.com/hcliucs/APSD.

2) Generation of Adversarial Examples

We employ 12 mainstream attacks to generate adversarial examples, encompassing 5 types of attacks: gradient-based [3], [4], [11], [24], optimization-based [7], generative model-based [10], [25]–[27], pixel-based [28], [29], and block-based [30]. In particular, we consider easily and hardly perceptible perturbations, ensuring that our objective assessment method encompasses the functionality of traditional FR-IQA techniques. Therefore, in Table I, we configure different parameters for most attacks to control the visual effect of the generated adversarial example. We use Torchattacks [4] to implement

[4]https://github.com/Harry24k/adversarial-attacks-pytorch

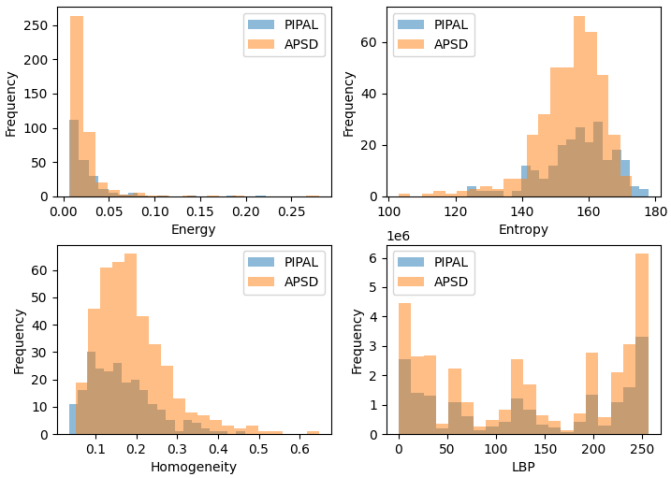


Fig. 5: Texture richness in APSD and PIPAL.

these attacks. Without loss of generality, we randomly choose Inception-V3 [67] as the victim model. We add a resizing layer before the original input layer of Inception-V3 to handle larger input dimensions. Regarding the inclusion of failed samples (i.e., adversarial examples that do not change the target model's output), there are two perspectives. One advocates for their inclusion, arguing that the effectiveness of an adversarial example does not affect stealthiness assessment. The other suggests excluding them, as failed samples are not strictly adversarial examples, and their exclusion avoids ambiguity, ensuring greater rigor. We follow the latter perspective, filtering out failure samples where $f(x') = f(x)$ and collecting **10586** adversarial examples, whose distribution is shown in Fig. 4.

B. Subjective Adversarial Stealthiness Assessment

We call for human observers to assess the stealthiness of all collected adversarial examples. A clear subjective scoring criterion is essential to ensure valid and consistent assessments, as its rationality affects the dataset's overall usability.

1) Motivation

In previous FR-IQA studies [16], [43], [59], [68], [69], observers view two fixed-size images and follow FDIS [17] to assess image quality subjectively. FDIS divides the impact of distortions into: 1) imperceptible, 2) perceptible but not annoying, 3) slightly annoying, 4) annoying, and 5) very annoying. This division pays more attention to the fine-grained assessment of easily perceptible distortions while treating all hardly perceptible distortions as identical. Considering that ❶ the resolution of objective metrics, particularly learning-based ones, depends on the granularity of the subjective score, and ❷ there is a growing demand for assessments of high-stealth perturbations as stealthy attacks evolve, this limitation in FDIS prevents metrics derived from it from delivering a fair evaluation and comparison of various adversarial attacks, particularly stealthy ones. Therefore, we argue that FDIS is unaligned with the need to develop stealth adversarial attacks. Refining the subjective criteria of assessing adversarial stealthiness is necessary, especially for hardly perceptible perturbations.

2) Our Subjective Assessment Criteria

We propose a ***two-step*** assessment method to cover the complete stealthiness spectrum. ❶ **Rough assessment**: observers roughly classify the stealthiness as *strong* (Level 1) or *weak* (Level 2) based on whether they can detect the perceptible difference between the shown two fixed-size images in a fixed time (5 seconds in this experiments). Any difference perceived at a glance indicates the perturbation is *easily perceptible*, corresponding to weak stealthiness. Otherwise, the perturbation is *hardly perceptible*, corresponding to strong stealthiness. ❷ **Fine-grained assessment**: observers refine the rough level by following our proposed fine-grained assessment criteria.

Assessing strong stealthiness subjectively. The overall visual degradation of distorted images relative to reference images serves as the assessing criterion in FDIS. However, assessing strong stealthiness in this way presents challenges since the corresponding perturbation is hardly perceptible, i.e., nearly negligible visual degradation. In light of this, alternative scoring criteria must be sought.

In this study, we establish subjective assessment criteria for strong stealthiness based on the intensity of detail changes. Detail changes can be depicted as the color differences between pixels at identical coordinates. This is because all distortions (including deformation and shifting) change the pixel values at some of the identical coordinates. When the pixel changes beyond a certain threshold, humans can perceive them as changes in color (see Fig. 7). Thus, *spotting significant detail changes is equivalent to finding corresponding significant color differences among all pairs of pixels.* Note that significant color difference is not equivalent to L_∞ since L_∞ treats each color channel equally, while the human eye's sensitivity differs on color channels.

To help observers intuitively perceive the intensity of color changes, we should expand their perceptual dimension beyond the two fixed-size images. As shown in Fig. 7, enlarging pixels into solid color blocks to highlight color differences, similar to the game named Spot The Difference Color [5]. However, identifying these significant color differences by pixel-by-pixel comparison is inefficient and highly labor-intensive.

To address this problem, we provide observers with scaled grayscale difference maps (SGDM) between the clean images and adversarial images, serving as an alternative to the enlargement operation. SGDM is defined as $s \times \text{gray}(x)$ where $\text{gray}(x)$=0.299$\times x_r$+0.587$\times x_g$+0.114$\times x_b$, and x_r, x_g, and x_b represent the RGB channels of image x. The gray operation accounts for the human eye's sensitivity to different color channels. We use s to amplify the color differences for all pairs of pixels simultaneously, gradually increasing s to highlight significant differences earlier. To sum up, our subjective assessment criterion for strong stealthiness can be described as follows:

- Level 1.1: imperceptible even seeing the 20$\times$ difference map (score: 100);
- Level 1.2: imperceptible until seeing the 20$\times$ difference map (score: 95);

[5]https://www.53lu.com/tool/spotcolor/

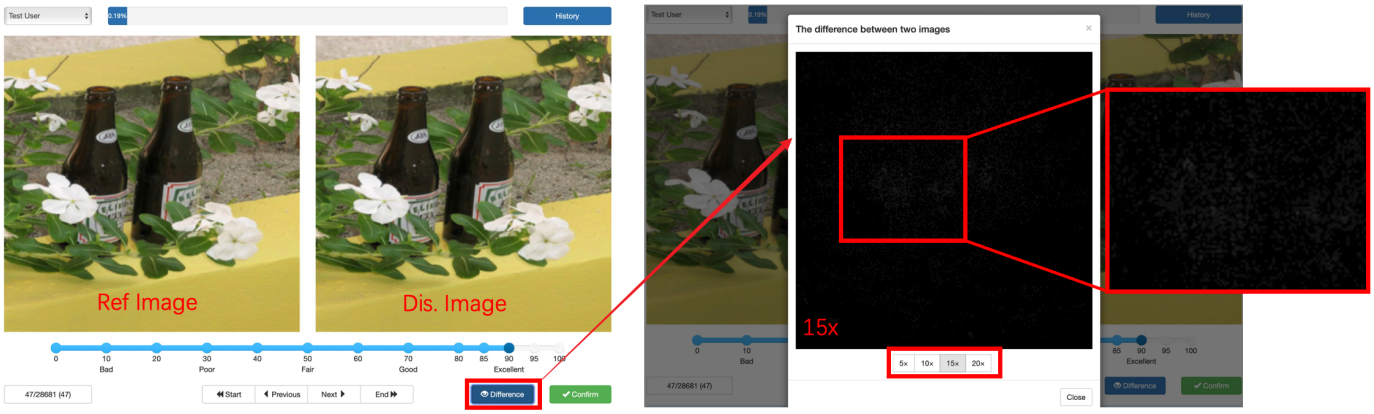


Fig. 6: Rating software. The left is the main interface, and the right is the interface of grayscale difference maps.

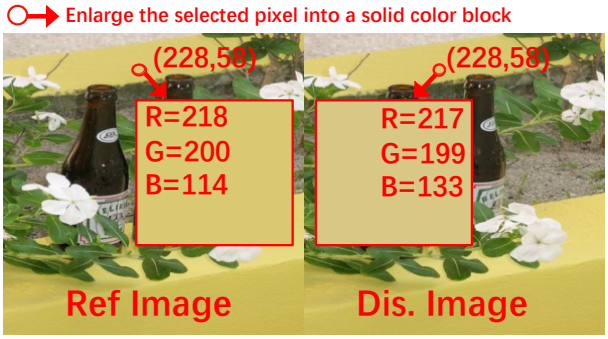
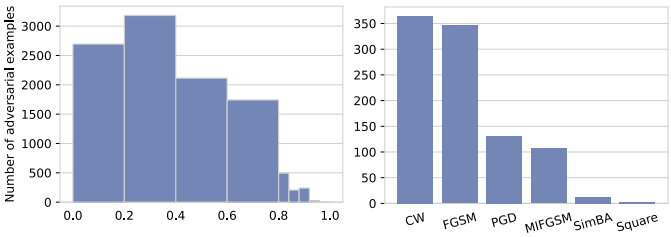


Fig. 7: The color difference between pixel pairs can be highlighted by enlargement. R, G, and B denote the red, green, and blue channels of RGB color space, respectively.



(a) (b)

Fig. 8: Performance of different adversarial attacks. (a) MOS histogram over all attacks. (b) The number of adversarial examples with hardly perceptible perturbations (MOS≥0.8).

- Level 1.3: imperceptible until seeing the $15\times$ difference map (score: 90);
- Level 1.4: imperceptible until seeing the $10\times$ difference map (score: 85);
- Level 1.5: imperceptible until seeing the $5\times$ difference map (score: 80).

Level 1.1 indicates the highest stealthiness. Note that the scaled rate corresponding to each sub-level is determined by extensive empirical observations. Moreover, SGDM can only be accessed when the observer determines that the perturbation is hardly perceptible.

Assessing weak stealthiness subjectively. We still assess weak stealthiness based on the overall visual degradation between the adversarial examples and clean samples [16], [59] since the degradation is distinguishable in this case. This also ensures that our subsequent objective metric applies to traditional FR-IQA tasks. The more severe the quality degradation, the worse the stealthiness in this case. Specifically, we use the last four levels of FDIS to quantify weak stealthiness subjectively:

- Level 2.1: perceptible, but not annoying (score: [60,80));
- Level 2.2: slightly annoying (score: [40,60));
- Level 2.3: annoying (score: [20,40));
- Level 2.4: very annoying (score: [0,20)).

We must emphasize that SGDM is inaccessible to observers when assessing weak stealthiness.

Our subjective assessment criteria proposed for strong and weak stealthiness address the limitations of FDIS by elucidating assessment methods across the entirety of the stealthiness spectrum, which are more in line with the development needs of adversarial attacks.

3) Subjective Experiments

We recruit human observers to score the stealthiness of all adversarial examples based on the proposed two-step assessment method.

Rating software. We design a rating software as depicted in Fig. 6. Its main interface (the left image in Fig. 6) displays a pair of fixed-size images, where the reference image is consistently positioned on the left and the adversarial image on the right. In the first assessment step, observers determine whether the stealthiness belongs to strong or weak stealthiness by looking at the two images. In the second step, *if the perturbation is regarded as strong stealthiness*, the observers are permitted to click the blue "difference" button for a new interface (the right image in Fig. 6), which shows the scaled grayscale difference maps. Observers can modify the scaling rate within the range of $5\times$ to $20\times$ via interface buttons below the disparity map. Note that observers are always instructed to start at the lowest scaling rate and gradually increase the scale factor until the scaled perturbation becomes perceptible or the maximum factor is reached. *If the perturbation possesses weak stealthiness*, observers can only continue to rate the stealthiness by observing the two fixed-size images.

Subjective scoring configurations. All subjective evaluations are conducted in a laboratory environment with normal indoor illumination conditions, where the incident light falling on

the screen is 90 lux and the environmental illumination from behind the monitor is 240 lux. The rating software is displayed on four identical 27-inch monitors, each with a resolution of 2560×1440 and default settings for color calibration, brightness, contrast, etc. To prevent extra distortions, all images are shown at the original resolution (512×512). We recruited 60 human observers for our subjective assessments, comprising 33 females and 27 males with normal or corrected vision, aged from 22 to 30 years old. Approximately two-fifths of observers have no prior background knowledge of adversarial attacks. We divided all observers into 4 groups, with 15 observers in each group. Similarly, we divided the clean samples in APSD into 4 sets, with 100 samples in each set. Each group of observers is assigned a set of clean samples and corresponding adversarial examples. Thus, each adversarial example is evaluated by 15 observers. We require all observers to maintain a fixed viewing distance of approximately four times the image height. A 20-minute rating is followed by a mandatory break to prevent visual fatigue.

Processing of raw scores. We process all raw subjective scores to generate the final MOS for each adversarial example. Specifically, we apply outlier detection as suggested in [17] to do this. A raw score $\mu_{i,j}$ of adversarial example i from observer j is considered an outlier if it falls outside the corresponding 95% confidence interval $[\bar{\mu_i} - \delta_i, \bar{\mu_i} + \delta_i]$. $\bar{\mu_i}$ and δ_i are defined as

$$\begin{aligned} \bar{\mu_i} &= \frac{1}{N}\sum_{j=1}^{N}\mu_{i,j}, \\ \delta_i &= 1.96 \times \frac{\sigma_i}{\sqrt{N}}, \end{aligned} \tag{3}$$

where

$$\sigma_i = \sqrt{\sum_{j=1}^{N}\frac{\left(\mu_{i,j} - \bar{\mu_i}\right)^2}{N-1}}, \tag{4}$$

and N is the number of observers ($N = 15$). After removing outliers, we average the remaining scores for each adversarial example to obtain its MOS. Note that all raw scores are divided by 100, making the MOS values in APSD range from 0 to 1. The MOS histogram in Fig. 8(a) reflects that adversarial perturbations in the APSD cover all levels of stealthiness except Level 1.1. In Fig. 8(b), we highlight the attacks that produce hardly perceptible perturbations, where CW [7] and FGSM [24] (mainly FGSM_0.005) account for the majority.

C. Analysis of APSD

We first demonstrate the validity of MOS values in APSD. Then, we identify the limitations of existing IQA metrics in assessing adversarial stealthiness.

Scoring consistency. To verify the validity of MOS values, we illustrate the distribution of $2 \times \delta_i$ in Fig. 9. It shows a high agreement among all observers in assessing the same adversarial examples, with $2 \times \delta_i$ mainly falling between 0.0824 and 0.1844. Meanwhile, in Fig. 10, we also show that each group of observers tends to agree on assessing the same attack. Both the two figures provide strong evidence supporting the high validity of APSD.

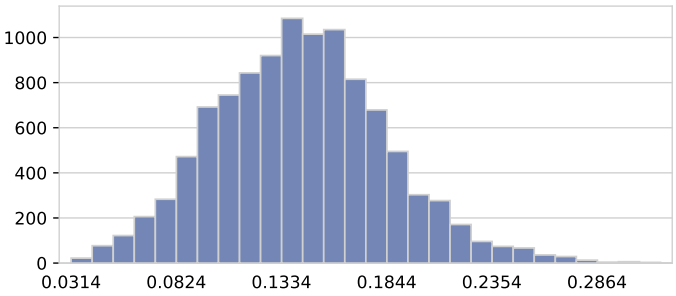


Fig. 9: Distribution of raw subjective scores' confidence intervals.

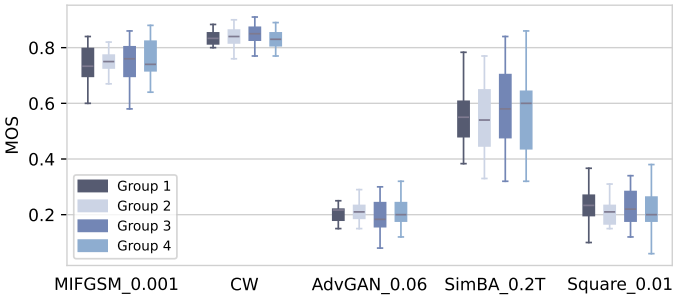
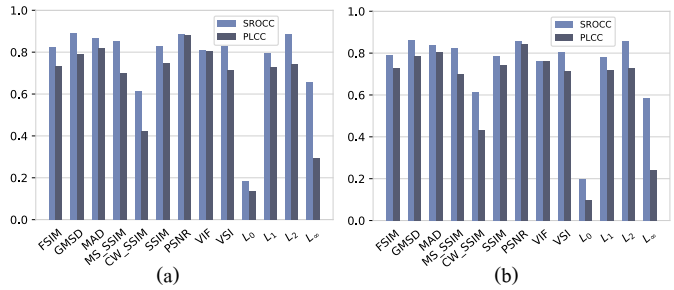


Fig. 10: Subjective scores of four groups of observers for the same attacks.

Performance of the existing IQA metrics on APSD. We use SROCC and PLCC, widely recognized metrics in IQA studies, to calculate the correlation between the objective scores and MOS values. A satisfactory objective metric should own a strong correlation. PLCC assesses the linear correlation between ground truth and the predicted scores, whereas SROCC describes the level of monotonic correlation. PLCC and SROCC range from 0 to 1 for positive correlations, with higher values indicating stronger correlation and more accurate objective assessment. Before computing PLCC and SROCC, we normalize the scores of each metric, as well as the subjective scores, by

$$s_i^j = \frac{s_i^j - s_{min}^j}{s_{max}^j - s_{min}^j}, \tag{5}$$

where s_i^j represents the score of i-th sample on j-th metric, and s_{max}^j and s_{min}^j are the maximum and minimum scores of the j-th metric. Fig. 11(a) confirms the inadequate performance of current metrics when assessing the adversarial stealthiness. Although PSNR, a commonly used metric, achieves the best overall performance, its SROCC and PLCC values remain below 0.89, suggesting room for improvement in objective adversarial stealthiness assessment. Moreover, we re-evaluate these metrics on easily perceptible perturbations (MOS<0.8) (Fig. 11(b)), revealing the struggle of current metrics in accurately assessing weak stealthiness even if we completely follow FDIS [17]. Hence, adversarial perturbations differ significantly from traditional distortions in perception. These experimental results strengthen our determination to develop a new objective metric aligned with subjective perception.



(a) (b)

Fig. 11: Performance of the existing IQA metrics and L_p norms on the proposed APSD. (a) Evaluation on the whole APSD. (b) Evaluation on adversarial examples with easily perceptible perturbations (MOS$<$0.8).

IV. Objective Method

Section III-C shows that current assessment metrics cannot effectively evaluate adversarial samples' stealthiness. This section introduces a novel learning-based objective assessment method for measuring adversarial stealthiness.

Fig. 12 shows the overall pipeline of the proposed **A**ttention-based **A**dversarial **S**tealthiness Assessment **M**odel (A2SM). A2SM contains three stages: (1) feature extraction, (2) feature fusion, and (3) score prediction. Suitable attention mechanisms are employed in each stage to enhance model performance.

- **Feature extraction.** We integrate two types of feature extractors: a vision transformer (ViT) extractor [70] and a convolutional neural network (CNN) extractor [71]. The ViT extractor captures global features for a holistic understanding, while the CNN extractor focuses on local features for detailed analysis.
- **Feature fusion.** We calibrate and integrate extracted features from a global perspective. First, we calibrate the local features using an offset module under the guidance of the global features. Then, we calibrate the difference map (F_f^d) using two specially designed modules, self-attention and cross-attention. The former calibrates the difference map based on itself, while the latter calibrates the difference map based on the reference image. After the calibrations, we fuse all necessary features and enter the last stage.
- **Score prediction**. Using a spatial attention module, we predict scores for all patches from a local perspective. The scores are then weighted and summed to yield the final predicted score.

In the subsequent sections, we detail our designs of A2SM.

A. Feature Extraction Stage

The completeness of basic features is crucial in achieving more accurate assessments. Most previous studies [19], [49], [72] solely rely on a single CNN extractor. Although CNNs excel at capturing rich low-level features like textures, they struggle to grasp the significance of each patch to the entire image, a critical aspect of the assessment. Lao et al. [18] proposed using another ViT extractor to address this limitation at this stage. The ViT can capture features over long distances, compensating for the shortcomings of the CNN extractor. Empirical evidence in [18] demonstrates that the additional extractor enhances the scoring model's performance. Thus, we adopt this approach in our study, combining both CNN and ViT extractors to capture a comprehensive feature set.

In particular, for the ViT extractor, we divide its inputs into 8×8 patches, where the input size is 224×224. The output of the ViT extractor is obtained from the first five blocks of ViT. We concatenate the outputs of each block along the channel dimension to form the global features (F_v^x and $F_v^{x'}$). Regarding the CNN extractor, we use ResNet50 [71] as its backbone. Similarly, we use the outputs of ResNet50's first three blocks in stage 1 to yield the local features (F_c^x and $F_c^{x'}$). We focus primarily on the shallow features of both the ViT and CNN because the shallow features concentrate on the low-level characteristics of the images, making them applicable for various vision tasks.

B. Feature Fusion Stage

Offset module. Generative models often produce slight errors and misalignments of edges in the corresponding adversarial examples. However, vanilla convolution operations have limitations in capturing such distortions, hurting model performance. In [18], [73], deformable convolution [74] has demonstrated its superiority in handling similar distortions. Therefore, we implement an offset module based on deformable convolution to calibrate extracted local features, as shown in Fig. 12.

Specifically, we employ a vanilla convolution layer ($k = 3$, $s = 1$, and $p = 1$, where k and s represent kernel size, stride, and padding) and a deformable convolution layer ($k = 3$, $s = 1$, and $p = 1$) to build the offset model. First, the vanilla convolution layer is applied to F_v^x to produce an offset map. Then, the deformable convolution layer is used to calibrate the local features under the guidance of the offset map. Generating the offset map based on F_v^x because F_v^x represents a holistic image understanding.

Fusion module. The two fusion modules in Fig. 12 (Fusion1 and Fusion2) are used to integrate necessary features. For each fusion module, we first concatenate all its inputs. Then, we pass the concatenated results through the corresponding fusion module. In experiments, we build Fusion1 using a vanilla convolution layer ($k = 1$, $s = 1$, and $p = 0$), while Fusion2 consists of two vanilla convolution layers ($k = 3$, $s = 1$, and $p = 1$).

Self-attention module and cross-attention module. In most previous studies [18], [49], the difference map F_f^d between $F_f^{x'}$ and F_f^x are directly used for the final prediction. However, the contribution of distortions in different regions to the final score varies as human observers tend to focus more on significant patches. After analysis, we believe that the significance of each patch is determined by two factors. The first is the relative strength of perturbation in different patches. The stronger the interference, the easier it will be to perceive. The second is the relative importance of each patch in the clean image. Distortions in key patches are often more annoying than those in non-key patches. We propose a self-attention module and a cross-attention module to capture two types of relative

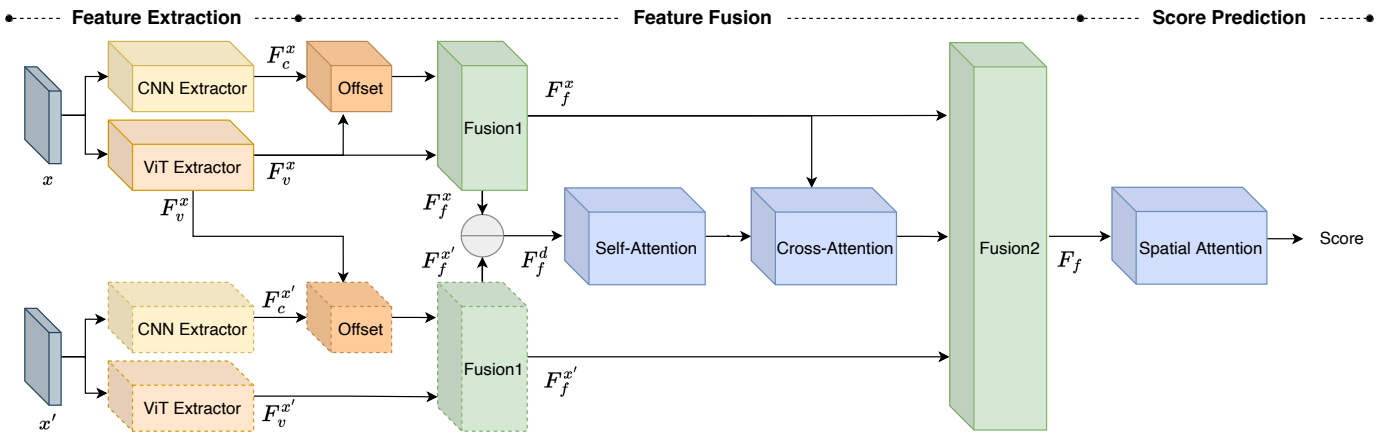


Fig. 12: Overview of A2SM. All dotted cubes in this figure mean that they share weights with their corresponding solid cubes. $\ominus$ denotes element-wise subtracting. F_c^x and $F_c^{x'}$ denote local features, while F_v^x and $F_v^{x'}$ are global features. F_f^x and $F_f^{x'}$ are initial fused features. F_f^d is a difference map between F_f^x and $F_f^{x'}$. F_f means the final fused features.

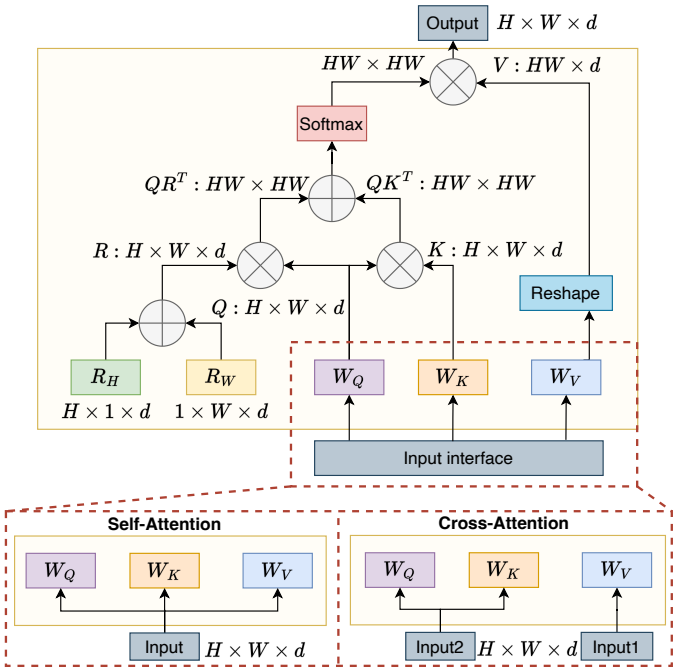


Fig. 13: Attention modules in the feature fusion stage of A2SM. Self-attention and cross-attention are multi-head, and there is a head of them. The self-attention module only accepts one input, while the cross-attention module accepts two. W_Q, W_K, and W_V are 1×1 convolution layers. R_H and R_W are two learnable parameters denoting the relative position encoding for height and width. $\oplus$, $\otimes$, H, W, and d mean the element-wise sum, matrix multiplication, height, width, and channel respectively.

relationships. Both modules capture the relative relationships from a global perspective. In contrast, works in [19], [72] only considered one of the two factors and calibrated the difference map from a local perspective.

We first calibrate the difference map using a self-attention module (Fig. 13) that captures the significance of patches according to the difference map. In experiments, we employ multi-head self-attention (MHSA) proposed in [75] to implement this module, which is similar to the self-attention in ViT [70]. In Fig. 13, we illustrate one head of the self-attention module, whose input is a chunk of F_f^d. Specifically, we split F_f^d into k ($k = 8$) chunks along the channel dimension for k heads. With c ($c = 1024$) denoting the channels of F_f^d, we have $d = c/k$. In each head, we first convert its input into query Q, key K, and value V through three vanilla convolution operations ($k = 1$, $s = 1$, and $p = 0$), i.e., W_Q, W_K, and W_V in Fig. 13. The first matrix multiplication in Fig. 13 calculates the significance of patches based on Q and K. Subsequently, we embed two-dimensional location information into Q through matrix multiplication ($\otimes$). The two-dimensional location information is learned by two parameters, R_H and R_W, representing the relative position encoding for height and width. Compared with the one-dimensional location encoding in ViT [70], two-dimensional position encoding is more suitable for images.

We then calibrate the difference map under the guidance of the clean sample using a cross-attention module. As illustrated in Fig. 13, we mainly modify the input interface so that the cross-attention module can capture the significance of patches determined by the clean sample. Each head of the cross-attention module has two inputs. The Input1 is a chunk of the self-attention module's output (i.e., initially calibrated difference map) while the Input2 is a chunk of F_f^x (i.e., features of the clean sample). In the cross-attention, chunks of F_f^x are converted into Q and K, which are used to calculate the significance of patches. The matrix multiplication at the top of Fig. 13 completes the calibration for the difference map under the guidance of the clean sample.

C. Score Prediction Stage.

In the last stage, we predict a score for each patch represented by $F_{f_{ij}}$ from a local perspective ($F_{f_{ij}} \in \mathbb{R}^{1\times1\times P}$, i and j are the coordinates for height and width of F_f). We propose a spatial attention module to do this. As illustrated in Fig. 14, the spatial attention module is two-branch. The above branch in Fig. 14 predicts an attention map (AM) for all patches, where GAP and GMP mean channel-wised global average pooling and channel-wised global max pooling. At the same time, the bottom branch predicts a score map

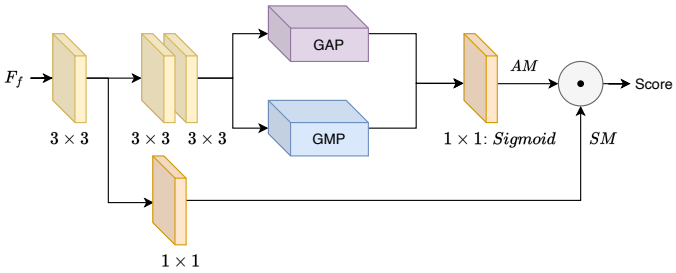


Fig. 14: Spatial attention. Each cuboid marked with 3×3 or 1×1 represents a vanilla convolution layer. GAP is channel-wise global average pooling, and GMP is channel-wise global max pooling. $\odot$ means dot product. For all patches, the top branch predicts an attention map (AM), and the bottom predicts a score map (SM).

TABLE II: Summary of popular IQA datasets and our APSD. DMOS is inversely proportional to MOS. Dis. is the abbreviation of distortion

Dataset	#Ref	#Dis.	Dis. Type	Rating Type	Score Range
LIVE [20]	29	779	5	DMOS	[0,100]
CSIQ [21]	30	886	6	DMOS	[0,1]
TID2013 [22]	25	3000	24	MOS	[0,9]
PIPAL [23]	250	25850	40	MOS	[917,1836]
APSD (ours)	400	10586	12	MOS	[0,1]

$SM \in \mathbb{R}^{H\times W\times 1}$ for all patches using a 1×1 vanilla convolution layer. The final score is computed by:

$$\text{score} = \frac{SM \odot AM}{\sum AM}, \tag{6}$$

where $\odot$ denotes the dot product.

To train A2SM, we use mean squared error (MES) between the predicted score and MOS as the training loss. Note that this distribution Fig. 4 is imbalanced because we only consider successful cases. To mitigate the potential negative impact of the imbalanced distribution on the training, we take the following two measures. First, we apply data augmentation (random rotation and flipping) to these adversarial examples generated by the last six methods in Fig. 4. The rotation and flipping do not affect the ground truth of MOS. Second, we apply the re-weighting scheme proposed in [76] to re-balance the loss among different attacks.

V. Experiments

In this section, we compare the performance of A2SM and other state-of-the-art IQA methods on various datasets.

A. Configuration

Datasets. We consider five datasets for evaluations, including our APSD, LIVE [20], CSIQ [21], TID2013 [22], and PIPAL [23]. The last four datasets are commonly used IQA datasets. We summarize all datasets in Table II. LIVE, CSIQ, and TID2013 only involve traditional image distortions, e.g., blurring. PIPAL involves traditional distorted images and images restored by multiple types of image restoration algorithms (e.g., denoising, super-resolution, etc). APSD is the first large-scale adversarial example dataset, while the other four datasets do not involve similar distortion types. For each dataset, we randomly split it into a training set (60%), a validation set (20%), and a test set (20%) according to reference images. Meanwhile, we normalize all datasets' MOS (or difference mean opinion score (DMOS)) between 0 and 1 using Eq. (5).

TABLE III: Performance comparison on APSD

Category	Method	PLCC	SROCC
Traditional	FSIM [52]	0.735	0.825
	GMSD [51]	0.790	0.889
	MAD [21]	0.818	0.868
	MS_SSIM [78]	0.698	0.853
	CW_SSIM [79]	0.424	0.615
	SSIM [15]	0.748	0.827
	PSNR	0.879	0.885
	VIF [53]	0.804	0.808
	VSI [54]	0.712	0.830
	L_0	0.137	0.182
	L_1	0.730	0.796
	L_2	0.741	0.885
	L_∞	0.292	0.657
Learning	AHIQ* [18]	0.453	0.457
	IQT [80]	0.941	0.942
	WaDIQaM-FR [50]	0.950	0.946
	RADN [73]	0.952	0.947
	DeepIQA [81]	0.964	0.955
	DeepQA [82]	0.972	0.966
	AHIQ [18]	0.976	0.972
	A2SM (ours)	**0.984**	**0.978**

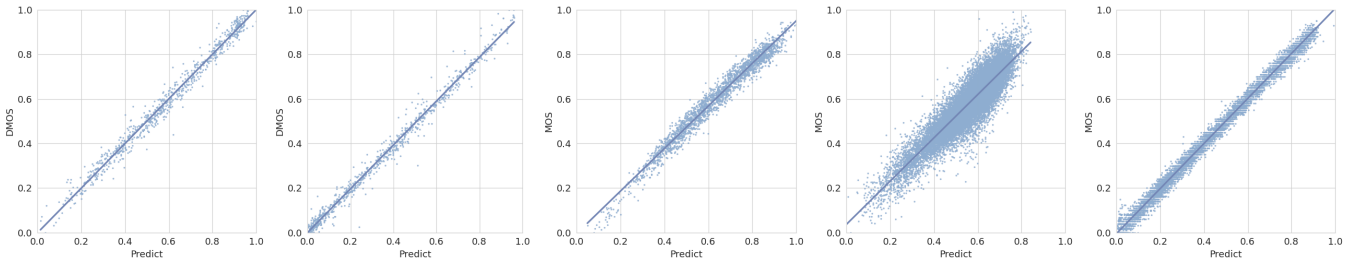
Implementation details. During the training phase, we assign pre-trained weights of ViT [70] and ResNet50 [71] on ImageNet to the ViT extractor and CNN extractor, respectively, and freeze them. We use Adam optimizer [77] whose initial learning rate is $1e^{-4}$ to optimize the remaining components of A2SM while the batch size is 8. We also apply random rotation and random flip (both operations do not affect the adversarial stealthiness) to augment all datasets. All input images are resized into a size of 224×224. During the test phase, we randomly crop images into a size of 224×224 and repeat the prediction 15 times. The average score of all cropped images is the final score. Such operation is common in previous studies [18], which can improve assessment accuracy to a certain extent. We implement A2SM using Pytorch and train it using a single NVIDIA RTX3090 GPU.

Metrics. We use PLCC and SROCC to evaluate the performance of all test methods as we have done in Section III-C. PLCC and SROCC quantify the correlation between the objective and subjective scores. *A higher correlation means that the corresponding objective method is well-aligned with the HVS, i.e., a more accurate objective assessment.*

B. Overall Evaluation

In this section, we first evaluate the performance of our A2SM on different datasets to show the effectiveness of A2SM. Then, we compare A2SM with the existing popular IQA methods, including conventional and learning-based ones, on our APSD and other IQA datasets. For conventional methods, we realize them using two open-source libraries: PYIQA [6] and PIQ [7]. For learning-based methods, we use their open-

[6]https://github.com/chaofengc/IQA-PyTorch
[7]https://github.com/photosynthesis-team/piq



(a) LIVE (b) CSIQ (c) TID2013 (d) PIPAL (e) APSD (ours)

Fig. 15: Objective scores of A2SM vs. the MOS values on different datasets.

source codes. In particular, when we conduct evaluations on APSD, we retrain all learning-based methods on APSD using their default hyper-parameters.

Effectiveness of A2SM on different datasets. Fig. 15 shows the effectiveness of A2SM on different datasets. In Fig. 15(e), all points lie very close to the line **MOS** $=$ **Predict**, indicating that 2ASM effectively predicts the stealthiness of different levels. Meanwhile, 2ASM also achieves satisfactory assessments on the other four datasets, as illustrated in Fig. 15(a)-(d). The scatter plot of 2ASM on PIPAL (Fig. 15(d)) appears slightly loose because PIPAL contains far more distortion types than other datasets, significantly increasing the fitting difficulty.

Superiority of A2SM. We compare A2SM with other state-of-the-art IQA methods in Table III and Table IV. Table III shows the PLCC and SROCC of A2SM and other IQA metrics on our APSD dataset, where LPIPS* and AHIQ* denote the utilization of pre-trained weights provided by PYIQA, without additional training on APSD. The results show that A2SM outperforms other methods in predicting adversarial stealthiness, achieving the highest PLCC and SROCC values of 0.984 and 0.978, respectively. It means that A2SM can be used to evaluate the stealthiness of adversarial samples effectively. Table IV shows the PLCC and SROCC results of different metrics on the traditional IQA datasets. It is worth noting from the results that A2SM consistently performs the best or at a competitive level. Notably, A2SM attains the highest values of PLCC and SROCC on the two complex datasets, TID2013 [22] and PIPAL [23], while remaining comparable to the best-performing methods on LIVE [20] and CSIQ [21]. In conclusion, these results demonstrate the versatility of A2SM's design, allowing its application to various objective assessment tasks. Besides, as anticipated, learning-based methods consistently outperform traditional methods in both Tables III and IV.

C. Ablation Study

In [18], the authors discussed the impact of the two feature extractors (ViT and CNN) and the deformable convolution for model performance. Here, we discuss the efficiency of the proposed self-attention and cross-attention modules. As shown in Table VI, both modules contribute to improving A2SM's performance, and their respective contributions to the improvement seem similar. A2SM achieves the best performance when the reference image and difference map are considered simultaneously during the calibration, supporting our previous analysis that both factors affect the significance of patches. In summary, modeling human scoring habits can improve the performance of the objective scoring model.

D. Cross-Dataset Evaluation

We evaluate the generalization of A2SM and verify the necessity of constructing APSD. Specifically, we train A2SM on a source dataset and test it on different destination datasets.

Generalization. In Table V, the A2SM trained on TID2013 [22] (denoted as A2SM-TID) exhibits the best cross-dataset capability. Particularly, for LIVE and CSIQ, A2SM-TID is highly available. In contrast, A2SM-TID generalization to complex destination datasets, PIPAL and APSD, is relatively limited. However, A2SM-TID still outperforms or is competitive with most traditional IQA methods on these two datasets. For example, when assessing adversarial stealthiness, A2SM-TID's PLCC and SROCC values are 0.873 and 0.875, while the PSNR values are 0.879 and 0.885. We also test A2SM's generalization to other victim models by constructing an additional testing dataset.

Necessity. Although A2SM-TID possesses good generalization, its performance still degrades heavily compared to A2SM trained on APSD (Table III and IV). Such an outcome suggests that it is necessary to construct a dedicated dataset for specific distortion types.

E. Interpretability in A2SM

We visualize the attention map in the spatial attention module of A2SM, noting that it employs different strategies for scoring images with varying levels of stealth. Specifically, the attention map is visualized as $Clip(0.6 \times x' + 0.8 \times AM_r)$, where AM_r is the resized result of AM to the size of x', and $Clip$ truncates values between 0 and 1. As shown in Fig. 16, A2SM attends to the entire image for high-stealth images, while it focuses on localized areas with noticeable perturbations in low-stealth images. This distinction is logical: ❶ high-stealth images must maintain quality across all areas, necessitating a global assessment to catch minor issues, resulting in distributed attention; ❷ low-stealth images often exhibit significant distortions, so local assessment of obvious distortions suffices to conclude low stealth, concentrating attention on those areas. **This adaptive attention highlights A2SM's ability to adjust to different stealth levels.**

Fig 17 shows that for different attacks, A2SM consistently demonstrates global attention to high-stealth images (top)

TABLE IV: Performance comparison on LIVE, CSIQ, TID2013, and PIPAL.

Category	Method	LIVE		CSIQ		TID2013		PIPAL	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
Traditional	FSIMc [52]	0.961	0.965	0.919	0.931	0.877	0.851	0.559	0.468
	GMSD [51]	0.957	0.960	0.945	0.950	0.855	0.804	0.619	0.591
	MAD [21]	0.968	0.967	0.950	0.947	0.827	0.781	0.626	0.608
	MS_SSIM [78]	0.940	0.951	0.889	0.906	0.830	0.786	0.563	0.486
	SSIM [15]	0.937	0.948	0.852	0.865	0.777	0.727	0.398	0.340
	PSNR	0.865	0.873	0.819	0.810	0.677	0.687	0.292	0.255
	VIF [53]	0.960	0.964	0.913	0.911	0.771	0.677	0.524	0.433
	VSI [54]	0.948	0.952	0.928	0.942	0.900	0.897	0.516	0.450
	NLPD [83]	0.932	0.937	0.923	0.932	0.839	0.800	0.511	0.498
Learning	LPIPS [84]	0.934	0.932	0.927	0.903	0.749	0.670	0.839	0.843
	JSPL [19]	0.983	0.980	0.970	**0.977**	0.949	0.940	0.877	0.874
	DISTS [85]	0.955	0.955	0.946	0.946	0.855	0.830	0.686	0.674
	WaDIQaM-FR [50]	0.980	0.970	0.951	0.960	0.946	0.940	0.654	0.678
	RADN [73]	0.878	0.884	0.846	0.828	0.845	0.830	0.867	0.866
	DeepQA [82]	0.982	0.981	0.965	0.961	0.947	0.939	0.795	0.785
	AHIQ [18]	**0.989**	**0.984**	0.978	0.975	0.968	0.962	0.865	0.852
	A2SM (ours)	0.982	0.976	**0.981**	0.973	**0.974**	**0.972**	**0.881**	**0.875**

TABLE V: Cross-dataset evaluation for A2SM

Source	Destination									
	LIVE		CSIQ		TID2013		PIPAL		APSD	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
LIVE	-	-	0.909	0.880	0.805	0.753	0.636	0.614	0.722	0.770
CSIQ	**0.950**	0.946	-	-	0.832	0.810	0.639	0.645	0.812	0.825
TID2013	**0.950**	**0.953**	**0.954**	**0.936**	-	-	**0.671**	**0.665**	**0.873**	**0.875**
PIPAL	0.842	0.855	0.867	0.872	0.795	0.770	-	-	0.677	0.767
APSD	0.858	0.892	0.855	0.870	0.719	0.728	0.528	0.531	-	-

TABLE VI: Impact of self-attention and cross-attention

No.	Self	Cross	PLCC	SROCC
1	✘	✘	0.973	0.970
2	✔	✘	0.981	0.975
3	✘	✔	0.981	0.976
4	✔	✔	**0.984**	**0.978**

TABLE VII: Increase in predicted scores after attack

Source	Max	Min	Average	Median
$\epsilon = 0.005$	0.167	0.002	0.090	0.088
$\epsilon = 0.01$	0.297	0.117	0.183	0.179
$\epsilon = 0.02$	0.269	0.139	0.197	0.197
$\epsilon = 0.04$	0.251	0.111	0.165	0.165
$\epsilon = 0.08$	0.154	0.048	0.089	0.088

Fig. 16: Adaptive attention in A2SM when scoring images with different levels of stealth. The top images are adversarial examples generated by FGSM with different perturbation budgets. The bottom images are the results of the combination of x' and AM in the spatial attention module.

while focusing on locally significant distorted areas in low-stealth images (bottom). The results illustrate that A2SM's adaptive scoring strategy is both effective and universally applicable across various attack methods.

VI. Limitations

A. Vulnerability to Adversarial Attacks

Like all previous learning-based scoring schemes [86], [87], as well as NR-IQA models [88], A2SM is also susceptible to being fooled by adversarial attacks. To illustrate this, we use adversarial examples generated by FGSM ($\epsilon = 0.005, 0.01, 0.02, 0.04, 0.08$) in APSD as clean samples and further perturbs them using targeted FGSM ($\epsilon = 0.005$, target score is 1) to attack A2SM. Notably, FGSM tends to generate hardly perceptible perturbations when $\epsilon = 0.005$ (Fig. 8(b)), so these new perturbations do not significantly affect the visual appearance of images already with weak stealthiness.

As shown in Table VII, A2SM is vulnerable to these new perturbations, generally increasing predicted scores by a sub-level. Thus, A2SM is suitable only for post-event evaluations and cannot guide adversarial example generation

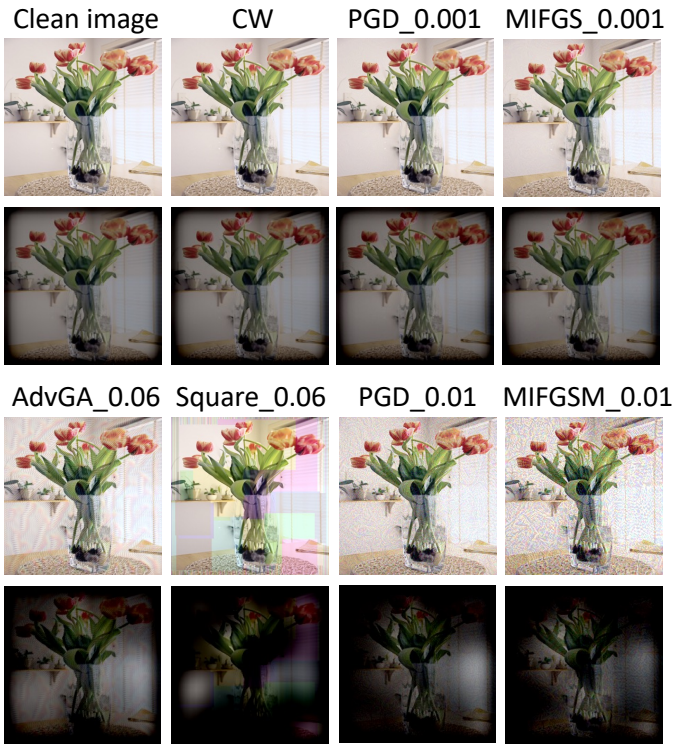


Fig. 17: The adaptive attention in A2SM is universal to different attack methods.

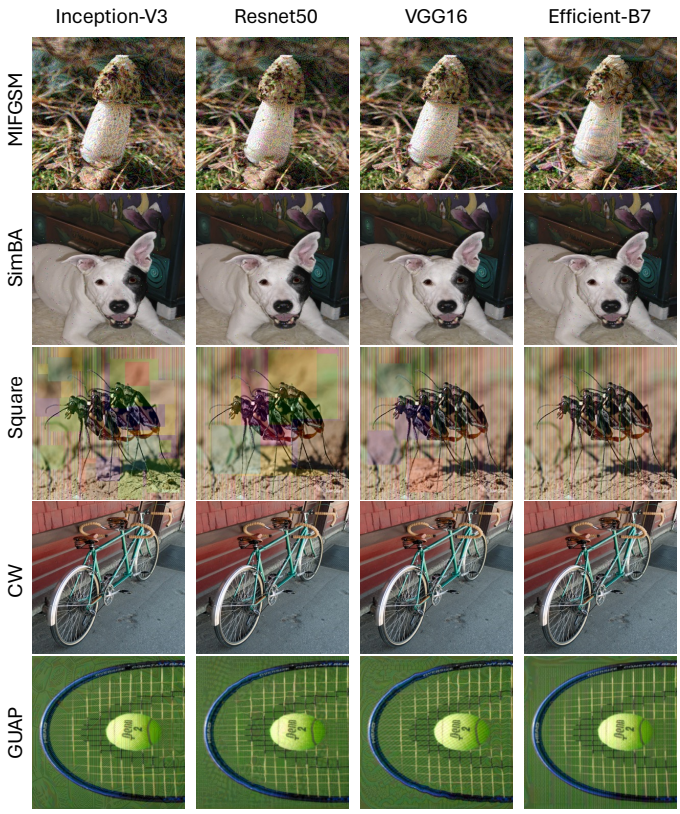


Fig. 18: Adversarial examples generated by different attack methods targeting different models. The attack methods have a more significant effect on the perturbation pattern than the choice of victim model. The perturbation patterns shown in each column differ significantly, while the patterns within each row remain similar.

like traditional methods such as L_p-norms. Despite this, the best approach is to use traditional metrics for generating adversarial examples and learning-based methods for objective evaluations. Our future research will focus on enhancing scoring model robustness and integrating them into adversarial example generation.

B. Limited Classification Models

In constructing APSD, we concentrated on the Inception-V3 model as the victim. However, different victim models yield varying perturbation patterns; therefore, APSD may not be sufficient to capture the full range of adversarial perturbations. Nevertheless, we find that A2SM has the potential to generalize to other classification models because, for most attack methods, the influence of the target model on the perturbation patterns is minimal, resulting in no significant changes (see Fig. 18). In contrast, adversarial attack methods have a more substantial impact on perturbation patterns in Fig. 18. This suggests that the diversity of attack methods is more critical than the diversity of models when constructing the dataset. Nevertheless, we still believe it is necessary to extend APSD by considering other victim models.

C. Limited Tasks

We focus solely on classification tasks in this study, which may limit the applicability of the dataset to other computer vision tasks, such as object detection and segmentation. This limitation primarily arises from the differentiated distribution of adversarial perturbations. For classification tasks, adversarial perturbations are generally spread across the entire image (see the right image in Fig. 19). In contrast, detection and segmentation tasks involve more localized perturbations (see the middle image in Fig. 19), concentrating on objects of interest, as these tasks focus on object-level or region-specific outputs. This discrepancy in perturbation patterns complicates the direct transfer of the visual quality evaluation framework from classification to detection or segmentation tasks. As existing subjective criteria are optimized for global image quality evaluation, they may not align with the region-specific requirements of detection and segmentation tasks. To address this limitation, future work could explore adapting subjective criteria to better capture region-specific assessments.

VII. Conclusion

This paper proposes novel subjective and objective assessment methods for adversarial stealthiness, bridging a long-standing technical gap. To evaluate stealthiness subjectively, we analyze the limitations of existing subjective criteria and expand human observers' perception dimension to refine stealthiness for subtle pixel changes. Then, we construct the first large-scale benchmark dataset, involving 12 state-of-the-art adversarial attacks and human subjective scores for the adversarial stealthiness. Moreover, we design a new objective assessment model that simulates human scoring habits as much

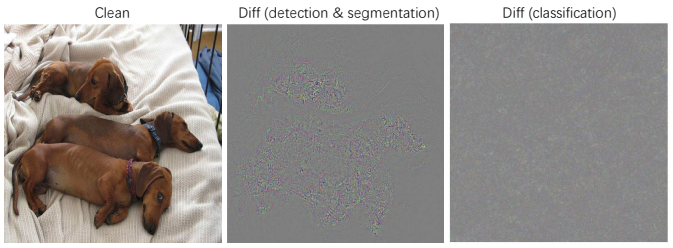


Fig. 19: Differences in adversarial perturbation distribution across tasks. The left image shows the clean image. The middle image illustrates the difference between the clean image and the adversarial example generated using DAG [89] for the object detection and segmentation task. The right image shows the difference between the clean image and the adversarial example generated by PGD for the classification task.

as possible using appropriate attention mechanisms. Extensive experimental results demonstrate that our objective scoring model exhibits significant superiority on APSD and performs well on the other IQA datasets. We anticipate that this research will bring increased focus to the importance of stealthiness in adversarial attacks, ultimately driving further advancements in developing imperceptible adversarial attack techniques.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.
[2] N. Hingun, C. Sitawarin, J. Li, and D. Wagner, "Reap: A large-scale realistic adversarial patch benchmark," in *ICCV*, 2023, pp. 4640–4651.
[3] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.
[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.
[5] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *CVPR*, 2019, pp. 4312–4321.
[6] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *CVPR*, 2019, pp. 2730–2739.
[7] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *S&P*, 2017, pp. 39–57.
[8] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," *NeurIPS*, vol. 31, 2018.
[9] A. Bhattad, M. J. Chong, K. Liang, B. Li, and D. A. Forsyth, "Unrestricted adversarial examples via semantic manipulation," in *ICLR*, 2020.
[10] O. Poursaeed, I. Katsman, B. Gao, and S. J. Belongie, "Generative adversarial perturbations," in *CVPR*, 2018, pp. 4422–4431.
[11] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *ICML*, vol. 80, 2018, pp. 2142–2151.
[12] S. Sarkar, A. R. Babu, S. Mousavi, S. Ghorbanpour, V. Gundecha, A. Guillen, R. Luna, and A. Naug, "Robustness with query-efficient adversarial attack using reinforcement learning," in *CVPR*, 2023, pp. 2330–2337.
[13] B. Huang and H. Ling, "Spaa: Stealthy projector-based adversarial attacks on deep image classifiers," in *IEEE VR*, 2022, pp. 534–542.
[14] Z. Chen, C. Wang, and D. Crandall, "Semantically stealthy adversarial attacks against segmentation models," in *WACV*, 2022, pp. 4080–4089.
[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process*, vol. 13, no. 4, pp. 600–612, 2004.
[16] S. A. Fezza, Y. Bakhti, W. Hamidouche, and O. Déforges, "Perceptual evaluation of adversarial attacks for cnn-based image classification," in *QoMEX*, 2019, pp. 1–6.
[17] R. I.-R. BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2002.
[18] S. Lao, Y. Gong, S. Shi, S. Yang, T. Wu, J. Wang, W. Xia, and Y. Yang, "Attentions help cnns see better: Attention-based hybrid image quality assessment network," in *CVPR*, 2022, pp. 1140–1149.
[19] Y. Cao, Z. Wan, D. Ren, Z. Yan, and W. Zuo, "Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment," in *CVPR*, 2022, pp. 5851–5861.
[20] H. R. Sheikh, M. F. Sabir, and A. 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.
[21] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, pp. 011 006–011 006, 2010.
[22] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal Process. Image Commun.*, vol. 30, pp. 57–77, 2015.
[23] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in *ECCV*, 2020, pp. 633–651.
[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.
[25] M. Naseer, S. H. Khan, M. H. Khan, F. S. Khan, and F. Porikli, "Cross-domain transferability of adversarial perturbations," in *NeurIPS*, 2019, pp. 12 885–12 895.
[26] C. Xiao, B. Li, J. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," in *IJCAI*, 2018, pp. 3905–3911.
[27] Y. Zhang, W. Ruan, F. Wang, and X. Huang, "Generalizing universal adversarial attacks beyond additive perturbations," in *ICDM*, 2020, pp. 1412–1417.
[28] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple black-box adversarial attacks," in *ICML*, vol. 97, 2019, pp. 2484–2493.
[29] J. Pomponi, S. Scardapane, and A. Uncini, "Pixle: a fast and effective black-box attack based on rearranging pixels," in *IJCNN*, 2022, pp. 1–7.
[30] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *ECCV*, pp. 484–501.
[31] M. Yin, S. Li, C. Song, M. S. Asif, A. K. Roy-Chowdhury, and S. V. Krishnamurthy, "Adc: Adversarial attacks against object detection that evade context consistency checks," in *WACV*, 2022, pp. 3278–3287.
[32] A. Wong, M. Mundhra, and S. Soatto, "Stereopagnosia: Fooling stereo networks with adversarial perturbations," in *AAAI*, vol. 35, no. 4, 2021, pp. 2879–2888.
[33] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *NeurIPS*, vol. 36, 2024.
[34] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *ICLR*, 2017.
[35] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, 2019.
[36] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *ICLR*, 2018.
[37] Z. Zhao, Z. Liu, and M. Larson, "Towards large yet imperceptible adversarial image perturbations with perceptual color distance," in *CVPR*, 2020, pp. 1039–1048.
[38] S. Yuan, Q. Zhang, L. Gao, Y. Cheng, and J. Song, "Natural color fool: Towards boosting black-box unrestricted attacks," *NeurIPS*, vol. 35, pp. 7546–7560, 2022.
[39] S. Jia, B. Yin, T. Yao, S. Ding, C. Shen, X. Yang, and C. Ma, "Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition," *NeurIPS*, vol. 35, pp. 34 136–34 147, 2022.
[40] J. Chen, H. Chen, K. Chen, Y. Zhang, Z. Zou, and Z. Shi, "Diffusion models for imperceptible and transferable adversarial attack," *arXiv preprint arXiv:2305.08192*, 2023.
[41] H. Hosseini and R. Poovendran, "Semantic adversarial examples," in *CVPRW*, 2018, pp. 1614–1619.
[42] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Full-reference image quality expression via genetic programming," *IEEE Trans. Image Process*, vol. 32, pp. 1458–1473, 2023.
[43] S. Guo, T. Xiang, X. Li, and Y. Yang, "Peid: A perceptually encrypted image database for visual security evaluation," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1151–1163, 2019.

[44] Q. Liu, H. Yuan, R. Hamzaoui, H. Su, J. Hou, and H. Yang, "Reduced reference perceptual quality model with application to rate control for video-based point cloud compression," *IEEE Trans. Image Process*, vol. 30, pp. 6623–6636, 2021.

[45] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE Trans. Image Process*, vol. 31, pp. 4149–4161, 2022.

[46] B. Yan, B. Bare, C. Ma, K. Li, and W. Tan, "Deep objective quality assessment driven single image super-resolution," *IEEE Trans. Multimed.*, vol. 21, no. 11, pp. 2957–2971, 2019.

[47] W. Liu, F. Zhou, T. Lu, J. Duan, and G. Qiu, "Image defogging quality assessment: Real-world database and method," *IEEE Trans. Image Process*, vol. 30, pp. 176–190, 2020.

[48] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, and X. Cao, "Single image deraining: A comprehensive benchmark analysis," in *CVPR*, 2019, pp. 3838–3847.

[49] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *CVPR*, 2021, pp. 433–442.

[50] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process*, vol. 27, no. 1, pp. 206–219, 2017.

[51] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process*, vol. 23, no. 2, pp. 684–695, 2013.

[52] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Trans. Image Process*, vol. 20, no. 8, pp. 2378–2386, 2011.

[53] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process*, vol. 15, no. 2, pp. 430–444, 2006.

[54] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process*, vol. 23, no. 10, pp. 4270–4281, 2014.

[55] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen, "Frequency-driven imperceptible adversarial attack on semantic similarity," in *CVPR*, 2022, pp. 15 315–15 324.

[56] M. Zhu, T. Chen, and Z. Wang, "Sparse and imperceptible adversarial attack via a homotopy algorithm," in *ICML*, 2021, pp. 12 868–12 877.

[57] Y. Tian, J. Pan, S. Yang, X. Zhang, S. He, and Y. Jin, "Imperceptible and sparse adversarial attacks via a dual-population-based constrained evolutionary algorithm," *Trans. Artif. Intell.*, vol. 4, no. 2, pp. 268–281, 2022.

[58] Z. Chen, Z. Wang, J.-J. Huang, W. Zhao, X. Liu, and D. Guan, "Imperceptible adversarial attack via invertible neural networks," in *AAAI*, vol. 37, no. 1, 2023, pp. 414–424.

[59] A. H. Mezher, Y. Deng, and L. J. Karam, "Visual quality assessment of adversarially attacked images," in *EUVIP*, 2022, pp. 1–5.

[60] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, vol. 30, 2017.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[63] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[64] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[65] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[66] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[67] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[68] Y. Yang, T. Xiang, S. Guo, X. Lv, H. Liu, and X. Liao, "Ehnq: Subjective and objective quality evaluation of enhanced night-time images," *IEEE Trans. Circuit Syst. Video Technol.*, 2023.

[69] Q. Wu, L. Wang, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Subjective and objective de-raining quality assessment towards authentic rain image," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 30, no. 11, pp. 3883–3897, 2020.

[70] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[72] S. M. Ayyoubzadeh and A. Royat, "(ASNA) an attention-based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment," in *CVPR*, 2021, pp. 388–397.

[73] S. Shi, Q. Bai, M. Cao, W. Xia, J. Wang, Y. Chen, and Y. Yang, "Region-adaptive deformable network for image quality assessment," in *CVPR*, 2021, pp. 324–333.

[74] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *CVPR*, 2017, pp. 764–773.

[75] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *CVPR*, 2021, pp. 16 519–16 529.

[76] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019, pp. 9268–9277.

[77] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[78] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC*, vol. 2, 2003, pp. 1398–1402.

[79] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *ICASSP*, vol. 2, 2005, pp. 573–576.

[80] S. Kastryulin, J. Zakirov, D. Prokopenko, and D. V. Dylov, "Pytorch image quality: Metrics for image quality assessment," *arXiv preprint arXiv:2208.14818*, 2022.

[81] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *CVPR*, 2017, pp. 1676–1684.

[82] S. Ahn, Y. Choi, and K. Yoon, "Deep learning-based distortion sensitivity prediction for full-reference image quality assessment," in *CVPR*, 2021, pp. 344–353.

[83] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelii, "Perceptual image quality assessment using a normalized laplacian pyramid," in *HVEI*, 2016, pp. 43–48.

[84] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[85] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, 2020.

[86] J. Korhonen and J. You, "Adversarial attacks against blind image quality assessment models," in *QoEVMA*, 2022, pp. 3–11.

[87] E. Shumitskaya, A. Antsiferova, and D. Vatolin, "Fast adversarial cnn-based perturbation attack on no-reference image-and video-quality metrics," *arXiv preprint arXiv:2305.15544*, 2023.

[88] W. Zhang, D. Li, X. Min, G. Zhai, G. Guo, X. Yang, and K. Ma, "Perceptual attacks of no-reference image quality models with human-in-the-loop," *NeurIPS*, vol. 35, pp. 2916–2929, 2022.

[89] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *ICCV*, 2017, pp. 1369–1378.