LEARNING TRANSFERABLE ANTI-FORENSIC PRIORS VIA ADVERSARIAL DIFFUSION TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image Manipulation Detection and Localization (IMDL) models have achieved remarkable progress in identifying and localizing manipulated regions, yet their robustness against anti-forensic attacks remains largely unexplored. Existing adversarial approaches typically rely on victim-specific information or image-by-image optimization, limiting their scalability and practical applicability. Inspired by recent advances that repurpose pretrained diffusion models as visual priors for diverse downstream tasks, we investigate whether such priors can be adversarially activated to acquire generalizable anti-forensic capability. To this end, we propose A-PAIR (Adversarial Prior Activation and Image Refinement), a framework that learns a transferable anti-forensic prior through adversarial adaptation of pretrained diffusion representations. Rather than optimizing perturbations independently for each image, A-PAIR trains a diffusion-based attack branch in competition with a manipulation localization branch. The two branches are optimized alternately: the localization branch learns to recover manipulation masks from generated adversarial images, while the attack branch learns to generate images that induce incorrect localization. Through this alternating competition, the diffusion representation acquires a transferable anti-forensic prior that can directly generate adversarial images across different inputs and localization models. The refinement stage improves image quality while preserving attack effectiveness, and further strengthens attacks when victim queries or gradients are available. Extensive experiments across multiple IMDL benchmarks and localization models demonstrate the advantages of A-PAIR over existing anti-forensic methods.

1 INTRODUCTION

Image Manipulation Detection and Localization (IMDL) models have become an essential component of digital forensics by identifying manipulated regions at the pixel level. Recent advances in deep learning have enabled IMDL systems (Kwon et al., 2022; Dong et al., 2023; Guillaro et al., 2023; Zhu et al., 2025; Yu et al., 2026) to capture subtle manipulation traces and achieve impressive localization accuracy. However, most existing studies focus on improving the localization accuracy and generalization of IMDL models, with limited attention to their robustness against *anti-forensic attacks*. Here, anti-forensics refers to deliberately modifying manipulated images to weaken or mislead forensic analysis while preserving their visual plausibility. Existing anti-forensic research has primarily focused on evading image-level forgery detectors (Nguyen et al., 2018; Neves et al., 2020; Li et al., 2021; Xie et al., 2022; Dehdashtian et al., 2025). Unlike image-level detection, manipulation localization produces a pixel-level prediction map. Suppressing the model's response in one part of a manipulated region may still leave other parts detectable. The challenge is to weaken localization across the manipulated region while keeping the adversarial image visually close to the original.

Even when an attack suppresses one model's localization responses, another model may still detect the manipulated region. Existing optimization-based attacks search for image-specific perturbations using a source localization model's responses (Cao et al., 2024; Zeng & Pun, 2024). The search is repeated for each input, and the resulting perturbations can depend on the particular image and source model. Learned generators reduce this repeated optimization by learning an attack mapping across images (Xiao et al., 2018; Zhuo et al., 2025). However, this mapping is still shaped by the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

training images and supervisory models: sharing a generator across inputs does not by itself ensure transfer to unseen localization models.

Recent studies have shown that pretrained diffusion representations can support visual perception, correspondence, and visual in-context adaptation (Zhao et al., 2023; Xu et al., 2023; Tang et al., 2023; Wang et al., 2023). Given the effectiveness of diffusion representations in visual tasks, we investigate whether adversarial training can adapt them for anti-forensic attacks. Following the conditional image generation framework (Tan et al., 2025), we train attack and localization branches with lightweight adaptation modules on a pretrained diffusion backbone. The localization branch predicts manipulation masks, while the attack branch learns to generate images that weaken these predictions. Their alternating optimization aims to learn anti-forensic patterns that transfer across inputs and localization models.

Although the learned generator produces transferable adversarial images, some perturbations may noticeably degrade visual quality while contributing little to attack effectiveness. We therefore introduce an adaptive perturbation refinement module that aims to reduce visual distortion while retaining attack effectiveness. The refiner is trained without victim feedback for black-box use. When victim information is available, it can be adapted using limited prediction feedback in the gray-box setting or victim gradients in the white-box setting.

Experiments on five IMDL benchmarks and nine localization models show that A-PAIR achieves higher average attack effectiveness than the compared baselines across black-box, gray-box, and white-box settings. A-PAIR transfers effectively to unseen models, while adaptive refinement improves visual fidelity and further strengthens attacks when additional victim information is available. Our contributions are summarized as follows:

- We show that manipulation localization supervision can guide pretrained diffusion models to learn transferable anti-forensic priors through adversarial training.
- We propose **A-PAIR**, a two-stage framework that learns transferable anti-forensic patterns through localization–attack interaction and improves visual fidelity through a shared refinement module supporting black-box, gray-box, and white-box access.
- Experiments across multiple IMDL benchmarks and localization models demonstrate the effectiveness and transferability of A-PAIR.

2 Related Work

2.1 Image Manipulation Detection and Localization

A central goal in IMDL is to identify manipulation traces that generalize across images and editing operations. Early CNN-based methods learned these traces directly for pixel-level localization (Zhuang et al., 2021). To obtain more reliable evidence, later approaches combined image content with compression artifacts, noise patterns, and boundary cues (Kwon et al., 2022; Dong et al., 2023; Guillaro et al., 2023). Beyond extracting individual cues, localization also requires comparing manipulated regions with their surroundings. Transformer-based methods address this need by modeling relationships across image regions (Wang et al., 2022; Ma et al., 2023), while hybrid architectures combine fine-grained forensic traces with broader semantic context across scales (Zhu et al., 2025). Subtle manipulations pose a further challenge because the available evidence can be weak. Recent methods therefore seek to make such evidence more apparent through noise-guided forgery amplification (Cai et al., 2026) and reconstruction-based inconsistency analysis (Yu et al., 2026).

Alongside improvements in localization, another line of research asks how forensic predictions can be explained. FakeShield (Xu et al., 2025) and SIDA (Huang et al., 2025) use multimodal large language models to provide textual explanations together with detection and localization results. These advances improve how manipulation evidence is identified and interpreted, but raise a complementary question: whether stronger localization and generalization also translate into robustness against anti-forensic perturbations.

2.2 Adversarial Attacks on Dense Prediction

The localization component of IMDL assigns a manipulation label to each pixel. Attacking it therefore shares a central problem with attacks on semantic segmentation: how to disrupt predictions across an image rather than change a single image-level decision. Dense prediction attacks provide relevant techniques for this purpose, although the goal in IMDL is to conceal manipulation evidence. The available victim information determines whether attacks can optimize against the model directly, use output feedback, or rely on transfer.

White-box: optimization with victim gradients. FGSM (Goodfellow et al., 2015) and PGD (Madry et al., 2018) provide general gradient-based procedures that can be applied to pixel-level losses. For dense outputs, the treatment of individual pixels also matters. SegPGD (Gu et al., 2022) addresses this by dynamically weighting correctly and incorrectly classified pixels, making optimization more effective than treating all pixel losses equally.

Gray-box: adaptation through victim queries. When gradients are unavailable, prediction feedback can instead guide attack updates. We refer to this query-based setting as gray-box. P-RGF (Cheng et al., 2019) estimates attack directions using a transfer-based prior, while Square Attack (Andriushchenko et al., 2020) uses random search. For dense prediction, EBAD (Cai et al., 2023) uses victim feedback to adjust surrogate-ensemble weights, and BadPart (Cheng et al., 2024) develops query-based patch optimization for pixel-wise regression. The form of feedback further determines the attack design: DLA (Chen et al., 2024) attacks semantic segmentation using discrete predictions, while AdvFor (Mo et al., 2026) studies hard-label feedback specifically for forgery localization.

Black-box: generation without victim feedback. Without victim queries or gradients, attacks must remain effective beyond the models used to construct them. Transfer-based methods pursue this through surrogate-model optimization. TranSegPGD (Jia et al., 2023) and TASS (He et al., 2024) study transfer across segmentation models. For IMDL, AttackITL (Cao et al., 2024) constructs adversarial images on source localizers and transfers them to other models, while RevAggAL (Zeng & Pun, 2024) combines gradient aggregation with frequency constraints to improve transferability and visual fidelity. Learned generators offer an alternative to optimizing each input separately. AdvGAN (Xiao et al., 2018) learns an image-to-perturbation mapping, and SEAR (Zhuo et al., 2025) trains a forgery-concealing network through self-supervised and adversarial learning. Diffusion-based counter-forensics instead uses purification to attenuate manipulation traces (Tailanián et al., 2024).

These approaches distinguish two related goals: generating attacks that transfer without victim feedback and exploiting victim information when it is available. A-PAIR connects them by learning an anti-forensic prior through localization–attack training of a pretrained diffusion model. The learned generator supports query-free attacks, while a separate refinement stage improves visual fidelity and can incorporate victim queries or gradients for further adaptation.

3 Methodology

3.1 Overview

Pretrained diffusion models learn representations of image content and spatial structure that are useful beyond image generation, supporting perception tasks such as segmentation and visual correspondence. These capabilities motivate us to adapt diffusion representations to manipulation localization and use localization supervision to guide the learning of transferable anti-forensic patterns. Given a manipulated image x, our goal is to generate an adversarial image x^* that remains visually close to x while degrading manipulation localization.

As illustrated in Fig. 1, A-PAIR follows a two-stage framework. Stage I learns a transferable anti-forensic prior through alternating adversarial training of the attack and localization branches, while Stage II refines the generated images to improve visual fidelity while preserving attack effectiveness. Ground-truth manipulation masks are used only during training, and the framework does not require ground-truth masks at inference. The same Stage I model and refinement architecture support black-box, gray-box, and white-box access settings, with additional victim information incorporated when available.

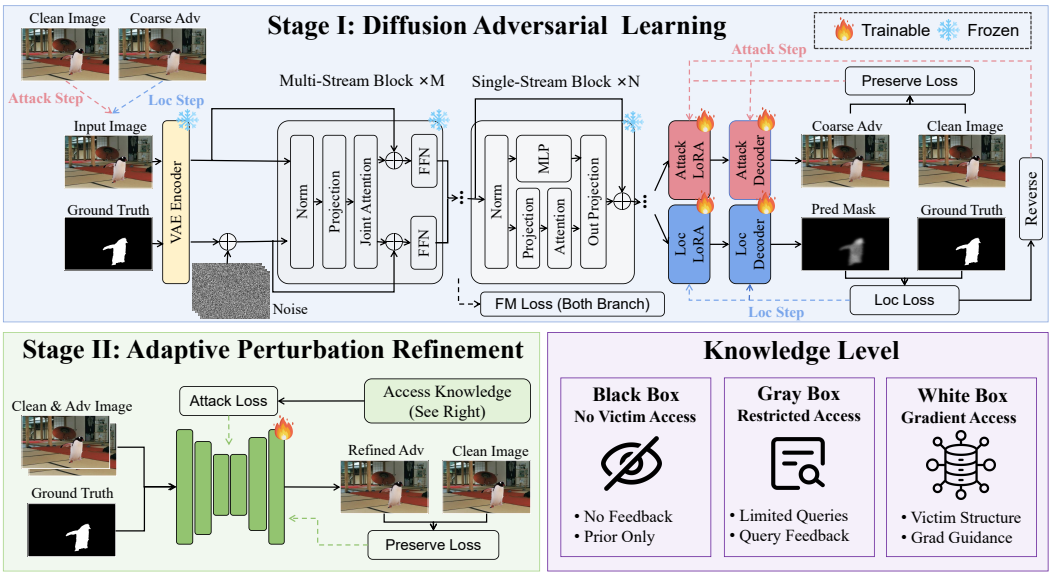


Figure 1: Overview of A-PAIR. Stage I learns a transferable anti-forensic prior for adversarial image generation through alternating adversarial training of attack and localization branches on a pretrained diffusion backbone. Stage II refines the generated images by suppressing visually redundant changes while preserving attack effectiveness. The refiner is trained without victim feedback and can be further adapted when victim information becomes available, supporting different access settings within the same architecture.

3.2 Manipulation-aware Diffusion Adversarial Learning

Motivation. We use manipulation localization to guide the adversarial adaptation of pretrained diffusion representations. The localization branch learns to recover manipulation masks, while the attack branch learns to generate images that interfere with this prediction. Alternating these two objectives encourages the diffusion model to acquire transferable anti-forensic patterns, rather than repeatedly optimizing perturbations for individual inputs.

Given an image–mask pair (x, m), we introduce an attack branch and a localization branch into a pretrained diffusion backbone Θ. Each branch uses its own LoRA parameters θ_b and a task-specific decoder D_b with parameters ϕ_b, where $b \in \{a, l\}$. The two branches share the same flow-matching parameterization, but use different diffusion targets and task objectives.

For each branch, we randomly sample a timestep t according to the timestep distribution of the pretrained diffusion model and obtain the corresponding noise level σ_t. Let $z_{\text{tar}}^{(b)}$ denote the latent representation of the branch-specific diffusion target. Given Gaussian noise $\epsilon_b \sim \mathcal{N}(0, I)$, the noisy training state and its target velocity are constructed as

$$z_t^{(b)} = \sigma_t \epsilon_b + (1 - \sigma_t) z_{\text{tar}}^{(b)}, \qquad v_t^{(b)} = \epsilon_b - z_{\text{tar}}^{(b)}. \tag{1}$$

The adapted diffusion branch predicts the corresponding velocity

$$\hat{v}_t^{(b)} = v_{\Theta,\theta_b}\left(z_t^{(b)}, t, c_b\right), \tag{2}$$

which is supervised by the branch-specific flow-matching objective

$$\mathcal{L}_{\text{FM}}^{(b)} = \mathbb{E}_{t,\epsilon_b}\left[\left\|\hat{v}_t^{(b)} - v_t^{(b)}\right\|_2^2\right], \qquad b \in \{a, l\}. \tag{3}$$

To apply task-specific supervision directly in the decoded output space, we reconstruct the target endpoint from the same velocity prediction, without unrolling the sampled state through the complete denoising trajectory:

$$\hat{z}_0^{(b)} = z_t^{(b)} - \sigma_t \hat{v}_t^{(b)}. \tag{4}$$

The predicted velocity is used both for flow-matching supervision and to estimate the clean latent $\hat{z}_0^{(b)}$. The task-specific decoder D_b then maps this latent to an adversarial image or a localization mask, on which the corresponding task loss is computed.

As illustrated in Fig. 1, the attack and localization branches are updated alternately with different diffusion targets. During an attack step, the input image x and its ground-truth mask m form the condition $c_a = (x, m)$. The diffusion target $\tilde{x}_{\text{adv}}$ is the image generated by the attack branch in the preceding attack step; for the first attack step, it is initialized to x. Conditioned on c_a, the attack branch predicts the clean latent $\hat{z}_0$, which is then decoded to obtain the adversarial image:

$$x_{\text{adv}} = D_a\left(\hat{z}_0^{(a)}; \phi_a\right). \tag{5}$$

During the subsequent localization step, the generated adversarial image x_{adv} serves as the image condition, and the ground-truth mask m serves as the diffusion target. With $c_l = (x_{\text{adv}}, m)$, the localization branch predicts:

$$\hat{m} = D_l\left(\hat{z}_0^{(l)}; \phi_l\right). \tag{6}$$

The localization objective encourages the decoded mask to match the ground truth:

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{BCE}}(\hat{m}, m) + \lambda_{\text{dice}}\mathcal{L}_{\text{Dice}}(\hat{m}, m) + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}}^{(l)}, \tag{7}$$

where the Dice Loss (Sudre et al., 2017) alleviates foreground imbalance. The attack objective instead encourages incorrect localization while constraining visual changes:

$$\mathcal{L}_{\text{adv}} = -\mathcal{L}_{\text{BCE}}(\hat{m}, m) + \lambda_w\mathcal{L}_{\text{wavelet}}^{\text{LF}}(x, x_{\text{adv}}) + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}}^{(a)}. \tag{8}$$

The negative BCE term promotes incorrect localization, while the wavelet term preserves visual fidelity by constraining differences between the low-frequency components of the adversarial and original images.

Training alternates three attack updates with one localization update. During attack updates, the localization parameters are frozen, while the localization output remains differentiable with respect to x_{adv} through its conditioning pathway. This allows the reversed localization objective to propagate into the attack branch. During localization updates, x_{adv} is detached so that only the localization LoRA and decoder are optimized. The pretrained diffusion backbone remains frozen throughout training.

At inference, we retain the standard iterative Euler sampler rather than the direct endpoint reconstruction used for training-time task supervision. Starting from Gaussian noise, the scheduler integrates the learned velocity field over a fixed trajectory to generate the coarse adversarial image and localization prediction. Flow matching therefore learns the velocity field from randomly sampled trajectory states during training, while inference integrates the same learned field over the complete sampling schedule.

3.3 Adaptive Perturbation Refinement

Motivation. Stage I learns a transferable anti-forensic prior and produces coarse adversarial images with strong attack effectiveness. However, these images may also contain visual changes that contribute little to suppressing localization responses. Stage II therefore refines the coarse adversarial image rather than learning a new attack from scratch. It suppresses visually redundant changes while preserving the attack-relevant patterns inherited from Stage I. Supporting analyses are provided in Appendix A.2.1.

As illustrated in Fig. 1, given the original image x, the coarse adversarial image x_{adv}, and a spatial guidance map $\tilde{m}$, the refinement module produces

$$x^* = R_\psi(\text{Concat}(x, x_{\text{adv}}, \tilde{m})), \tag{9}$$

where $\tilde{m} = m$ during Stage II training and $\tilde{m} = \hat{m}_1$ during inference and victim-specific adaptation. The original image provides the visual reference, while the spatial guidance helps the refiner focus on manipulation-relevant regions. The detailed network architecture is provided in Appendix A.2.2.

The refinement module is optimized to improve visual fidelity while preserving attack effectiveness. We define the preservation objective as

$$\mathcal{L}_{\text{pres}} = \mathcal{L}_{\text{MSE}}(x, x^*) + \mathcal{L}_{\text{SSIM}}(x, x^*), \tag{10}$$

where the MSE term constrains pixel-level deviations and the SSIM term preserves structural consistency.

Victim-independent refinement training. Stage II is first trained without accessing any victim model. The frozen internal localization branch provides attack supervision, while the ground-truth mask m is used as spatial guidance. The training objective is

$$\mathcal{L}_{\text{ref}}^{\text{bb}} = \mathcal{L}_{\text{pres}} + \lambda_{\text{atk}} \mathcal{L}_{\text{atk}}^{\text{bb}}, \tag{11}$$

where $\mathcal{L}_{\text{atk}}^{\text{bb}}$ is the response-margin objective computed using the frozen internal localization branch. This training produces the victim-independent refinement parameters ψ_{bb}^*, which are directly used under black-box access.

Victim-specific adaptation. Gray-box and white-box settings start from the same ψ_{bb}^* and perform per-image test-time adaptation. Ground-truth masks are unavailable at this stage, so the Stage I localization prediction $\hat{m}_1$ is used as spatial guidance. Under gray-box access, victim queries are used to estimate the image-space gradient of the erasure objective with NES, and the resulting gradient surrogate updates only the refinement parameters. Under white-box access, victim gradients are directly available, and the refiner is optimized using the preservation objective together with the victim-based response margin and erasure loss:

$$\mathcal{L}_{\text{ref}}^{\text{wb}} = \mathcal{L}_{\text{pres}} + \lambda_{\text{atk}} \mathcal{L}_{\text{atk}}^{\text{wb}} + \mathcal{L}_{\text{erase}}^{f_v}(x^*; \hat{m}_1). \tag{12}$$

Only the refinement parameters are updated during victim-specific adaptation; Stage I and all supervising localization models remain frozen. Detailed gray-box and white-box objectives are provided in Appendix A.2.3.

4 Experiments

4.1 Experimental Setup

Datasets and Victim Models. We evaluate A-PAIR on CASIAv1 (Dong et al., 2013), COVERAGE (Wen et al., 2016), Columbia (Hsu & Chang, 2006), NIST16 (Guan et al., 2019), and IMD2020 (Novozámský et al., 2020). These benchmarks cover diverse manipulation scenarios. For each victim model, the values reported in the main comparison tables are the unweighted arithmetic means of the dataset-level RelDrop scores over these five benchmarks.

Following previous IMDL studies, we select nine representative localization models as victim models, including CAT-Net (Kwon et al., 2022), MVSS-Net (Dong et al., 2023), ObjectFormer (Wang et al., 2022), PSCC-Net (Liu et al., 2022), TruFor (Guillaro et al., 2023), IML-ViT (Ma et al., 2023), Mesorch (Zhu et al., 2025), NFA-ViT (Cai et al., 2026), and EARG-Net (Yu et al., 2026). These models cover a broad range of representative IMDL architectures, enabling comprehensive evaluation of the transferability and adaptability of our approach.

Evaluation Metrics. We measure attack effectiveness by the relative drop in pixel-level F1: $\text{RelDrop} = (F1_{\text{clean}} - F1_{\text{adv}})/F1_{\text{clean}}$, where the two scores correspond to the original and adversarial images. We measure image fidelity using PSNR between these images. Higher values indicate stronger attacks and better fidelity, respectively.

Implementation Details. We implement A-PAIR using the pretrained FLUX-1.dev backbone (Labs, 2024) and train it on CASIAv2 (Dong et al., 2013). During Stage I training, the diffusion backbone is kept frozen and the attack and localization branches are alternately optimized with a 3:1 update ratio for 6,000 iterations. Flow-matching supervision is performed with randomly sampled timesteps, while inference uses a fixed Euler sampling trajectory to generate branch outputs. For victim-specific adaptation, we use 20 victim-model queries per image under gray-box access and 10 gradient updates per image under white-box access. Additional training and implementation details are provided in the appendix.

4.2 White-box Optimization with Full Access

We first evaluate our framework under the white-box setting, where the victim localization models and their gradients are fully accessible. We compare A-PAIR with representative gradient-based attacks, including FGSM (Goodfellow et al., 2015), MI-FGSM (Dong et al., 2018), TI-FGSM (Dong et al., 2019), NI-FGSM (Lin et al., 2019), PGD (Madry et al., 2018), SegPGD (Gu et al., 2022), and TranSegPGD (Jia et al., 2023). All baseline attacks use an ℓ_∞ perturbation budget of $16/255$. The iterative FGSM variants run for 10 iterations, while PGD variants run for 50 iterations. All iterative attacks use a step size of $2/255$.

Table 1: White-box anti-forensic performance comparison with gradient-based adversarial attacks. Results are reported as RelDrop (%). The best results for each victim model are highlighted in **bold**.

Victim	FGSM-based				PGD-based			Ours
	FGSM	MI-FGSM	TI-FGSM	NI-FGSM	PGD	SegPGD	TranSegPGD	
CAT-Net	61.0	96.5	**99.6**	96.0	65.7	68.1	68.8	99.5
TruFor	15.1	59.8	67.3	58.5	73.0	83.0	88.4	**93.4**
IML-ViT	73.2	98.9	94.1	98.9	**99.8**	99.5	99.4	98.8
MVSS-Net	85.7	82.4	82.7	82.5	40.2	40.2	54.9	**95.5**
ObjectFormer	47.1	**99.7**	93.1	99.6	99.5	98.6	99.1	99.4
PSCC-Net	56.0	**100.0**	99.9	**100.0**	99.0	97.4	97.3	98.9
Mesorch	71.7	96.6	99.6	**99.7**	92.3	93.0	90.0	99.2
NFA-ViT	91.7	93.3	**99.7**	**99.7**	93.1	92.2	90.2	99.5
EARG-Net	97.7	98.5	**100.0**	**100.0**	99.5	99.7	99.7	**100.0**
Mean	66.6	91.7	92.9	92.8	84.7	85.7	87.5	**98.2**

As shown in Table 1, A-PAIR achieves strong and consistent anti-forensic performance across different victim localization models. The results suggest that the learned anti-forensic prior provides an effective initialization for gradient-guided adaptation, allowing the refinement stage to further specialize the transferable prior toward target localization models under full model access.

4.3 Gray-box Adaptation with Limited Feedback

We evaluate A-PAIR under gray-box access and compare it with P-RGF (Cheng et al., 2019), Square Attack (Andriushchenko et al., 2020), EBAD (Cai et al., 2023), BadPart (Cheng et al., 2024), DLA (Chen et al., 2024), and AdvFor (Mo et al., 2026). Each baseline is evaluated using the query budget specified in its original paper. For A-PAIR, we restrict adaptation to 20 victim-model queries per image to reflect practical scenarios in which access to deployed models is tightly controlled.

Table 2: Gray-box anti-forensic comparison with query-based baselines. Results are reported as RelDrop (%). The best results for each victim model are highlighted in **bold**.

Victim	P-RGF	Square	EBAD	BadPart	DLA	AdvFor	Ours
CAT-Net	85.2	76.1	80.5	42.0	46.5	38.3	**88.7**
TruFor	45.9	53.0	45.1	28.5	30.7	31.0	**81.7**
IML-ViT	65.8	35.8	24.5	35.0	45.9	12.3	**77.9**
MVSS-Net	80.4	34.7	33.0	21.3	45.3	15.6	**83.4**
ObjectFormer	65.7	67.6	56.1	26.1	**91.3**	63.2	78.3
PSCC-Net	81.6	49.8	72.2	22.8	72.7	47.2	**82.4**
Mesorch	47.5	13.4	27.9	23.9	33.4	72.7	**75.4**
NFA-ViT	89.4	47.4	41.9	27.6	80.8	93.2	**95.7**
EARG-Net	19.8	52.5	**86.9**	47.7	47.4	31.7	77.5
Mean	64.6	47.8	52.0	30.5	54.9	45.0	**82.3**

As shown in Table 2, A-PAIR achieves strong attack effectiveness across the evaluated localization models with limited victim feedback. These results suggest that the learned anti-forensic prior provides an effective initialization for target-specific adaptation. Rather than constructing perturbations from scratch during victim interaction, the refinement network uses the available queries to adapt the generated adversarial images to the target model.

4.4 Black-box Anti-forensic Generation

We compare A-PAIR with representative black-box baselines from IMDL and related adversarial attack research, including AttackITL (Cao et al., 2024), RevAggAL (Zeng & Pun, 2024), TASS (He et al., 2024), SEAR (Zhuo et al., 2025), AdvGAN (Xiao et al., 2018), and Diff-CFG (Tailanián et al., 2024). Since CAT-Net is used as the surrogate model for the transfer-based baselines, we exclude CAT-Net from the cross-model transfer evaluation to avoid reporting source-model performance as transferability. Under the zero-query protocol, all resulting adversarial images are evaluated on unseen victim models without target-specific queries or adaptation.

Table 3: Black-box anti-forensic comparison with representative baselines. Results are reported as RelDrop (%). The best results for each victim model are highlighted in **bold**.

Victim	AttackITL	TASS	RevAggAL	AdvGAN	Diff-CFG	SEAR	Ours
TruFor	42.2	53.8	36.2	47.3	18.2	17.2	**67.0**
IML-ViT	44.5	32.8	41.9	67.3	10.1	31.1	**68.0**
MVSS-Net	51.5	23.1	47.5	48.3	10.7	29.7	**72.7**
ObjectFormer	24.1	25.9	22.1	6.7	2.4	6.1	**63.3**
PSCC-Net	36.3	19.1	33.7	35.7	17.9	8.7	**83.0**
Mesorch	20.7	27.8	19.5	29.3	12.3	12.2	**63.0**
NFA-ViT	50.9	39.2	57.7	30.5	8.1	29.5	**90.4**
EARG-Net	25.7	31.2	18.6	14.0	1.4	9.8	**75.3**
Mean	37.0	31.6	34.7	34.9	10.1	18.0	**72.8**

As shown in Table 3, A-PAIR achieves consistently stronger attack effectiveness across heterogeneous localization architectures. In particular, its advantage is maintained across models with different forensic representations and network designs, rather than being concentrated on a small subset of victims. These results indicate that the anti-forensic prior learned in Stage I, together with victim-independent refinement in Stage II, generalizes beyond the models involved during training and provides effective cross-model transfer under black-box access.



Figure 2: Qualitative results under the black-box setting. The first three columns show the input image, ground-truth manipulation mask, and adversarial image, while the remaining columns show localization responses from different victim models.

Figure 2 further provides qualitative evidence for this transferability. The generated adversarial images remain visually consistent with their corresponding inputs, whereas the localization responses become fragmented, weakened, displaced, or spatially diffuse, and in several cases nearly vanish.

Importantly, such failure patterns are observed across diverse victim architectures without using any victim-specific feedback. Overall, the quantitative and qualitative results demonstrate that A-PAIR effectively suppresses manipulation localization even under the most restrictive attacker-access condition.

4.5 Ablation Study

We ablate the two key designs of A-PAIR: the localization–attack interaction in manipulation-aware diffusion adversarial learning and the adaptive perturbation refinement. RelDrop and PSNR are used to evaluate attack effectiveness and visual quality, respectively. The results are summarized in Table 4.

Table 4: Ablation studies of A-PAIR. RelDrop is reported in % and PSNR in dB. The random-noise baseline replaces Stage I with bounded random perturbations and retains Stage II.

(a) Localization–attack interaction

Configuration	RelDrop↑	PSNR↑
Random noise	19.6	**34.5**
Pretrained diffusion	22.4	26.0
Attack adaptation only	45.4	24.0
Full interaction	**73.9**	32.4

(b) Adaptive perturbation refinement

Configuration	RelDrop↑	PSNR↑
Without refinement	68.2	28.4
$\mathcal{L}_{\mathrm{atk}}$ only	70.4	28.8
$\mathcal{L}_{\mathrm{pres}}$ only	69.7	30.3
Full refinement	**73.9**	**32.4**

Localization–attack interaction. We compare four configurations: random-noise initialization, pretrained diffusion without attack-specific fine-tuning, attack-only fine-tuning, and full localization–attack training. For the random-noise baseline, we replace only the coarse adversarial image fed to Stage II with a randomly perturbed version of the input image, while keeping the localization mask from the full localization–attack configuration unchanged. As shown in Table 4(a), random-noise initialization preserves image fidelity but yields limited attack effectiveness. Pretrained diffusion is also insufficient without attack adaptation. Attack-only fine-tuning strengthens the attack at the cost of image fidelity, whereas full localization–attack training improves both metrics over attack-only fine-tuning. These results support the contribution of localization–attack interaction to learning effective anti-forensic perturbations while maintaining image fidelity.

Adaptive perturbation refinement. We next examine the refinement stage and its two objectives, while keeping the full localization–attack configuration in Stage I unchanged. As shown in Table 4(b), using the coarse adversarial image directly yields weaker attacks and lower visual quality than the refined variants. Optimizing $\mathcal{L}_{\mathrm{atk}}$ alone improves attack effectiveness, but provides only a modest gain in visual fidelity. In contrast, $\mathcal{L}_{\mathrm{pres}}$ alone yields a larger improvement in image quality, although its attack effectiveness remains below that of attack-only refinement. Combining the two objectives achieves the best performance on both metrics, improving upon the coarse output and either single-objective variant. These trends support their complementary roles: $\mathcal{L}_{\mathrm{atk}}$ encourages the refined image to continue disrupting localization, while $\mathcal{L}_{\mathrm{pres}}$ constrains its deviation from the original image.

5 Conclusion

We presented A-PAIR, a two-stage framework that learns transferable anti-forensic priors by adapting pretrained diffusion models. Alternating localization–attack training guides adversarial generation, while adaptive refinement improves visual fidelity and incorporates additional victim information when available. Experiments across multiple IMDL benchmarks demonstrate strong transferability to unseen localization models and higher average attack effectiveness than the evaluated baselines across different access settings. These findings suggest that a learned anti-forensic prior can support attacks without victim feedback while providing an effective starting point for query-based or gradient-based adaptation. Defenses against transferable attacks and extensions to other dense prediction tasks remain important directions for future work.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, literature retrieval, and research ideation and execution. For manuscript preparation, these tools supported organization, drafting, language polishing, translation of technical descriptions, and the presentation of mathematical formulations and algorithms. They also assisted with identifying and summarizing related literature, discussing methodological choices and experimental designs, interpreting results, and preparing scientific figures. We additionally used AI to assist in writing code for this research.

The authors reviewed and revised the AI-assisted content, checked cited sources against the original publications, and checked the methodological descriptions and reported results against the implementation and experimental records. The authors take full responsibility for the final manuscript, code, experimental results, and conclusions.

ETHICS STATEMENT

This work investigates vulnerabilities in image manipulation localization to support robustness evaluation and the development of more reliable forensic systems. The proposed method has dual-use implications: it could reduce the effort required to conceal manipulation traces and undermine the reliability of image-based evidence, including in journalism and other settings that depend on trustworthy visual information.

Our evaluation uses public benchmark datasets and existing localization models. It does not involve recruiting human participants or collecting private user data. The reported results concern the evaluated models and access settings; successful evasion of a localization model does not establish image authenticity or imply that an image can evade all forensic analyses.

The accompanying code documentation will describe the intended use for authorized robustness evaluation, the assumed access conditions, known limitations, and foreseeable misuse risks. It will also provide guidance for reproducing the benchmark experiments. Potential defensive responses include adversarial evaluation and the use of complementary forensic evidence, but their effectiveness against the proposed method has not been established by this study. Documentation and intended-use guidance alone cannot prevent deliberate misuse.

REPRODUCIBILITY STATEMENT

The main text describes the proposed framework, training objectives, attacker access settings, and evaluation protocol. The appendix provides the detailed algorithms, implementation details, and complete hyperparameter settings. We also submit anonymized source code as supplementary material to support reproduction of the experiments. Together, the manuscript, appendix, and supplementary code document the method and its experimental evaluation.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, 2020.

Lvpan Cai, Haowei Wang, Jiayi Ji, YanShu ZhouMen, Shen Chen, Taiping Yao, and Xiaoshuai Sun. Zooming in on fakes: A novel dataset for localized ai-generated image detection with forgery amplification approach. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, 2026.

Zikui Cai, Yaoteng Tan, and M. Salman Asif. Ensemble-based blackbox attacks on dense prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023.

Gang Cao, Yuqi Wang, Haochen Zhu, Zijie Lou, and Lifang Yu. Transferable adversarial attack on image tampering localization. *J. Vis. Commun. Image Represent.*, 2024.

Zhaoyu Chen, Zhengyang Shan, Jingwen Chang, Kaixun Jiang, Dingkang Yang, Yiting Cheng, and Wenqiang Zhang. Delving into decision-based black-box attacks on semantic segmentation. *CoRR*, 2024.

Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Zhiyuan Cheng, Zhaoyi Liu, Tengda Guo, Shiwei Feng, Dongfang Liu, Mingjie Tang, and Xiangyu Zhang. Badpart: Unified black-box adversarial patch attacks against pixel-wise regression tasks. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Sepehr Dehdashtian, Mashrur Mahmud Morshed, Jacob H. Seidman, Gaurav Bharaj, and Vishnu Boddeti. Polyjuice makes it real: Black-box, universal red teaming for synthetic image detectors. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP 2013, Beijing, China, July 6-10, 2013*, 2013.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip H. S. Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIX*, 2022.

Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N. Yates, Andrew Delgado, Daniel Zhou, Timothée Kheyrkhah, Jeff Smith, and Jonathan G. Fiscus. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, 2019.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023.

Mengqi He, Jing Zhang, and Xin Yu. Transferable attacks for semantic segmentation. In *Databases Theory and Applications - 35th Australasian Database Conference, ADC 2024, Gold Coast, QLD, Australia, December 16-18, 2024, Proceedings*, 2024.

Yu-Feng Hsu and Shih-Fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006, July 9-12 2006, Toronto, Ontario, Canada*, 2006.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: social media image deepfake detection, localization and explanation with large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 2025.

Xiaojun Jia, Jindong Gu, Yihao Huang, Simeng Qin, Qing Guo, Yang Liu, and Xiaochun Cao. Transegpgd: Improving transferability of adversarial examples on semantic segmentation. *CoRR*, 2023.

Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 2022.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. Exploring adversarial fake images on face manifold. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2019.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Trans. Circuits Syst. Video Technol.*, 2022.

Xiaochen Ma, Bo Du, Zhuohang Jiang, Ahmed Y. Al Hammadi, and Jizhe Zhou. Iml-vit: Benchmarking image manipulation localization by vision transformer, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Xianbo Mo, Shunquan Tan, Rongxuan Peng, Bin Li, and Jiwu Huang. Query-efficient hard-label attacks against black-box image forgery localization model via reinforcement learning. *IEEE Trans. Inf. Forensics Secur.*, 2026.

João C. Neves, Ruben Tolosana, Rubén Vera-Rodríguez, Vasco Lopes, Hugo Proença, and Julian Fiérrez. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE J. Sel. Top. Signal Process.*, 2020.

Huy H. Nguyen, Ngoc-Dung T. Tieu, Hoang-Quoc Nguyen-Son, Junichi Yamagishi, and Isao Echizen. Transformation on computer-generated facial image to avoid detection by spoofing detector. In *2018 IEEE International Conference on Multimedia and Expo, ICME 2018, San Diego, CA, USA, July 23-27, 2018*, 2018.

Adam Novozámský, Babak Mahdian, and Stanislav Saic. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2020, Snowmass Village, CO, USA, March 1-5, 2020*, 2020.

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support - Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings*, 2017.

Matías Tailanián, Marina Gardella, Álvaro Pardo, and Pablo Musé. Diffusion models meet image counter-forensics. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022.

Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang (Atlas) Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. COVERAGE - A novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, 2016.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 2018.

Hao Xie, Jiangqun Ni, Jian Zhang, Weizhe Zhang, and Jiwu Huang. Evading generated-image detectors: A deep dithering approach. *Signal Process.*, 2022.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Yanpu Yu, Zhaoxin Shi, Hanqing Zhao, Tianyi Wei, Wenbo Zhou, and Nenghai Yu. Earg-net: Edge-aware reconstruction-guided network for image manipulation detection and localization. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, 2026.

Yijia Zeng and Chi-Man Pun. Frequency-constrained transferable adversarial attack on image manipulation detection and localization. *Vis. Comput.*, 2024.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Thirty-Ninth AAAI Conference on Artificial*

Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025, 2025.

Peiyu Zhuang, Haodong Li, Shunquan Tan, Bin Li, and Jiwu Huang. Image tampering localization using a dense fully convolutional network. *IEEE Trans. Inf. Forensics Secur.*, 2021.

Long Zhuo, Shenghai Luo, Shunquan Tan, Han Chen, Bin Li, and Jiwu Huang. Evading detection actively: Toward anti-forensics against forgery localization. *IEEE Trans. Dependable Secur. Comput.*, 2025.

A APPENDIX

A.1 ALGORITHMIC DESCRIPTION OF A-PAIR

Algorithm 1 Stage I: Adversarial Diffusion Learning

Require: Training set $\mathcal{D}$; frozen diffusion backbone Θ
Require: LoRA parameters θ_a, θ_l; decoders D_a, D_l with parameters ϕ_a, ϕ_l
Require: Attack/localization update ratio $\rho = 3$
Ensure: Trained parameters $(\theta_a, \phi_a, \theta_l, \phi_l)$

Freeze Θ
for each training iteration **do**
  Sample $(x, m) \sim \mathcal{D}$
  Initialize attack target $\tilde{x}_{\mathrm{adv}} \leftarrow x$
  Attack updates
  Enable (θ_a, ϕ_a); freeze (θ_l, ϕ_l)
  for $j = 1, \ldots, \rho$ **do**
    Sample timesteps t_a, t_l and noise ϵ_a, ϵ_l
    Construct $(z_{t_a}^{(a)}, v_{t_a}^{(a)})$ from target $\tilde{x}_{\mathrm{adv}}$ using Eq. 1
    $\hat{v}_{t_a}^{(a)} \leftarrow v_{\Theta,\theta_a}(z_{t_a}^{(a)}, t_a; x, m)$
    $\hat{z}_0^{(a)} \leftarrow z_{t_a}^{(a)} - \sigma_{t_a} \hat{v}_{t_a}^{(a)}$
    $x_{\mathrm{adv}} \leftarrow D_a(\hat{z}_0^{(a)}; \phi_a)$
    Construct $(z_{t_l}^{(l)}, v_{t_l}^{(l)})$ from target m using Eq. 1
    $\hat{v}_{t_l}^{(l)} \leftarrow v_{\Theta,\theta_l}(z_{t_l}^{(l)}, t_l; x_{\mathrm{adv}}, m)$
    $\hat{z}_0^{(l)} \leftarrow z_{t_l}^{(l)} - \sigma_{t_l} \hat{v}_{t_l}^{(l)}$
    $\hat{m} \leftarrow D_l(\hat{z}_0^{(l)}; \phi_l)$
    Compute $\mathcal{J}_a \leftarrow \mathcal{L}_{\mathrm{adv}}$
    Update (θ_a, ϕ_a) using AdamW
    $\tilde{x}_{\mathrm{adv}} \leftarrow \mathrm{sg}(x_{\mathrm{adv}})$
  end for
  Localization update
  Enable (θ_l, ϕ_l); freeze (θ_a, ϕ_a)
  Obtain a fixed x_{adv} from the attack branch and detach it
  Sample timestep t_l and noise ϵ_l
  Construct $(z_{t_l}^{(l)}, v_{t_l}^{(l)})$ from target m using Eq. 1
  $\hat{v}_{t_l}^{(l)} \leftarrow v_{\Theta,\theta_l}(z_{t_l}^{(l)}, t_l; x_{\mathrm{adv}}, m)$
  $\hat{z}_0^{(l)} \leftarrow z_{t_l}^{(l)} - \sigma_{t_l} \hat{v}_{t_l}^{(l)}$
  $\hat{m} \leftarrow D_l(\hat{z}_0^{(l)}; \phi_l)$
  Compute $\mathcal{J}_l \leftarrow \mathcal{L}_{\mathrm{loc}}$
  Update (θ_l, ϕ_l) using AdamW
end for
return $(\theta_a, \phi_a, \theta_l, \phi_l)$

In Stage I training, each branch samples a timestep and constructs the corresponding noisy state from its branch-specific diffusion target. The predicted velocity is supervised by the flow-matching objective and simultaneously used to reconstruct the clean endpoint $\hat{z}_0$, from which the task decoder produces the attack image or localization mask. Thus, Stage I training does not require unrolling the full Euler sampling trajectory for task-specific losses. During attack updates, the localization parameters are frozen, but localization branch's conditioning path remains differentiable with respect to x_{adv} through the conditioning pathway, allowing the negated localization loss to update the attack branch. During the localization update, x_{adv} is detached so that only the localization LoRA and decoder are optimized. At inference, no ground-truth manipulation mask is available: the attack branch starts from Gaussian noise and performs K-step Euler sampling conditioned only on the input image x. The resulting coarse adversarial image x_1 then serves as the sole image condition for the localization branch, which independently starts from Gaussian noise and performs the same

iterative sampling procedure to obtain $\hat{m}_1$. The predicted mask is subsequently used as the spatial guidance for Stage II.

Algorithm 2 Stage II: Refinement Training and Test-Time Adaptation

Require: Training set $\mathcal{D}$; Stage I model and probe f_{p}; refiner R_ψ; frozen victim f_v; budgets Q, T_g
Ensure: Victim-independent refiner ψ^*_{bb}; adapted refiner ψ

for refinement iteration **do**
  Sample $(x, m) \sim \mathcal{D}$; obtain Stage I coarse image x_1
  $x_2 \leftarrow R_\psi(\mathrm{Concat}(x, x_1, m))$
  Update ψ with $\mathcal{J}_{\mathrm{bb}}$
end for
$\psi^*_{\mathrm{bb}} \leftarrow \psi$
procedure ADAPTREFINER($x, x_1, \hat{m}_1, f_v, s$)
  if $s = \mathrm{BlackBox}$ **then**
    return ψ^*_{bb}
  end if
  $\psi \leftarrow \psi^*_{\mathrm{bb}}$; initialize optimizer for ψ
  if $s = \mathrm{GrayBox}$ **then**
    $q \leftarrow 0$
    while $q < Q$ **do**
      $x_2 \leftarrow R_\psi(\mathrm{Concat}(x, x_1, \hat{m}_1))$
      $(\hat{g}, q_t) \leftarrow \mathrm{NESGrad}(\mathcal{L}^{f_v}_{\mathrm{erase}}, x_2, Q - q)$
      if $q_t = 0$ **then**
        break
      end if
      Update ψ with $\mathcal{J}_{\mathrm{gb}}$
      $q \leftarrow q + q_t$
    end while
  else if $s = \mathrm{WhiteBox}$ **then**
    for $t = 1, \ldots, T_g$ **do**
      $x_2 \leftarrow R_\psi(\mathrm{Concat}(x, x_1, \hat{m}_1))$
      Update ψ with $\mathcal{J}_{\mathrm{wb}}$
    end for
  end if
  return ψ
end procedure

Stage II is first trained offline without accessing any victim model, yielding a victim-independent refinement initializer ψ^*_{bb}. This initializer is used directly under black-box access and serves as the common starting point for victim-specific test-time adaptation. For each test image, gray-box adaptation updates only the refinement parameters using NES-estimated image-space gradients obtained from victim queries, whereas white-box adaptation uses direct victim gradients.

Algorithm 3 Inference of A-PAIR under Different Access Settings

Require: Input image x; trained Stage I attack/localization branches; victim-independent refiner ψ^*_{bb}
Require: Access setting s; victim model f_v when available
Ensure: Final adversarial image x^*

Sample Gaussian noise ϵ_a
$z^a_0 \leftarrow \mathrm{EulerSample}_K(\epsilon_a; x)$
$x_1 \leftarrow D_a(z^a_0; \phi_a)$
Sample Gaussian noise ϵ_l
$z^l_0 \leftarrow \mathrm{EulerSample}_K(\epsilon_l; x_1)$
$\hat{m}_1 \leftarrow D_l(z^l_0; \phi_l)$
$\psi_s \leftarrow \mathrm{ADAPTREFINER}(x, x_1, \hat{m}_1, f_v, s)$
$x^* \leftarrow R_{\psi_s}(\mathrm{Concat}(x, x_1, \hat{m}_1))$
return x^*

Inference is fully self-contained and does not require the ground-truth manipulation mask. Stage I first generates the coarse adversarial image and then predicts its localization mask, which is used as the spatial guidance for Stage II and subsequent victim-specific adaptation.

A.2 Details of Perturbation Refinement

This appendix provides further details of the Stage II refinement module, including its motivation, network architecture, optimization objectives, and adaptation under different victim-access settings. Stage I already produces effective coarse adversarial images, but some introduced changes may contribute little to localization suppression while degrading visual quality. Stage II therefore learns an image refinement mapping that removes visually unnecessary variations and preserves attack-relevant patterns instead of generating a new perturbation from scratch.

A.2.1 Motivation

Stage I produces effective adversarial images, but the generated perturbations can also degrade visual fidelity. We therefore examine whether this distortion can be reduced by directly constraining the perturbations, and how doing so affects attack effectiveness. Specifically, we consider three fixed strategies: restricting perturbations to manipulated regions, limiting their magnitude, and concentrating them near manipulation boundaries.

Table 5: Effect of constraining the Stage I perturbation to different structured perturbation spaces on CASIAv1. RelDrop denotes the average relative F1-score drop over nine victim models.

Variant	PSNR ↑	LPIPS ↓	RelDrop ↑
Original Stage I output	25.26	0.276	82.4
Manipulation-region focused	28.96	0.067	67.4
Magnitude constrained	27.09	0.083	38.5
Boundary focused	30.23	0.063	34.4
Stage II Refinement	**34.36**	**0.050**	**88.3**

As shown in Table 5, constraining the Stage I perturbation to different structured spaces substantially improves visual fidelity, while some attack effectiveness is preserved. In particular, restricting the perturbation to manipulation-related regions yields a favorable trade-off, considerably improving PSNR and LPIPS while retaining strong attack performance. These results indicate that the full perturbation produced by Stage I is not necessary for maintaining the attack, and that effective anti-forensic signals can remain within a more restricted perturbation space.

This observation motivates Stage II to learn such an attack-preserving refinement automatically. Rather than retaining the entire Stage I residual, the refinement module learns to suppress changes that contribute little to the attack while preserving those required for effective localization suppression. In this way, Stage II seeks a more favorable balance between attack effectiveness and visual fidelity than the fixed constraints considered above.

A.2.2 Multi-Scale Image Refinement

Figure 3 illustrates the architecture and data flow of the Stage II refinement module. Stage I provides a coarse adversarial image x_1 together with a localization prediction $\hat{m}_1$. During training, the ground-truth manipulation mask m is used as spatial guidance, whereas at inference it is replaced by the localization mask predicted by Stage I. We denote the spatial guidance supplied to the refiner as

$$\tilde{m} = \begin{cases} m, & \text{during training,} \\ \hat{m}_1, & \text{during inference and adaptation.} \end{cases} \tag{13}$$

Accordingly, the refiner maintains the same seven-channel input structure in both cases by concatenating the original image x, the coarse adversarial image x_1, and the corresponding spatial guidance $\tilde{m}$.

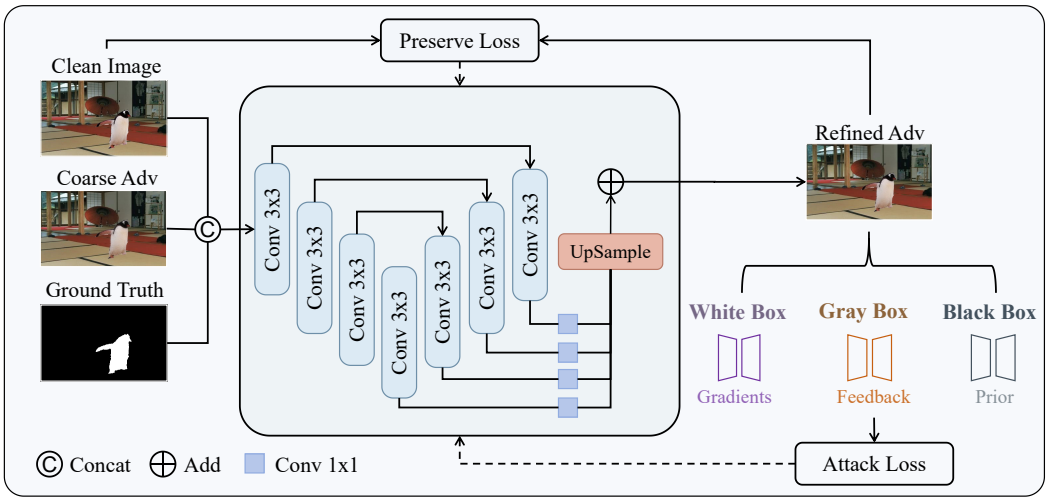


Figure 3: Stage II multi-scale refinement network. The original image x, coarse adversarial image x_1, and spatial guidance are processed by a U-shaped encoder–decoder with skip connections. Multi-scale features are projected by 1×1 convolutions, spatially aligned through the *UpSample* module, and fused to predict the refined adversarial image x_2. Preservation and attack objectives jointly supervise the final output. Ground-truth masks provide spatial guidance during training and are replaced by Stage I localization predictions at inference. The trained refiner is shared across black-box, gray-box, and white-box settings, with victim information used only for additional adaptation when available.

Encoder–decoder refinement. The concatenated input is first projected into a feature representation and processed by a lightweight U-shaped encoder–decoder. The encoder progressively reduces the spatial resolution and extracts features at multiple scales, allowing the network to aggregate both local appearance information and broader contextual cues. Intermediate encoder features are retained and connected to the corresponding decoder levels through skip connections.

Starting from the lowest-resolution representation, the decoder progressively restores spatial resolution. At each level, the upsampled feature is combined with the encoder feature at the same scale and further processed by convolutional blocks. These skip connections preserve fine-grained spatial information from the original and coarse adversarial images while the deeper features provide contextual information for refinement.

Multi-scale fusion. In addition to the main decoding pathway, intermediate features at multiple resolutions are routed to auxiliary projection heads. Each head first projects its feature representation and then resizes the resulting feature to the output resolution. The *UpSample* module shown in Fig. 3 therefore includes both convolutional projection and spatial upsampling.

The aligned multi-scale representations are fused and passed through the final output head to directly predict the refined RGB image:

$$x_2 = R_\psi(\mathrm{Concat}(x, x_1, \tilde{m}))\,, \tag{14}$$

where ψ contains the parameters of the encoder–decoder, multi-scale projection heads, fusion layers, and output head. No independent supervision is applied to the intermediate scales; all branches are jointly optimized through objectives evaluated on the final refined image x_2.

The Stage I model remains frozen throughout refinement learning, and both x_1 and its predicted localization map $\hat{m}_1$ are treated as fixed outputs when used for test-time adaptation. The same refinement architecture is shared across all access settings; the difference lies only in the source of attack-related supervision and whether victim-specific adaptation is available.

A.2.3 Attack Supervision under Different Access Levels

Before access-specific inference, the refinement network is first trained without accessing any victim model. This victim-independent training produces parameters ψ^*_{bb}, which are used directly under black-box access and also serve as the initialization for gray-box and white-box adaptation.

All settings share the same preservation objective

$$\mathcal{L}_{\mathrm{pres}}(x, x_2) = \mathcal{L}_{\mathrm{MSE}}(x, x_2) + \mathcal{L}_{\mathrm{SSIM}}(x, x_2), \tag{15}$$

which encourages the refined image to remain visually close to the original manipulated image.

Localization response and erasure objective. Let $\mathcal{M}(\tilde{m})$ denote the region indicated by the spatial guidance map $\tilde{m}$, and let $\ell_f(y)_p$ denote the localization logit produced by a localizer f at pixel p for image y. We measure the average localization response inside the indicated region as

$$A_f(y; \tilde{m}) = \frac{1}{|\mathcal{M}(\tilde{m})|} \sum_{p \in \mathcal{M}(\tilde{m})} \sigma(\ell_f(y)_p)\,, \tag{16}$$

where σ denotes the sigmoid function. During victim-independent training, $\tilde{m} = m$. During victim-specific adaptation, the Stage I prediction $\hat{m}_1$ provides the spatial guidance instead.

For gray-box gradient estimation and white-box adaptation, we further define the region-aware erasure objective

$$\mathcal{L}^f_{\mathrm{erase}}(y; \tilde{m}) = \frac{1}{|\mathcal{M}(\tilde{m})|} \sum_{p \in \mathcal{M}(\tilde{m})} \mathrm{BCEWithLogits}\left(\ell_f(y)_p, \tau\right), \qquad \tau = 0.1, \tag{17}$$

which encourages low manipulation responses within the region indicated by the available spatial guidance.

Victim-independent refinement. The black-box refiner is trained using the internal localization probe f_{p} learned in Stage I. During this offline training, ground-truth masks are available and the attack objective compares the probe responses of the refined and coarse adversarial images:

$$\mathcal{L}^{\mathrm{bb}}_{\mathrm{atk}} = \left[A_{f_{\mathrm{p}}}(x_2; m) - A_{f_{\mathrm{p}}}(x_1; m) + \kappa\right]_+, \qquad \kappa = 0.03, \tag{18}$$

where $[u]_+ = \max(u, 0)$. The complete victim-independent objective is

$$\mathcal{L}^{\mathrm{bb}}_{\mathrm{ref}} = \mathcal{L}_{\mathrm{pres}}(x, x_2) + \lambda_{\mathrm{atk}} \mathcal{L}^{\mathrm{bb}}_{\mathrm{atk}}. \tag{19}$$

The probe parameters remain frozen, but its computation is differentiable with respect to x_2. Consequently, the attack objective propagates through the probe to the refined image and then through R_ψ, while the preservation objective provides a second gradient path directly from x_2 to the refiner. Only ψ is updated.

After this training, strict black-box inference requires no victim-model interaction. Given x, Stage I produces x_1 and its localization prediction $\hat{m}_1$, and the trained refiner directly produces

$$x^* = R_{\psi^*_{\mathrm{bb}}}\left(\mathrm{Concat}(x, x_1, \hat{m}_1)\right). \tag{20}$$

Thus, the absence of parameter updates under black-box access refers only to test-time inference; the Stage II model has already been trained using victim-independent supervision.

Query-guided adaptation. Under gray-box access, the victim model can be queried but its gradients are unavailable. Adaptation starts from $\psi^{(0)} = \psi^*_{\mathrm{bb}}$. The Stage I localization prediction $\hat{m}_1$ provides the spatial guidance used throughout adaptation:

$$x_2^{(t)} = R_{\psi^{(t)}}\left(\mathrm{Concat}(x, x_1, \hat{m}_1)\right). \tag{21}$$

Queried values of the victim erasure objective are used to estimate its image-space gradient with NES:

$$\widehat{g}^{(t)} \approx \nabla_y \mathcal{L}^{f_v}_{\mathrm{erase}}(y; \hat{m}_1)\big|_{y = x_2^{(t)}}. \tag{22}$$

Since the victim model's gradients are unavailable to the attacker, the estimated gradient is injected into the refinement network through the surrogate objective

$$\mathcal{L}_{\text{atk}}^{\text{gb}} = \frac{1}{N} \left\langle x_2^{(t)}, \text{sg}\left(\widehat{g}^{(t)}\right) \right\rangle, \tag{23}$$

where sg denotes stop-gradient and N denotes the number of scalar elements in the refined image, including all color channels, i.e., $N = \text{numel}(x_2)$.

The gray-box adaptation objective is

$$\mathcal{L}_{\text{ref}}^{\text{gb}} = \mathcal{L}_{\text{pres}}(x, x_2^{(t)}) + \lambda_{\text{atk}} \mathcal{L}_{\text{atk}}^{\text{gb}}. \tag{24}$$

Accordingly, the attack-related contribution to the refinement-parameter gradient is

$$\nabla_\psi \mathcal{L}_{\text{atk}}^{\text{gb}} = \frac{1}{N} \left(\frac{\partial x_2}{\partial \psi} \right)^\top \widehat{g}^{(t)}. \tag{25}$$

NES therefore estimates only the image-space direction using victim queries, while ordinary back-propagation through the differentiable refiner converts this estimate into an update of ψ. The erasure BCE serves only as the queried scalar objective and is not directly differentiated through the victim model. The preservation objective continues to be differentiated normally through R_ψ.

Gradient-guided adaptation. White-box adaptation also starts from ψ_{bb}^*, but full access to the victim model allows direct differentiation through its localization outputs. As in gray-box adaptation, the Stage I prediction $\hat{m}_1$ provides the spatial guidance. The victim-based attack margin is

$$\mathcal{L}_{\text{atk}}^{\text{wb}} = \left[A_{f_v}(x_2; \hat{m}_1) - A_{f_v}(x_1; \hat{m}_1) + \kappa \right]_+, \tag{26}$$

and the complete white-box objective is

$$\mathcal{L}_{\text{ref}}^{\text{wb}} = \mathcal{L}_{\text{pres}}(x, x_2) + \lambda_{\text{atk}} \mathcal{L}_{\text{atk}}^{\text{wb}} + \mathcal{L}_{\text{erase}}^{f_v}(x_2; \hat{m}_1). \tag{27}$$

The victim parameters remain fixed, while gradients from the victim-based attack terms are propagated through the victim model to x_2 and then through R_ψ to update the refinement parameters. The preservation objective follows the same direct path from x_2 to the refiner.

The three access settings therefore share the same Stage I generator, trained Stage II initialization, refinement architecture, and forward data flow. Their difference lies in the information available for test-time adaptation. Strict black-box inference directly uses the victim-independent refiner without further updates; gray-box access uses query-estimated image gradients to adapt the same parameters; and white-box access directly propagates victim gradients through the refined output. In all cases, Stage I and the supervising localization models remain fixed, and only the refinement parameters are updated when victim-specific adaptation is performed.

A.3 More Ablation Experiments

A.3.1 Implementation Details

Table 6 summarizes the key training configurations of A-PAIR. We build A-PAIR on the pretrained FLUX-1.dev backbone and train separate LoRA branches for localization and attack learning. The localization and attack branches use ranks of 32 and 16, respectively, and are alternately optimized at a 1:3 ratio. During Stage I training, flow matching supervises velocity predictions at randomly sampled timesteps, while task-specific outputs are obtained by directly reconstructing the clean endpoint from the predicted velocity. At inference, the learned velocity field is instead integrated over the Euler sampling trajectory to generate the final branch outputs.

The two Stage I branches share the same optimization configuration but use different LoRA capacities. The attack target is initialized with the input image and refreshed using previously generated adversarial images after each attack update, while the localization branch uses the ground-truth mask as its diffusion target. The wavelet loss adopts a two-level Haar decomposition and constrains the LL component with an L_1 distance.

Table 6: Key training configurations of A-PAIR.

Component	Parameter	Value
Stage I	Optimizer	AdamW
	Learning rate	1×10^{-4}
	Weight decay	1×10^{-4}
	AdamW betas	$(0.9, 0.999)$
	Batch size	1
	precision	bf16
	LoRA rank (Localization / Attack)	$32/16$
	LoRA alpha (Localization / Attack)	$64/32$
Stage I loss	λ_{dice}	0.5
	$\lambda_{\text{FM}}^{(a)}$	0.5
	$\lambda_{\text{FM}}^{(l)}$	0.5
	λ_w	0.5
	Wavelet configuration	Haar, 2 levels, LL subband
	Wavelet distance	L_1
Stage II	Optimizer	AdamW
	Learning rate	1×10^{-4}
	Training steps	2,000
	Weight decay	1×10^{-4}
	AdamW betas	$(0.9, 0.999)$
	precision	bf16
	λ_{atk}	1.5
	Margin κ	0.03
	Preservation loss	MSE + SSIM (equal weights)

Stage II uses a lightweight U-shaped refinement network and is trained with the frozen Stage I model. It is optimized using AdamW and combines preservation and attack objectives with $\lambda_{\text{atk}} = 1.5$ and $\kappa = 0.03$. The preservation loss assigns equal weights to MSE and SSIM. During victim-specific adaptation, only the refinement parameters are updated, while Stage I and the supervising localization models remain frozen.

Table 7: Access-conditioned adaptation and evaluation configurations.

Setting	Parameter	Value
Gray-box	Query budget	20 queries/image
	NES scale σ	0.01
	Sampling directions	2
	Antithetic sampling	Yes
White-box	Gradient update budget	10 updates/image
	Learning rate	1×10^{-4}
Evaluation	Input resolution	512×512
	VAE downsampling factor	8
	Hardware	NVIDIA A100-PCIE-40GB
	Mask threshold	0.5
	PSNR data range	1.0
	LPIPS backbone	VGG

Table 7 summarizes the access-conditioned adaptation and evaluation configurations. Under black-box access, the trained victim-independent refiner is directly applied without further updates. For gray-box adaptation, A-PAIR uses 20 victim-model queries per image and estimates image-space gradients with NES using two antithetic directions and $\sigma = 0.01$. White-box adaptation performs 10 gradient-based refinement updates per image.

All experiments are conducted at a resolution of 512×512 on NVIDIA A100-PCIE-40GB GPUs. Localization masks are binarized using a threshold of 0.5. PSNR is computed with a data range of

1.0, and LPIPS is evaluated using the VGG backbone. Additional implementation details and source code will be provided to facilitate reproduction.

A.3.2 Effect of Sampling Steps

We investigate the effect of the number of Euler sampling steps K used to generate task outputs. This ablation changes the discretization of the inference trajectory, while the training objective remains unchanged and continues to supervise velocity predictions at randomly sampled timesteps. Table 8 reports the results for different sampling budgets.

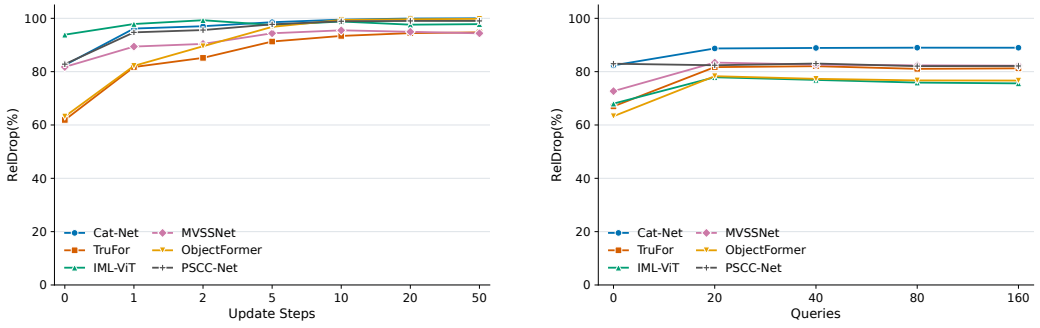
Table 8: Effect of the number of Euler sampling steps. RelDrop (%) measures attack effectiveness, while PSNR (dB) measures visual fidelity.

Sampling Steps K	RelDrop ↑	PSNR ↑
1	52.7	24.05
5	67.3	27.07
10	70.2	30.62
20	**73.9**	**32.42**
28	66.6	31.74

Increasing the number of sampling steps from 1 to 20 consistently improves both attack effectiveness and visual fidelity, with the largest gain appearing at the lower end of the sampling range. However, further increasing the number of steps to 28 does not yield additional benefits and instead reduces attack effectiveness while slightly degrading visual quality. This suggests that excessively fine Euler discretization is not necessarily beneficial for the learned anti-forensic prior. Among the evaluated settings, $K = 20$ provides the best overall trade-off and is therefore used in the main experiments.

A.3.3 Effect of Victim-Specific Adaptation Budget

We further examine how the amount of victim-specific information affects the adaptation of the Stage II refiner. All configurations start from the same victim-independent initialization ψ^*_{bb}, with zero budget corresponding to direct black-box inference. Figure 4 reports the sensitivity to the number of gradient updates under white-box access and victim-model queries under gray-box access.



(a) White-box adaptation (b) Gray-box adaptation

Figure 4: Effect of victim-specific adaptation budget on different localization models. (a) White-box adaptation with increasing numbers of gradient updates. (b) Gray-box adaptation with increasing victim-model query budgets. A zero budget corresponds to the victim-independent Stage II refiner before target-specific adaptation. Results are reported as RelDrop (%).

White-box adaptation. As shown in Fig. 4(a), direct gradient guidance rapidly specializes the victim-independent refiner to the target localization model. Most models obtain substantial gains within the first few updates, after which the improvement gradually saturates. This behavior indicates that the learned initialization already provides a strong anti-forensic starting point, requiring only limited gradient-based adaptation to capture model-specific vulnerabilities.

Gray-box adaptation. Figure 4(b) shows a more moderate and model-dependent response to additional victim queries. Most of the improvement appears at relatively small query budgets, while further interaction generally produces diminishing or less consistent gains. Compared with white-box adaptation, this slower saturation reflects the weaker supervision provided by query-estimated gradients. Nevertheless, the improvement over the zero-query initialization shows that limited victim feedback can still effectively specialize the transferable refinement prior.

A.3.4 Robustness to Common Image Transformations

We further evaluate the robustness of A-PAIR under JPEG compression and image resizing. Each transformation is applied identically to clean and adversarial images, and RelDrop is computed from their corresponding post-processed F1 scores. The results are summarized in Table 9.

Table 9: Robustness of A-PAIR under JPEG compression and image resizing. Results are reported as RelDrop (%).

Transformation	Setting	RelDrop ↑
JPEG	Quality 95	**72.0**
JPEG	Quality 85	69.3
JPEG	Quality 75	67.6
Resize	Scale 0.8	74.2
Resize	Scale 0.6	**74.3**
Resize	Scale 0.4	71.5

A-PAIR maintains strong attack effectiveness across the tested JPEG compression and resizing settings, indicating limited sensitivity to these common post-processing operations. Performance remains largely stable under resizing and decreases only moderately with stronger JPEG compression, with neither transformation substantially compromising attack effectiveness.

A.3.5 Visual Quality Comparison

In addition to attack effectiveness, we evaluate the visual quality of the generated adversarial images. We report PSNR and LPIPS between the original images and the generated outputs, where higher PSNR and lower LPIPS indicate better visual fidelity. The results are summarized in Table 10.

Table 10: Visual quality of adversarial images under three attacker access settings. PSNR is reported in dB. Best results within each setting are bolded.

Setting	Method	PSNR↑	LPIPS↓
Black-box	AdvGAN	29.4	0.115
	Diff-CFG	30.8	**0.036**
	A-PAIR(BB)	**32.4**	0.084
Gray-box	Square Attack	27.7	0.154
	AdvFor	29.4	0.121
	A-PAIR(GB)	**33.6**	**0.062**
White-box	FGSM	27.8	0.151
	PGD	30.5	0.060
	A-PAIR(WB)	**36.7**	**0.032**

As shown in Table 10, A-PAIR generally achieves higher visual fidelity under different access settings. The black-box model already maintains competitive perceptual quality without any victim-specific optimization, demonstrating the effectiveness of the learned anti-forensic prior. Victim-specific adaptation further improves visual fidelity while preserving attack capability, with gray-box and white-box adaptation producing progressively refined outputs. Compared with iterative gradient-based attacks, which often introduce larger visible distortions, A-PAIR benefits from the diffusion-based generation and refinement process to maintain more natural image appearance.

A.4 More Visualization Results

We provide additional qualitative results to further demonstrate the effectiveness of A-PAIR. These visualizations complement the quantitative results from three perspectives. First, we show that adaptive perturbation refinement substantially improves the visual quality of the coarse adversarial images while preserving their attack effectiveness. Second, extensive examples under black-box, gray-box, and white-box access illustrate that A-PAIR consistently suppresses manipulation localization across diverse victim models while maintaining high visual fidelity. Third, comparisons with representative baseline attacks further highlight the stronger and more consistent anti-forensic behavior of A-PAIR across different attacker-access settings.

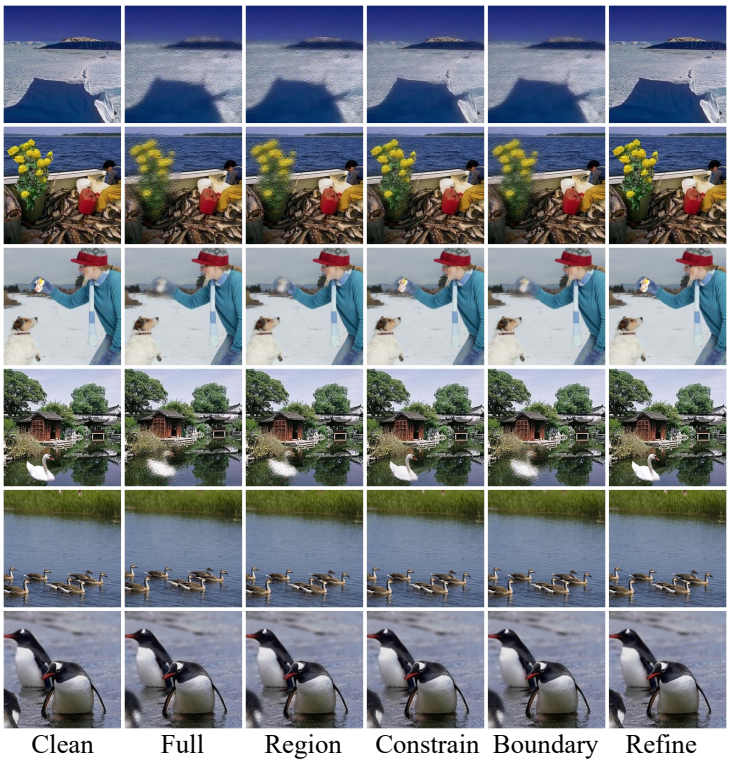


Figure 5: Visual comparison between the coarse Stage I output, several representative fixed refinement strategies, and the learned Stage II refinement.

A.4.1 Effect of Adaptive Perturbation Refinement

Figures 5 and 6 visualize how adaptive perturbation refinement improves the coarse adversarial outputs produced by Stage I. The fixed constrained variants can partially remove visible perturbations, but their effectiveness is limited by the predefined form of the constraint. In contrast, the learned refinement consistently produces images closer in appearance to the original input while retaining the attack-relevant changes. The improvement is particularly evident in the manipulated regions, where visually noticeable artifacts in the coarse adversarial images are substantially reduced after refinement. Together with the quantitative results in Table 5, these examples demonstrate that Stage II provides a substantially better attack–fidelity trade-off than hand-crafted perturbation constraints.

A.4.2 Qualitative Results of A-PAIR

Figures 7–12 provide additional qualitative results of A-PAIR under different victim-access settings. Across all examples, the generated adversarial images remain visually close to the corresponding inputs, while the localization predictions are strongly suppressed or substantially misaligned with the ground-truth manipulation regions. Under black-box access, the same adversarial image is evaluated across victim models without target-specific adaptation. Under gray-box and white-box access,

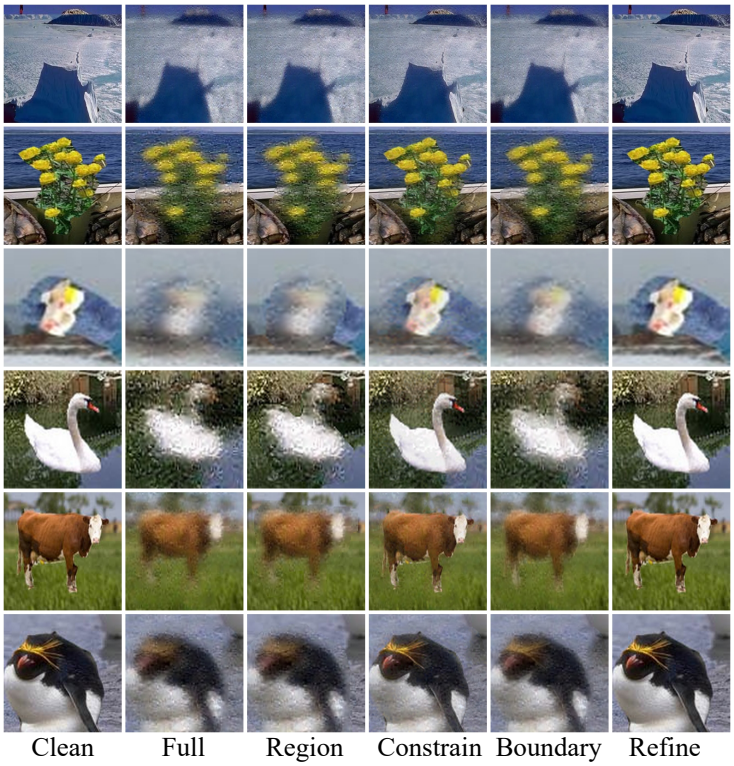


Figure 6: Enlarged comparison of the manipulated regions corresponding to Fig. 5, highlighting local visual differences before and after Stage II refinement.

attacks are adapted separately to each victim; the displayed RGB image is the CAT-Net example, while each prediction map corresponds to its own victim-specific adversarial image.

Under black-box access, A-PAIR already produces highly transferable adversarial effects across different localization architectures without using any victim-specific feedback. The predicted manipulation regions become weakened, fragmented, displaced, or dominated by irrelevant responses, demonstrating that the learned anti-forensic prior generalizes beyond the model used during training. With limited prediction feedback, gray-box adaptation further suppresses model-specific localization responses and produces more pronounced failures on the target models. White-box refinement provides the strongest target-specific adaptation: direct gradient guidance substantially reduces responses within the manipulated regions and, in many cases, causes the predicted masks to nearly disappear or deviate entirely from the ground truth.

These visual results clearly illustrate the progressive benefit of victim information in A-PAIR. The transferable prior alone is sufficient to induce strong cross-model localization failures, while query and gradient feedback further specialize the same refinement mechanism toward individual victim models without compromising the visual quality of the adversarial images.

A.4.3 Comparison with Baseline Attacks

Figures 13–21 provide qualitative comparisons between A-PAIR and representative baseline attacks under black-box, gray-box, and white-box settings. We use the same input images across all three settings to illustrate how additional victim feedback or gradients can improve both attack effectiveness and visual fidelity. Across the illustrated examples, A-PAIR more effectively weakens or misaligns localization responses while maintaining high visual similarity to the inputs.

Under black-box access, each method's adversarial image is evaluated across victim models without target-specific adaptation. Under gray-box and white-box access, adversarial images are generated separately for each victim. The displayed A-PAIR image corresponds to CAT-Net, while each lo-

calization response is obtained from the adversarial image generated for the respective victim. The baselines follow the same access-specific evaluation protocol.

In the black-box setting, existing transfer-based attacks often remain effective only on a subset of victim models, with clear localization responses still preserved around the manipulated regions. In contrast, A-PAIR consistently weakens or displaces these responses across heterogeneous localization architectures, demonstrating markedly stronger cross-model transferability. The example comparison also shows that baseline methods frequently retain concentrated responses around the ground-truth region, whereas A-PAIR breaks this correspondence across multiple victims simultaneously.

The advantage remains evident under gray-box access. Query-based baselines typically require iterative victim interaction to gradually search for effective perturbations, whereas A-PAIR starts from an already transferable anti-forensic prior and uses the limited feedback primarily for target-specific refinement. Consequently, its localization suppression is stronger and more consistent across victim models under the same restricted-access setting.

Under white-box access, A-PAIR also compares favorably with conventional gradient-based attacks. While existing methods directly optimize perturbations from the input image, A-PAIR begins from a learned anti-forensic prior and further specializes it using victim gradients. This results in more complete suppression or stronger misalignment of the localization responses across the illustrated examples. Overall, these qualitative comparisons complement the quantitative results by demonstrating the effectiveness of A-PAIR across different access settings. The matched examples further illustrate the benefit of additional victim information: query feedback and gradient guidance enable more effective interference with manipulation localization while producing adversarial images closer to the original inputs.

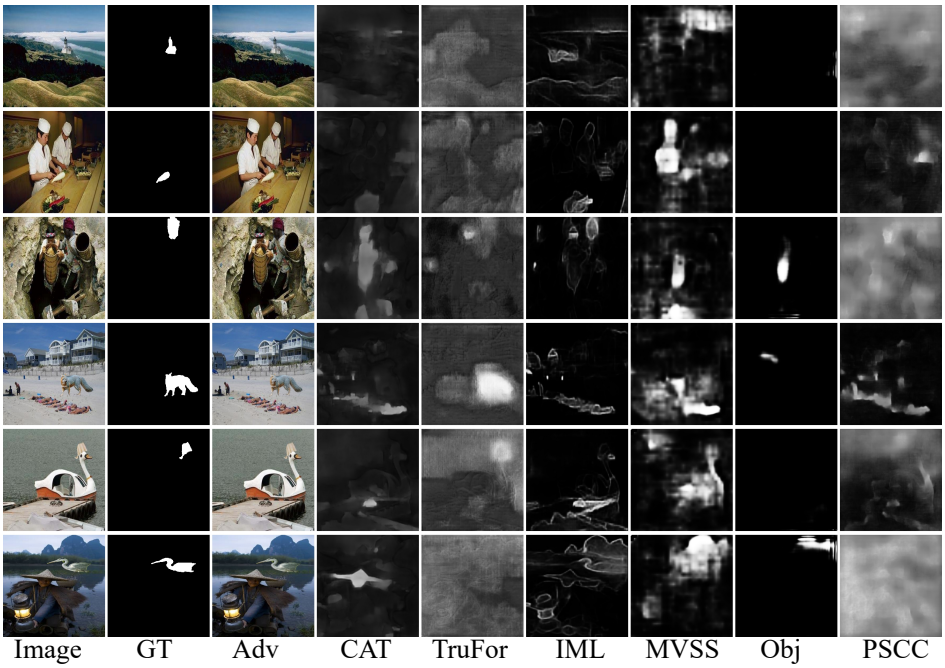


Figure 7: Additional qualitative results under the black-box setting.

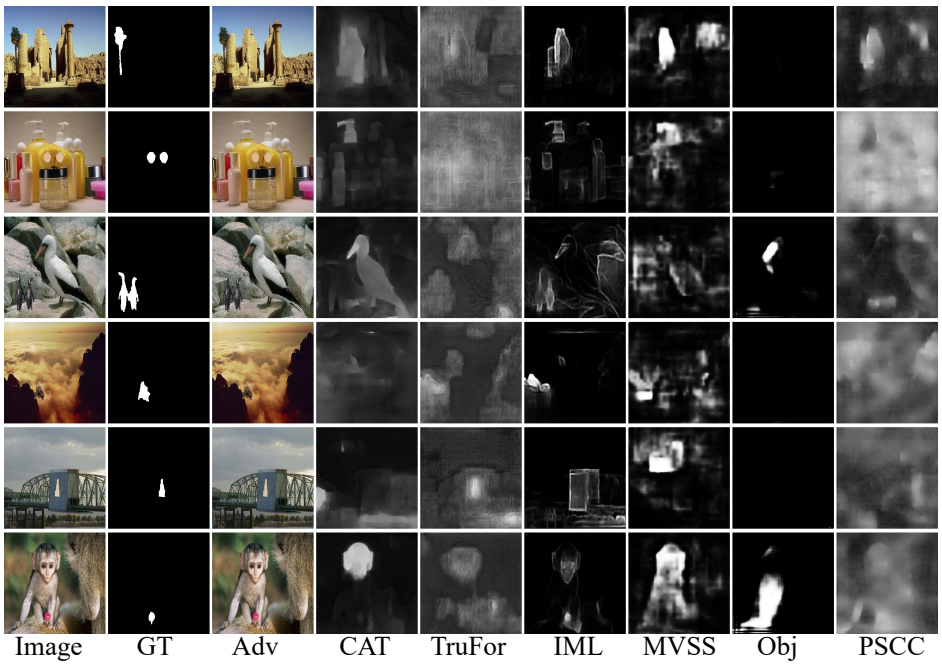


Figure 8: Additional qualitative results under the black-box setting.

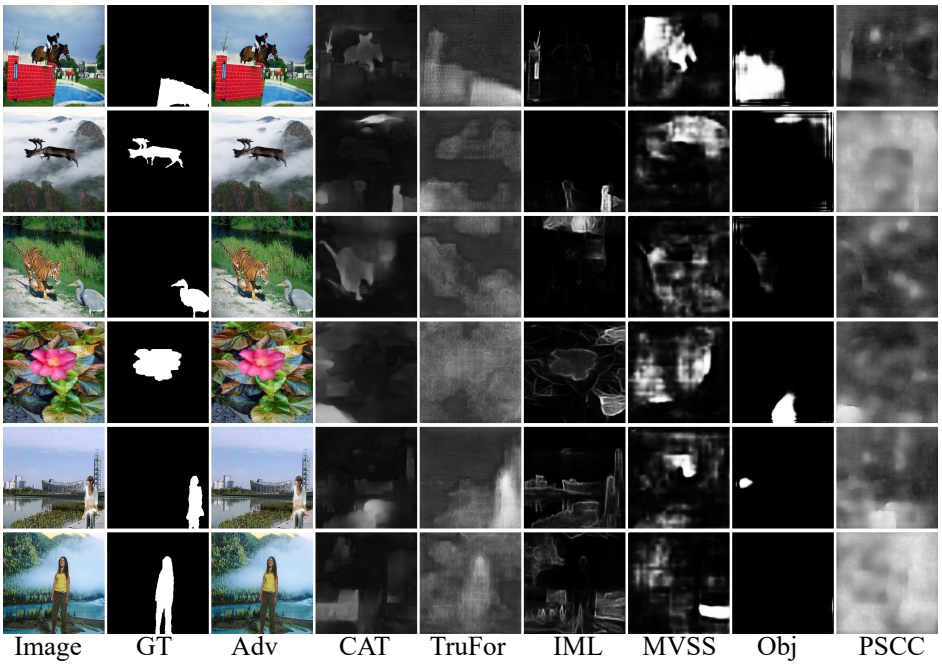


Figure 9: Additional qualitative results under the gray-box setting. The example adversarial images are optimized on CAT-Net.

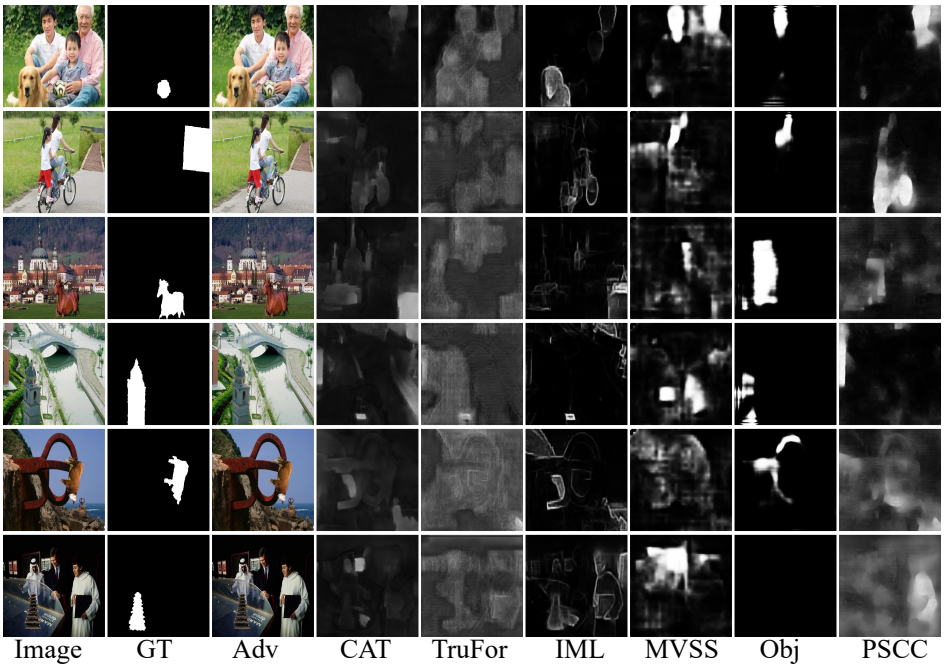


Figure 10: Additional qualitative results under the gray-box setting. The example adversarial images are optimized on CAT-Net.

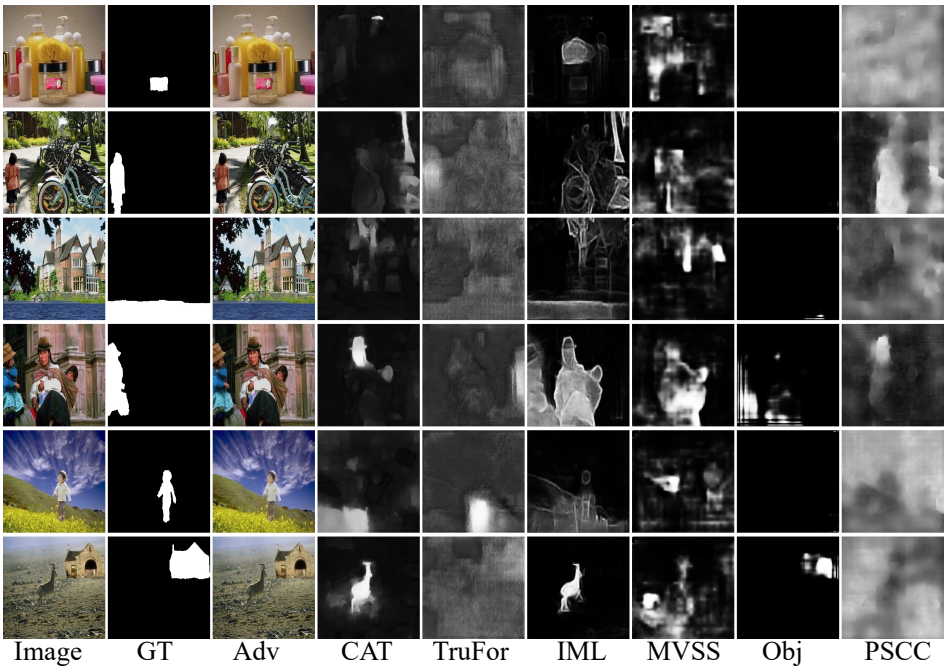


Figure 11: Additional qualitative results under the white-box setting. The example adversarial images are optimized on CAT-Net.

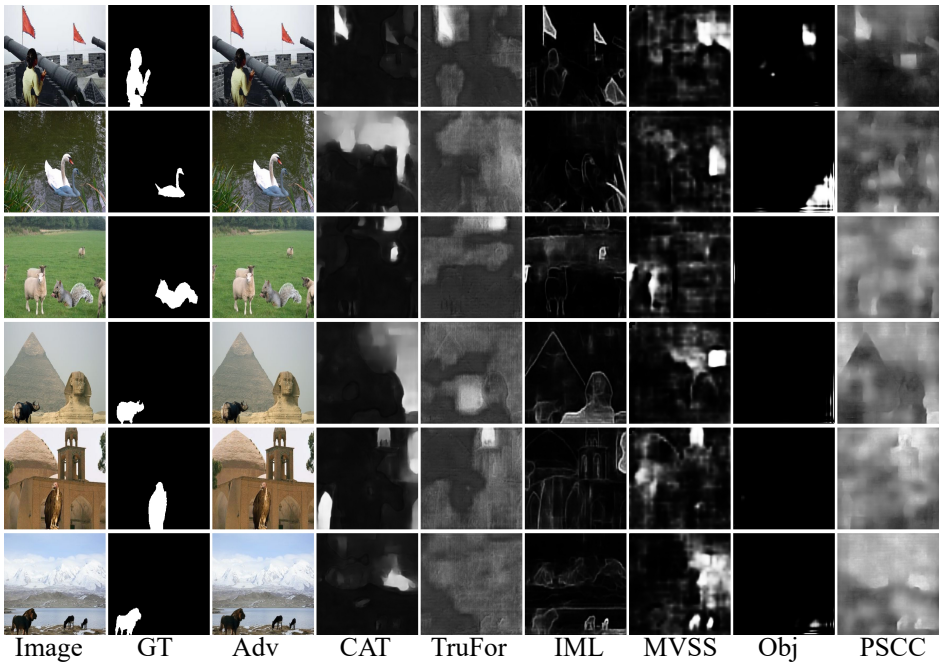


Figure 12: Additional qualitative results under the white-box setting. The example adversarial images are optimized on CAT-Net.

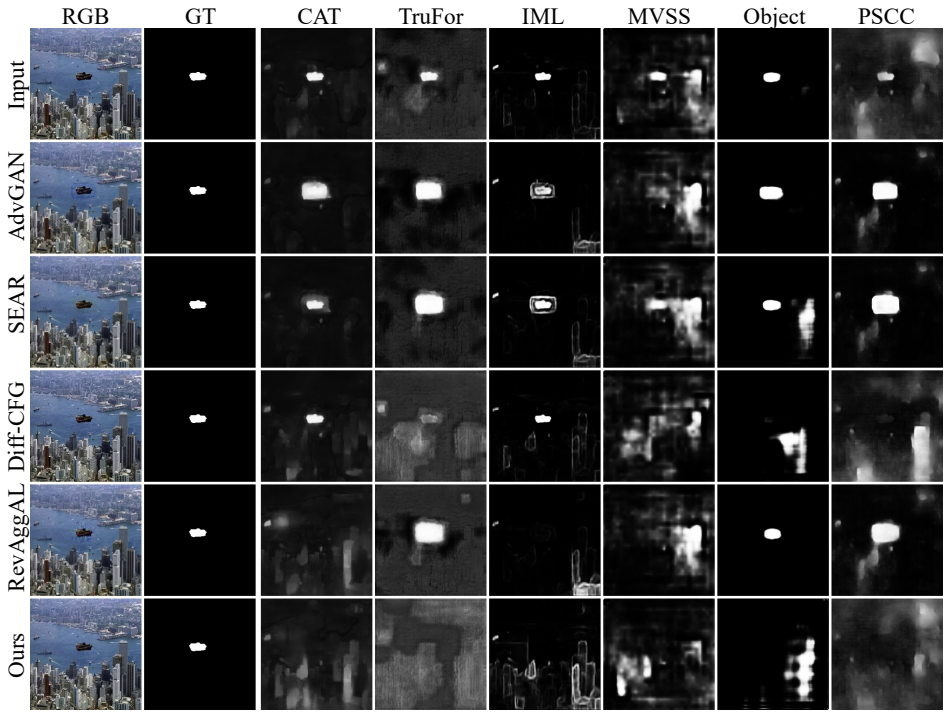


Figure 13: Qualitative comparison with representative baseline attacks under the black-box setting.

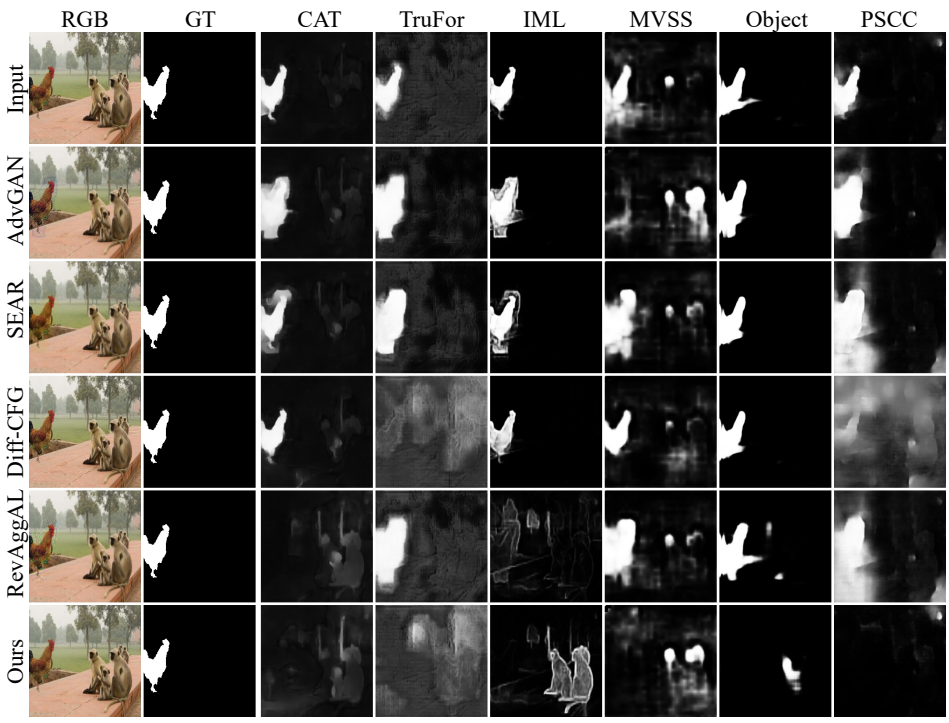


Figure 14: Qualitative comparison with representative baseline attacks under the black-box setting.

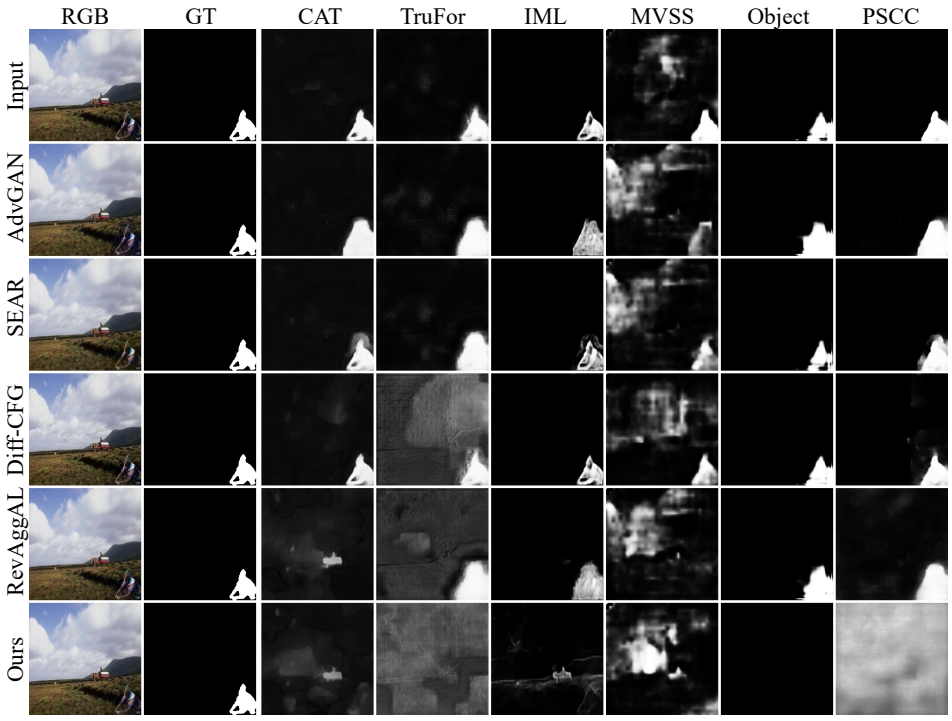


Figure 15: Qualitative comparison with representative baseline attacks under the black-box setting.

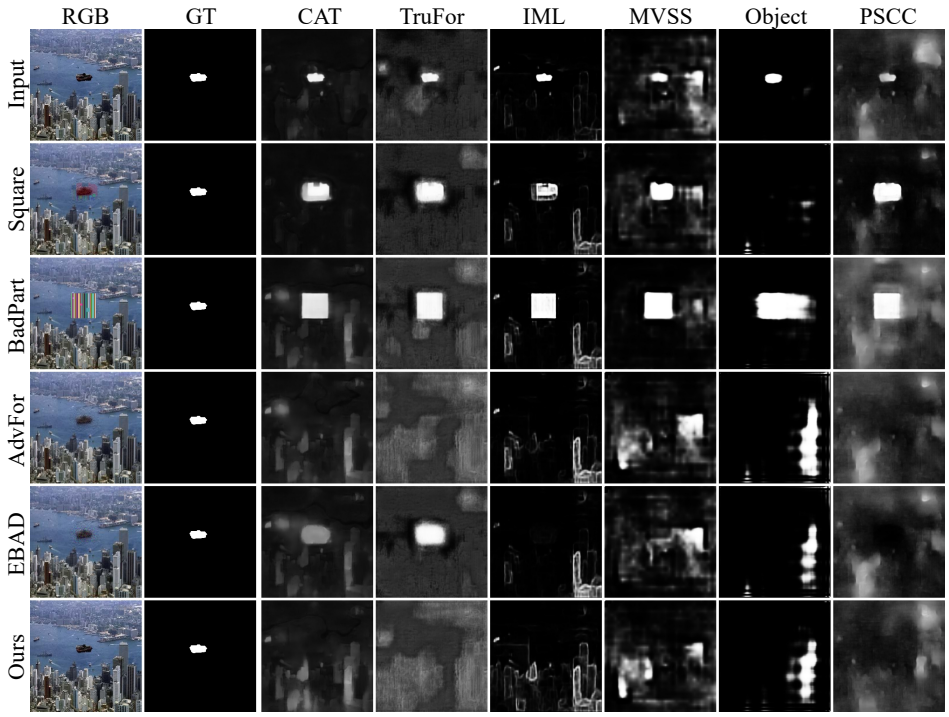


Figure 16: Qualitative comparison with representative baseline attacks under the gray-box setting. The example adversarial images are optimized on CAT-Net.

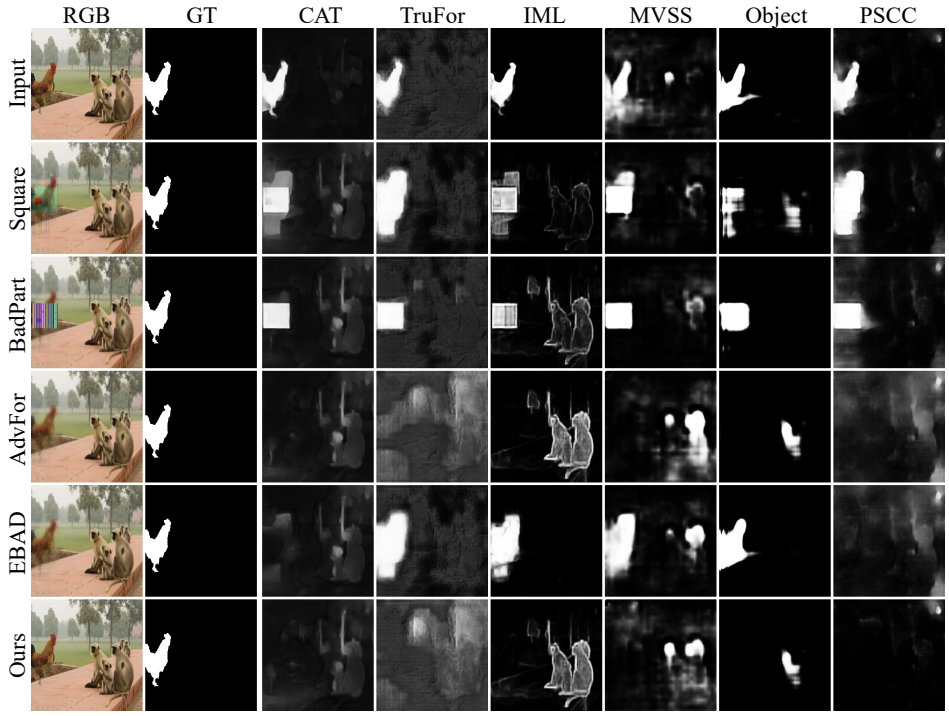


Figure 17: Qualitative comparison with representative baseline attacks under the gray-box setting. The example adversarial images are optimized on CAT-Net.

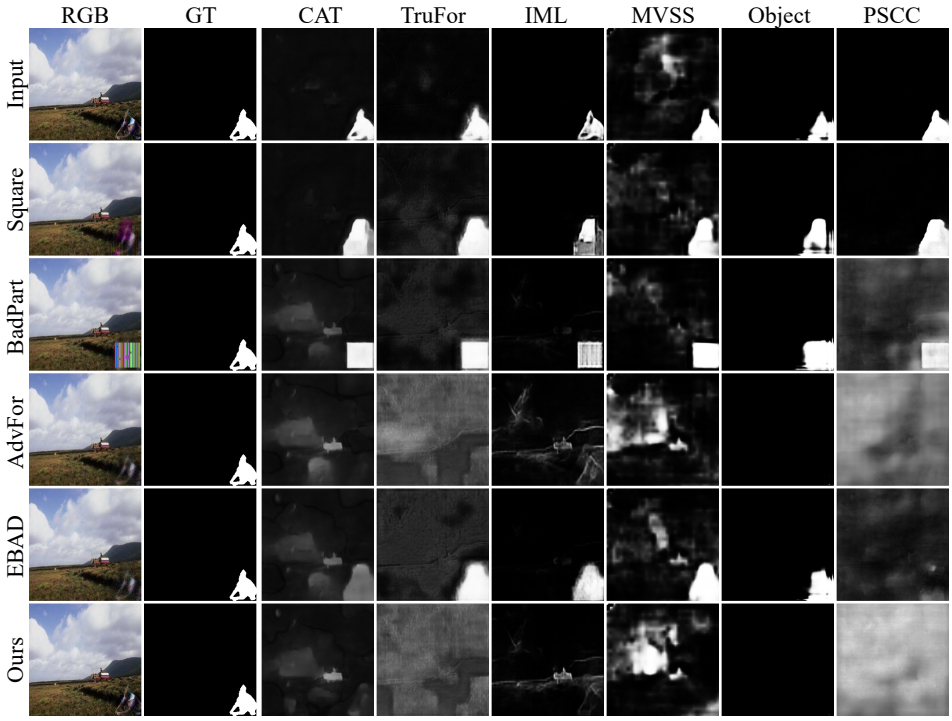


Figure 18: Qualitative comparison with representative baseline attacks under the gray-box setting. The example adversarial images are optimized on CAT-Net.

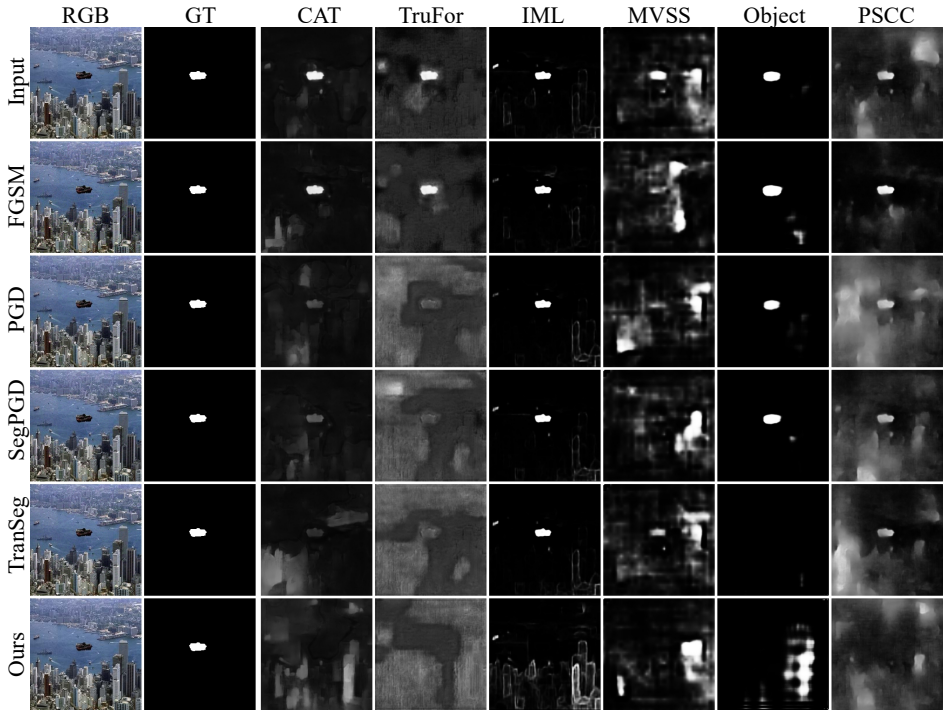


Figure 19: Qualitative comparison with representative baseline attacks under the white-box setting. The example adversarial images are optimized on CAT-Net.

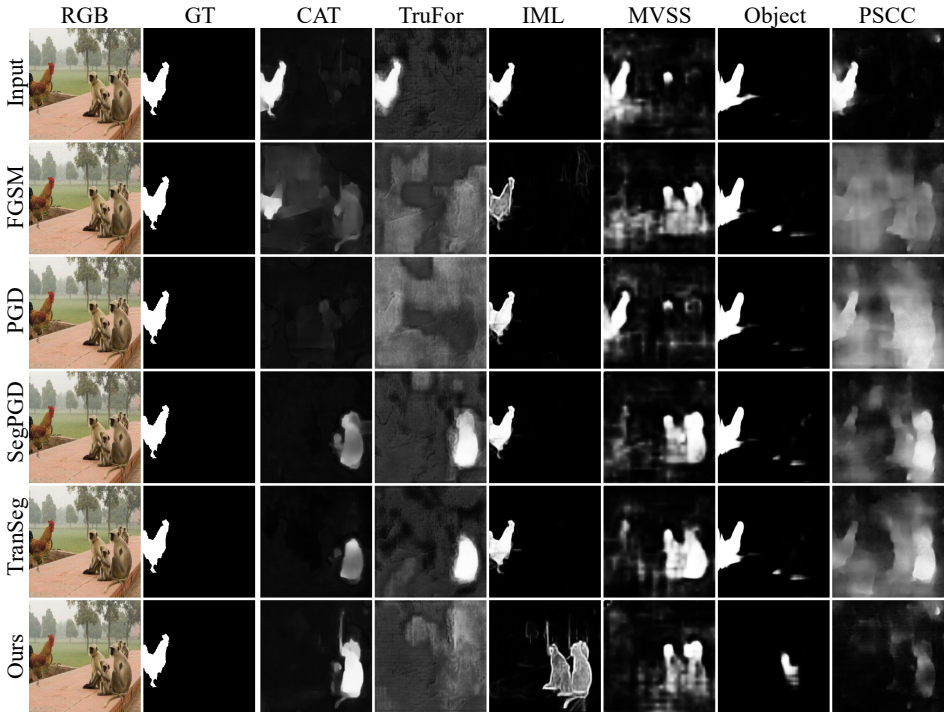


Figure 20: Qualitative comparison with representative baseline attacks under the white-box setting. The example adversarial images are optimized on CAT-Net.

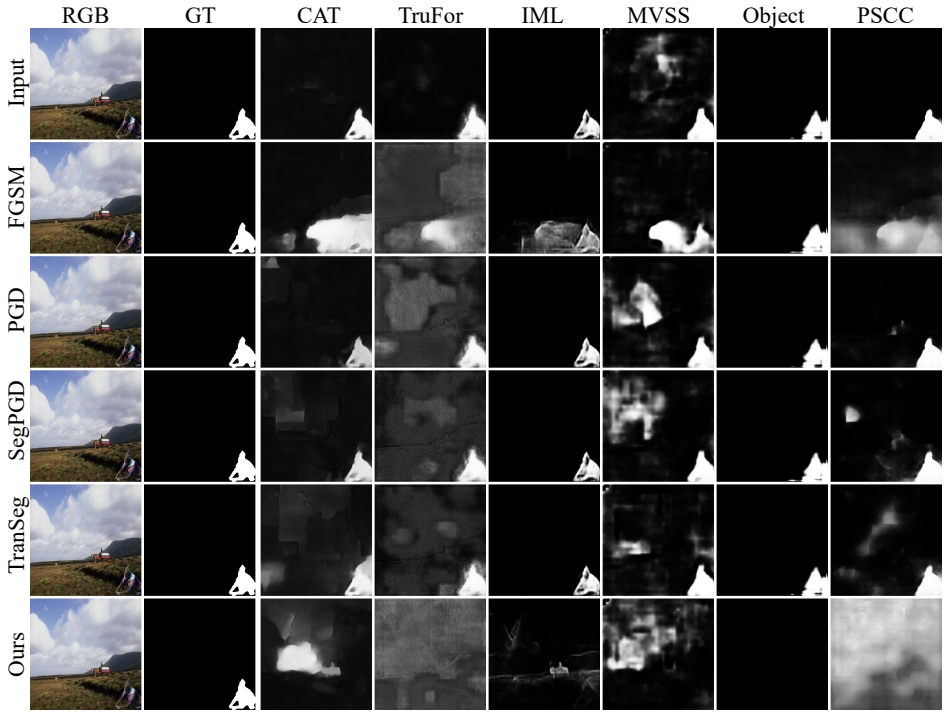


Figure 21: Qualitative comparison with representative baseline attacks under the white-box setting. The example adversarial images are optimized on CAT-Net.