DASH: LEARNING ADVERSARIAL DISTRIBUTION SHIFTS FOR TRANSFERABLE BLACK-BOX ATTACKS ON OBJECT DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object detectors are widely deployed in safety-critical applications. Since deployed models conceal internal information, transferable black-box attacks provide a practical robustness evaluation. Unlike image-classification attacks, detection attacks must suppress multiple objects while transferring across heterogeneous architectures. Our gradient analysis identifies two obstacles: detector gradients are poorly aligned, encouraging surrogate-specific solutions, while object gradients can conflict and interfere with joint suppression. Distribution-level generation provides a natural mechanism for this goal: by aggregating attack supervision over the training distribution and encoding it into generator parameters, it can capture reusable adversarial patterns rather than instance- and surrogate-specific optimization directions, while enabling efficient inference-time generation. Based on this insight, we propose Detector Adversarial Distribution Shift (DASH), which aggregates attack supervision over the training distribution and encodes reusable knowledge into an input-conditioned mapping from natural to adversarial images. DASH combines class-agnostic confidence suppression to coordinate multiple objects with CLIP feature disruption to reduce surrogate dependence. We instantiate this formulation with conditional GAN and Stable Diffusion generators. Across three datasets and thirteen detectors, DASH-GAN and DASH-SD achieve average ASRs of 83.6% and 80.08%, outperforming the strongest baseline by up to 36.14 percentage points with generation times of 0.02 and 0.75 seconds per image. DASH-SD also averages 28.14% ASR on two ODDefense-protected models, exceeding the strongest prior baseline by 10.76 percentage points. Thus, DASH balances black-box transferability, generation efficiency, and defense robustness.

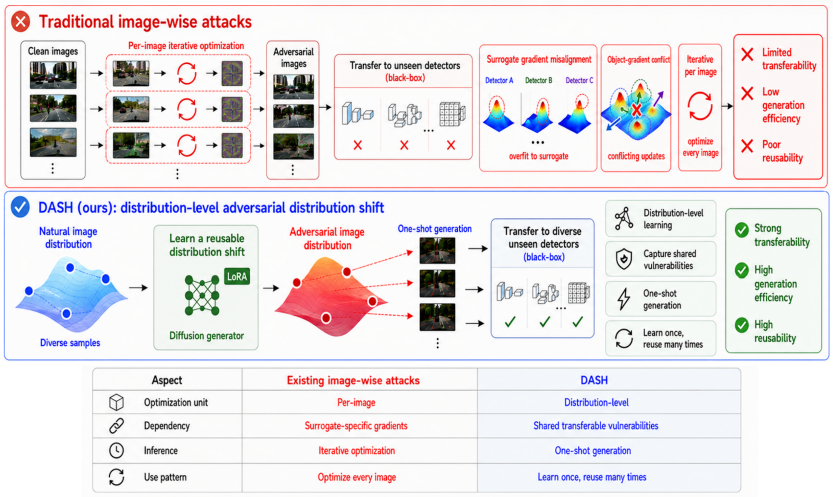
1 INTRODUCTION

Object detectors are increasingly deployed in safety-critical systems(Liu et al., 2020), including autonomous driving(Arnold et al., 2019), intelligent surveillance(Sreenu & Durai, 2019), and robotic perception(Ahmad & Rahimi, 2022), making adversarial robustness a critical security concern. In realistic black-box settings, attackers cannot access the deployed detector's architecture or parameters and must craft adversarial examples using a surrogate model, requiring cross-architecture transferability. This is particularly challenging for detection: unlike classification(Xie et al., 2017)(Cai et al., 2022), which involves a single image-level prediction, detection must handle multiple objects with varying categories, locations, and scales. Thus, a transferable attack must both suppress multiple objects within a scene and generalize beyond its surrogate detector.

Most existing attacks employ image-wise optimization (Xie et al., 2017; Li et al., 2018; Chow et al., 2020), using surrogate gradients to optimize each input independently (Liao et al., 2021; Ding et al., 2024). Our analysis in Sec. 3.1 identifies two transferability barriers: cross-detector gradient inconsistency and inter-object gradient conflict. These findings motivate learning reusable attack knowledge across samples instead of relying on image-specific optimization.

Universal perturbation methods take a first step toward sharing attack knowledge across samples by optimizing a common perturbation over a training set(Wang et al., 2024), thereby avoiding repeated

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.



Aspect	Existing image-wise attacks	DASH
Optimization unit	Per-image	Distribution-level
Dependency	Surrogate-specific gradients	Shared transferable vulnerabilities
Inference	Iterative optimization	One-shot generation
Use pattern	Optimize every image	Learn once, reuse many times

Figure 1: A high-level comparison between DASH and existing image-wise attack methods.

image-wise optimization. However, this knowledge is compressed into a single image-agnostic template and cannot adapt to varying object layouts or visual content. Conditional generative attacks offer a more flexible alternative by amortizing attack optimization over the training distribution, encoding shared knowledge into generator parameters while producing input-dependent perturbations. This distribution-level learning is particularly suitable for transferable attacks, since diverse training samples jointly update the same transformation, reducing reliance on image-specific optimization trajectories and encouraging reusable adversarial patterns. However, distribution-level learning alone does not guarantee that the shared knowledge is transferable. Existing generative attacks such as UEA(Wei et al., 2019) still rely on category-dependent predictions and surrogate-specific representations, so the learned transformation may encode conflicting object-level signals or surrogate-specific vulnerabilities. The central question is therefore not merely how to avoid iterative optimization, but what attack signals should be shared across samples so that the learned transformation is both consistent across objects and transferable across detectors.

We address this question with Detector Adversarial Distribution Shift (DASH). Instead of independently optimizing perturbations in pixel space, DASH learns a shared input-conditioned transformation from the natural image distribution to an adversarial distribution. By optimizing this transformation jointly over diverse training samples, attack knowledge is accumulated in the generator parameters rather than in image-specific perturbations, enabling the model to capture reusable adversarial regularities and reducing instance-specific overfitting. This distribution-level formulation constitutes the first component of DASH and provides the basis for efficient and transferable attack generation.

To ensure that the learned attack knowledge itself is transferable, we further introduce transfer-oriented supervision corresponding to the two gradient inconsistencies identified above. A class-agnostic confidence suppression objective establishes a common suppression goal across objects, alleviating inter-object gradient conflict, while an object-region CLIP feature disruption objective provides detector-external semantic supervision, reducing dependence on surrogate-specific representations. In this way, distribution-level learning determines how attack knowledge is shared across samples, whereas the two objectives determine what transferable knowledge is learned. Together, they encourage adversarial transformations that generalize across images, objects, and detector architectures.

This separation is also empirically important. When the same class-agnostic and CLIP objectives are optimized independently for each image using PGD, strong white-box performance does not translate into comparable black-box transferability. In contrast, optimizing these objectives through a shared generator over the training distribution substantially improves transfer to unseen detectors. This controlled comparison shows that the gain of DASH does not arise solely from the proposed objectives: distribution-level generative learning itself plays a critical role by converting transfer-oriented supervision from image-specific optimization signals into reusable attack knowledge across samples. We instantiate this formulation with both a conditional GAN and a pretrained Stable Diffusion model, resulting in DASH-GAN and DASH-SD.

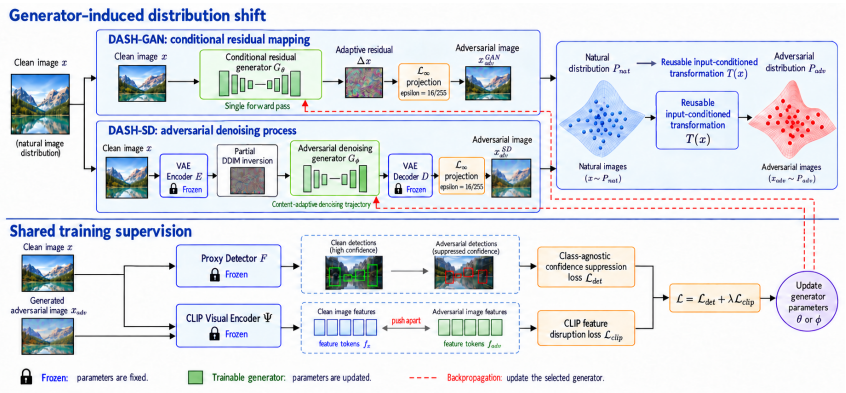


Figure 2: Pipeline of DASH. A generator learns an input-conditioned mapping from the natural image distribution to an adversarial distribution using a frozen surrogate detector and CLIP visual encoder.

Our contributions are threefold. First, we identify cross-detector gradient inconsistency and inter-object gradient conflict as two key obstacles to transferable object detection attacks. Second, we propose DASH, a distribution-level framework that learns reusable, input-conditioned adversarial transformations, together with class-agnostic suppression and detector-external semantic disruption to improve transferability across objects and detector architectures. Finally, we instantiate DASH with GAN and Stable Diffusion generators and demonstrate strong cross-model, cross-dataset, and defense transferability across three datasets and thirteen detectors, while enabling efficient inference without per-image optimization.

2 Related Work

Object Detection. Object detection jointly predicts categories and locations. Representative architectures include two-stage proposal-based detectors such as Faster R-CNN Ren et al. (2015) and Mask R-CNN He et al. (2017); one-stage dense predictors such as SSD Liu et al. (2016), YOLO Ultralytics (2020); Redmon & Farhadi (2017), VFNet Zhang et al. (2021), and TOOD Feng et al. (2021); and Transformer-based set predictors such as DETR Carion et al. (2020) and Deformable DETR Zhu et al. (2020). Their architectural diversity provides a demanding testbed for adversarial transferability.

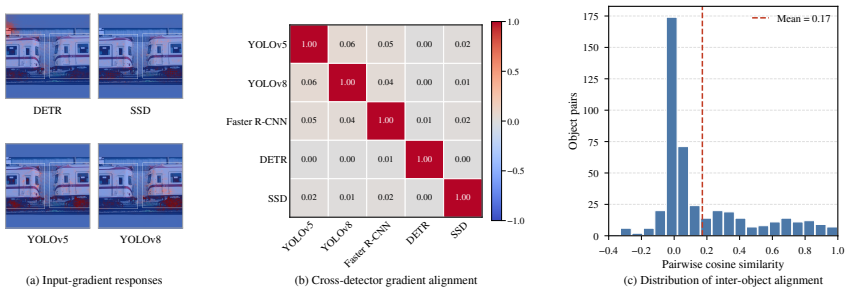
Transferable Adversarial Attacks. Transfer attacks craft examples on a surrogate and apply them to unknown victims without queries. Classification transferability has been improved with momentum (Dong et al., 2018), input diversification (Xie et al., 2019), translation-invariant gradients (Dong et al., 2019), and feature disruption (Zhou et al., 2018). For detection, CW-Attack (Liao et al., 2021) uses category-wise optimization across anchor-free and anchor-based models, while G-UAP (Wu et al., 2019) suppresses RPN foreground proposals with a reusable perturbation. Image-wise attacks nevertheless require iterative optimization, whereas universal attacks remain tied to one fixed surrogate-optimized template.

Generative Adversarial Attacks. UEA (Wei et al., 2019) replaces iterative optimization with a conditional GAN, combining the DAG classification loss with surrogate-specific attention features; it evaluates transfer from F-RCNN to SSD within one generator family. Diffusion-based methods such as Diffusion to Confusion (Lin et al., 2025), AdvLogo (Miao et al., 2024), and BadPatch (Wang et al., 2025) improve naturalness but mainly generate localized physical patches. DASH instead learns an input-conditioned distribution mapping instantiated by both GAN and diffusion generators, enabling reusable full-image attacks with stronger cross-detector and cross-dataset transferability.

3 Method: Detector Adversarial Distribution Shift

3.1 Motivation: Gradient Barriers to Transferability

To understand why image-wise attacks often exhibit limited black-box transferability, we analyze their input gradients from two complementary perspectives: consistency across detector architectures and compatibility among objects within an image. We use the same DAG attack objective for



(a) Input-gradient responses (b) Cross-detector gradient alignment (c) Distribution of inter-object alignment

Figure 3: Gradient responses across detectors (a), inter-detector gradient correlations (b), and inter-object gradient correlations (c).

all detectors and objects, ensuring that the observed differences arise from the optimization landscapes rather than from different loss functions. Let

$$g_d = \nabla_x \ell_{\mathrm{DAG}}(F_d(x), y), \qquad g_{d,i} = \nabla_x \ell^{(i)}_{\mathrm{DAG}}(F_d(x), y), \tag{1}$$

denote the image-level gradient of detector F_d and the gradient associated with object i, respectively. Their alignment is measured using cosine similarity.

Cross-detector gradient inconsistency. As shown in Fig. 3(a), heterogeneous detectors produce spatially different gradient-response patterns for the same image. The pairwise cosine similarities in Fig. 3(b) are also close to zero, quantitatively showing that their effective optimization directions are poorly aligned. Consequently, a perturbation optimized from the local gradient of one surrogate detector can easily encode architecture-specific sensitivities and transfer poorly to unseen detectors.

Inter-object gradient conflict. Object detection requires a single perturbation to suppress multiple objects simultaneously. However, Fig. 3(c) shows that object-level gradients are only weakly aligned: their average cosine similarity is 0.17, with many object pairs exhibiting near-zero or negative similarity. Therefore, an update that suppresses one object may provide little benefit to, or even interfere with, the suppression of another object.

Design implications. These observations suggest two requirements for transferable detection attacks. First, attack knowledge should be learned across samples rather than stored in the local gradient trajectory of each image–surrogate pair. DASH therefore uses an input-conditioned generator as a shared parameterized carrier: diverse training images jointly update the same generator parameters, converting image-specific attack signals into reusable knowledge and mitigating overfitting to local surrogate gradients. Second, the attack objective should coordinate multiple objects without introducing competing category-specific directions. Accordingly, class-agnostic confidence suppression alleviates inter-object conflict, while detector-external CLIP supervision reduces dependence on surrogate-specific representations.

3.2 Distribution-Level Adversarial Transformation

Let $(x, y) \sim P_{\mathrm{nat}}$, where $y = \{(b_i, c_i)\}_{i=1}^{M}$, and let F_s be an accessible surrogate detector. Conventional image-wise attacks independently optimize a perturbation δ_x for every input, whereas universal attacks reuse one fixed perturbation for all inputs. The former can overfit the local gradient of an image–surrogate pair, while the latter cannot adapt its perturbation to image content. In particular, image-wise optimization solves

$$\delta_x^* = \arg\min_{\|\delta\|_\infty \le \epsilon} \ell_{\mathrm{atk}}\big(F_s(x+\delta), y\big), \tag{2}$$

so the learned attack knowledge is confined to the temporary variable δ_x^* and cannot be reused for another input.

DASH instead learns a shared input-conditioned transformation T_θ:

$$\begin{aligned} x_{\mathrm{adv}} &= T_\theta(x) = \Pi_{\mathcal{B}_\infty(x,\epsilon)\cap[0,1]^d}\big(\widetilde{T}_\theta(x)\big), \qquad \delta_\theta(x) = T_\theta(x) - x, \\ P^\theta_{\mathrm{adv}} &= (T_\theta)_{\#} P_{\mathrm{nat}}. \end{aligned} \tag{3}$$

Here, $(T_\theta)_{\#} P_{\mathrm{nat}}$ is the adversarial distribution induced by the transformation, rather than an explicitly estimated density. Different inputs may receive different residuals, but the shared parameters

θ accumulate attack signals over the training distribution. DASH therefore stores reusable attack knowledge in the generator instead of an independently optimized perturbation or a fixed universal template.

3.3 Class-Agnostic Suppression for Inter-Object Conflict

Category-dependent objectives can assign different optimization targets to different objects, causing their gradients to compete. Since our goal is object disappearance rather than targeted misclassification, we replace these targets with a common objective: suppress the strongest foreground response around every object. Let $F_s(x_{\text{adv}}) = \{(\widehat{b}_n, o_n, \mathbf{p}_n)\}_{n=1}^{N}$ denote the differentiable predictions of F_s. We define

$$s_n = o_n \max_c p_{n,c}, \qquad \mathcal{L}_{\text{CA}} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|\mathcal{R}_i|}\sum_{n\in\mathcal{R}_i} s_n, \tag{4}$$

where $o_n = 1$ for detectors without an explicit objectness branch. To avoid relying on a detector-specific matching rule, we construct the candidate set using only box geometry and foreground confidence. Let $\mathcal{C}_i$ contain predictions whose centers lie in an enlarged version of b_i or whose IoU with b_i exceeds a small threshold. We use $\pi_i^{\text{iou}}(k)$ to denote the index of the prediction with the k-th largest IoU to b_i, and $\pi_i^{s}(k)$ for the index with the k-th largest foreground score in $\mathcal{C}_i$. The selected set is

$$\mathcal{R}_i = \left\{\pi_i^{\text{iou}}(k)\right\}_{k=1}^{K_{\text{iou}}} \cup \left\{\pi_i^{s}(k)\right\}_{k=1}^{K_s}. \tag{5}$$

Combining the two criteria retains spatially related predictions even when the attack has already reduced their confidence. The maximum in s_n follows the strongest current category even if the predicted label changes. Consequently, all objects share the same suppression goal and contribute equally, directly targeting the category competition underlying inter-object gradient conflict.

3.4 CLIP Disruption for Cross-Detector Transfer

Although $\mathcal{L}_{\text{CA}}$ coordinates multiple objects, its gradient still originates from F_s and may encode surrogate-specific proposal and prediction mechanisms. We therefore introduce a frozen CLIP visual encoder as detector-external supervision. Let $\mathbf{h}_j(x)$ be its j-th patch token and m_j indicate whether that patch belongs to an annotated object region. The object-region feature discrepancy is

$$\mathcal{D}_{\text{CLIP}} = \frac{1}{\sum_j m_j}\sum_{j=1}^{P} m_j\left(1 - \frac{\mathbf{h}_j(x)^\top \mathbf{h}_j(x_{\text{adv}})}{\|\mathbf{h}_j(x)\|_2\|\mathbf{h}_j(x_{\text{adv}})\|_2}\right). \tag{6}$$

Using patch tokens instead of a single global embedding allows the loss to focus on annotated objects, preventing unrelated background regions from dominating the semantic disruption signal. The clean features are detached and the CLIP encoder remains frozen. DASH jointly minimizes detector confidence and maximizes this detector-external feature discrepancy:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim P_{\text{nat}}}\left[\mathcal{L}_{\text{CA}} - \lambda_{\text{clip}}\mathcal{D}_{\text{CLIP}}\right]. \tag{7}$$

Thus, the two objectives have distinct roles: $\mathcal{L}_{\text{CA}}$ alleviates conflict among objects, whereas $\mathcal{D}_{\text{CLIP}}$ discourages exclusive reliance on the surrogate detector's representation.

3.5 GAN and Diffusion Instantiations

The formulation above specifies what transformation is learned without tying it to a particular generator. We instantiate it independently with two architectures; they share the perturbation constraint and the objective in Eq. equation 7, but are not connected or jointly optimized.

DASH-GAN. A conditional residual generator directly predicts a bounded, content-dependent perturbation:

$$T_\phi^{\text{GAN}}(x) = \text{clip}_{[0,1]}\left(x + \epsilon\tanh(G_\phi(x))\right). \tag{8}$$

A discriminator regularizes local appearance during training. It is combined with the shared attack objective as

$$\mathcal{L}_G^{\text{GAN}} = \mathcal{L}_{\text{CA}} - \lambda_{\text{clip}}\mathcal{D}_{\text{CLIP}} - \lambda_{\text{gan}}\mathbb{E}[D_\psi(x_{\text{adv}})], \tag{9}$$

where the final term encourages the local patch statistics of adversarial images to remain close to those of natural images. The discriminator is discarded at inference, when one generator forward pass produces x_{adv}.

DASH-SD. We encode the image into the latent space of Stable Diffusion, partially invert it, and adapt the frozen U-Net with spatially gated LoRA(Hu et al., 2022):

$$z_s = \mathrm{DDIMInv}(\mathcal{E}(x), s), \qquad W'(h) = W_0 h + M \odot \frac{\alpha}{r} BAh,$$
$$T^{\mathrm{SD}}_{\Delta\omega}(x) = \Pi_{\mathcal{B}_\infty(x,\epsilon)\cap[0,1]^d}\left(\mathcal{V}\big(\mathrm{DDIM}_{\omega_0+\Delta\omega}(z_s)\big)\right), \tag{10}$$

where $\mathcal{E}$ and $\mathcal{V}$ are the frozen VAE, W_0 is a frozen attention projection, and M is an object-region gate obtained from clean image boxes. Only the LoRA parameters $\Delta\omega = \{A, B\}$ are updated. Hence, DASH-GAN realizes the shared transformation through one-step residual generation, whereas DASH-SD realizes it through an adapted denoising trajectory. In both cases, the CA–CLIP losses supervise generator training; no image-wise gradient optimization is performed at test time.

4 Experiments

4.1 Experiments Setup

Datasets. We train DASH on 11,987 images from the VOC2012(Everingham et al., 2010) training set, enabling the generator to learn an input-conditioned mapping from the natural image distribution to an adversarial image distribution. We then evaluate its attack performance on 1,713 images from the VOC2012 test set. To further assess its generalization to unseen data distributions, we evaluate cross-dataset attack transferability on INRIA Person(Dalal & Triggs, 2005) and COCO-500(Lin et al., 2014) without any additional training or fine-tuning. All images are first padded to a square aspect ratio and then resized to 640×640. The data splits and source code used in our experiments will be released in a public open-source repository upon acceptance.

Surrogate and victim detectors. We use YOLOv5, YOLO11s, and Faster R-CNN (F-RCNN) as surrogate detectors for training. The resulting attacks are evaluated on 13 representative victim detectors from four architectural categories: one-stage detectors, including YOLOv2, YOLOv5, YOLOv8, YOLO11s, SSD, VFNet, and TOOD; end-to-end detectors, including YOLOv10 and YOLO26s; two-stage detectors, including F-RCNN and Mask R-CNN; and Transformer-based detectors, including DETR and Deformable DETR (D-DETR). This diverse model suite enables a comprehensive evaluation of the cross-detector and cross-architecture black-box transferability of our method.

Baselines. We compare DASH with six representative object detection attacks: DAG(Xie et al., 2017), UEA(Wei et al., 2019), TOG(Chow et al., 2020), LGP(Li et al., 2024), CAT(Cai et al., 2022), and AFOG(Yahn et al., 2025). We uniformly set the perturbation budget to $16/255$ but retain each method's original surrogate model and remaining settings. This is because several baselines rely on detector-specific components or objectives, and forcing a common surrogate would require modifying their original designs and could unfairly weaken their performance.

Evaluation metrics. We use object-level attack success rate (ASR) as the primary metric and $\mathrm{mAP50:95}$ as a complementary metric. Object-level ASR is computed over objects correctly detected in clean images; an attack succeeds if an object vanishes, is misclassified, or is incorrectly localized. For ablation studies, we adopt a stricter image-level ASR, where an image is counted as successfully attacked only if all initially detected objects fail. This prevents improvements from being dominated by easily attacked objects and better evaluates multi-object suppression. Higher ASR and lower adversarial $\mathrm{mAP50:95}$ indicate stronger attacks. We additionally report inference time per image and LPIPS to measure generation efficiency and perceptual similarity, respectively.

Implementation details. We set the perturbation budget to $\epsilon = 16/255$ for all experiments. DASH-GAN is trained for 20 epochs while DASH-SD is built on Stable Diffusion v1.4(CompVis & RunwayML, 2022) and optimizes rank-16 LoRA modules for 3,000 steps. The surrogate detector and CLIP ViT-B/16(Radford et al., 2021) encoder remain frozen throughout training. For object-level ASR evaluation, we use a confidence threshold of 0.5 and an IoU threshold of 0.5 for matching predictions to ground-truth boxes, with class agreement also required for a correct detection.

4.2 CROSS-MODEL TRANSFERABILITY

Table 1 evaluates cross-model transfer from three surrogate detectors to 13 heterogeneous victims. DASH-GAN and DASH-SD consistently outperform competing attacks across one-stage, two-stage, end-to-end, and Transformer-based detectors, demonstrating broad cross-architecture transferability. Table 4 isolates the contribution of distribution-level generation by comparing DASH with PGD-CA+CLIP, which uses identical attack losses but optimizes each image independently. Despite strong white-box performance, PGD achieves black-box averages of only 47.54% and 61.92%, with standard deviations of 8.44 and 20.61 under the YOLOv5 and F-RCNN surrogates, respectively. In contrast, DASH-GAN raises the averages to 85.21% and 86.24% while reducing the standard deviations to 4.39 and 2.68. This comparison shows that image-wise PGD overfits the white-box surrogate, whereas the generator-induced distribution shift captures more model-agnostic attack patterns, yielding stronger and more stable black-box transferability. Complementarily, Table 2 shows that DASH achieves the lowest $\mathrm{mAP}_{50:95}$ on three of four representative victims and the second-lowest on Mask R-CNN, further corroborating the object-level ASR results.

4.3 CROSS-DATASET TRANSFERABILITY

Table 5 evaluates the models trained on VOC2012 directly on INRIA Person and COCO-500 without additional training or fine-tuning. DASH maintains the strongest overall attack performance across both unseen datasets and diverse victim architectures, indicating that the learned adversarial mapping generalizes beyond the training distribution. These results further validate the cross-dataset transferability of the proposed distribution-level attack paradigm.

Table 1: Object-level ASR (%) on PASCAL VOC2012. Y5, Y11, and FR denote YOLOv5, YOLO11s, and F-RCNN surrogates. The best and second-best results are shown in bold and underlined, respectively.

Victim	Comparison methods						DASH-SD			DASH-GAN		
	UEA	DAG	LGP	AFOG	CAT	TOG	Y5	Y11	FR	Y5	Y11	FR
One-stage detectors												
YOLOv2	49.40	9.62	19.92	46.14	34.97	25.58	83.75	84.22	74.01	89.35	**95.45**	82.34
YOLOv5	59.62	12.14	28.32	55.78	43.70	33.26	92.37	90.48	80.45	**95.83**	95.62	87.99
YOLOv8	51.36	10.28	22.05	46.43	36.59	27.53	90.95	88.55	78.32	92.66	**94.13**	86.63
YOLO11s	46.87	8.54	21.77	46.94	34.05	26.40	93.15	86.08	74.46	90.34	**96.57**	85.34
SSD	17.19	3.16	6.38	18.14	13.19	11.72	**69.43**	67.43	64.61	23.85	26.30	20.90
VFNet	34.87	15.00	33.81	39.21	64.43	22.80	77.78	80.84	71.45	80.06	**86.93**	82.19
TOOD	35.36	17.99	39.06	42.32	71.26	20.50	81.45	83.89	76.89	87.70	**91.25**	89.83
End-to-end detectors												
YOLOv10	48.21	8.74	20.21	44.21	35.15	25.19	90.79	88.58	77.86	89.94	**94.75**	84.45
YOLO26s	46.13	9.62	20.80	45.88	33.62	24.70	92.64	86.12	72.89	88.36	**96.02**	85.79
Two-stage detectors												
F-RCNN	40.09	21.56	35.78	41.26	69.70	25.82	72.73	81.13	70.00	79.76	**85.44**	84.53
Mask R-CNN	36.21	17.30	37.44	38.66	73.81	23.64	74.83	81.57	69.25	78.48	**86.08**	84.46
Transformer-based detectors												
DETR	46.15	34.06	47.58	53.99	69.37	46.44	74.66	81.80	70.63	88.60	**91.96**	86.74
D-DETR	39.81	16.17	38.47	42.48	65.45	20.80	83.61	84.76	78.96	89.74	**93.49**	90.87
Average	42.41	14.17	28.58	43.18	49.64	25.72	82.93	83.50	73.83	82.66	**87.23**	80.92

Table 2: Representative adversarial detection performance on VOC2012. Lower $\mathrm{mAP}_{50:95}$ indicates a stronger attack. The best and second-best attack results are shown in **bold** and underline, respectively.

Method	YOLOv5	YOLOv10	Mask R-CNN	DETR
Clean	0.547	0.592	0.505	0.553
UEA	0.172	0.365	0.364	0.296
TOG	0.375	0.522	0.441	0.493
CAT	0.303	0.465	**0.064**	0.251
AFOG	0.195	0.349	0.281	0.359
DASH	**0.034**	**0.060**	0.117	**0.050**

Table 3: Transfer attack success rate (ASR, %) against detectors protected by ODDefense. The best and second-best results are shown in bold and underlined, respectively.

Method	Defended victim detector	
	F-RCNN–R50	F-RCNN–ConvNeXt
UEA	19.39	15.38
CAT	8.56	4.78
AFOG	10.53	8.05
DASH-GAN	17.82	16.14
DASH-SD	**31.04**	**25.24**

4.4 TRANSFERABILITY TO DEFENDED DETECTORS

To evaluate DASH against adversarial defenses, we adopt ODDefense (Li et al., 2025), a state-of-the-art robust object detection method. ODDefense combines adversarially pre-trained backbones with adversarial fine-tuning of the detector. As shown in Table 3, DASH maintains the highest attack success rate even when the victim detectors are adversarially trained, consistently outperforming the selected competing methods. These results demonstrate that the effectiveness of DASH does not merely rely on vulnerabilities specific to undefended models. Instead, its learned adversarial

Table 4: Object-level ASR (%) comparison of image-wise PGD-CA+CLIP, DASH-GAN, and DASH-SD under YOLOv5 and F-RCNN surrogates. PGD-CA+CLIP uses the same losses as DASH but directly optimizes image-specific perturbations. Black-box statistics exclude the corresponding surrogate detector.

Victim	YOLOv5 surrogate			F-RCNN surrogate		
	PGD	GAN	SD	PGD	GAN	SD
YOLOv5	**98.60**	95.83	92.37	43.19	**87.99**	80.45
YOLOv10	59.26	89.94	**90.79**	31.44	**84.45**	77.86
F-RCNN	40.83	**79.76**	72.73	**97.74**	84.53	70.00
DETR	56.14	**88.60**	74.66	80.39	**86.74**	70.63
VFNet	38.87	**80.06**	77.78	81.28	**82.19**	71.45
TOOD	42.59	**87.70**	81.45	73.28	**89.83**	76.89
Black-box Avg. ↑	47.54	**85.21**	79.48	61.92	**86.24**	75.46
Black-box Std. ↓	8.44	**4.39**	6.38	20.61	**2.68**	3.80

Table 5: Cross-dataset transfer ASR (%) on INRIA Person and COCO-500. Y5, Y11, and FR denote YOLOv5, YOLO11s, and F-RCNN surrogates. Bold and underlined values indicate the best and second-best results in each row, respectively.

Victim	Comparison methods						DASH-SD			DASH-GAN		
	UEA	DAG	LGP	AFOG	CAT	TOG	Y5	Y11	FR	Y5	Y11	FR
INRIA Person												
YOLOv5	30.44	10.50	19.27	40.78	28.29	24.15	92.56	94.04	80.73	**95.86**	95.62	87.59
YOLO11s	25.34	7.51	16.95	38.98	24.54	23.89	90.38	87.01	65.46	92.25	**97.42**	85.47
VFNet	21.36	12.56	25.83	32.88	56.26	20.91	74.50	80.45	64.40	78.12	**86.86**	80.45
YOLOv10	22.55	6.94	16.13	32.87	22.72	20.21	91.07	90.29	70.77	90.98	**94.36**	83.35
YOLO26s	24.98	6.74	13.91	37.47	22.40	22.06	93.67	84.01	60.70	86.84	**96.59**	83.01
F-RCNN	26.63	17.17	30.54	36.49	64.44	23.51	68.93	79.07	63.51	75.61	**81.76**	80.88
DETR	34.87	32.61	45.86	54.47	62.13	47.53	73.08	80.05	66.54	91.30	**92.57**	88.13
Average	26.60	13.43	24.07	39.13	40.11	26.04	83.46	84.99	67.44	87.28	**92.17**	84.13
COCO-500												
YOLOv5	51.09	11.36	26.19	48.10	36.98	39.97	73.15	82.08	69.30	**91.25**	86.50	78.23
YOLO11s	43.32	9.04	21.95	41.54	29.61	32.37	70.04	78.09	63.98	82.77	**89.45**	75.38
VFNet	30.52	12.48	29.31	36.48	58.23	31.48	56.90	70.74	60.41	69.38	**73.98**	71.33
YOLOv10	42.27	8.78	20.81	39.92	30.16	32.36	70.13	79.90	67.54	82.15	**84.31**	72.71
YOLO26s	43.86	9.64	20.27	42.32	29.50	35.14	71.68	78.01	63.55	82.32	**89.95**	75.64
F-RCNN	35.95	19.00	32.74	39.50	67.48	34.88	54.28	72.98	58.75	71.88	73.91	**75.72**
DETR	43.88	38.35	50.41	56.18	69.88	56.89	57.28	72.54	58.77	81.81	**82.67**	80.27
Average	41.56	15.52	28.81	43.43	45.98	37.58	64.78	76.33	63.19	80.22	**82.97**	75.61

mapping transfers effectively to robustified detectors, indicating stronger resistance to adversarial defenses and more stable attack performance.

4.5 Efficiency and Perceptual Quality

As shown in Table 6, both DASH variants eliminate per-image gradient optimization at test time. DASH-GAN requires only 0.02 seconds per image, ranking second only to UEA while achieving substantially higher ASR, whereas DASH-SD takes 0.75 seconds due to its denoising trajectory but remains faster than all iterative baselines. By amortizing attack generation into model parameters, DASH-GAN prioritizes low-latency generation, while DASH-SD better balances efficiency and perceptual quality. DASH-SD achieves the third-best average LPIPS of 0.136, and DASH-GAN obtains 0.160; both outperform UEA, TOG, CAT, and AFOG, whereas LGP and DAG attain lower LPIPS at the cost of substantially weaker ASR and higher generation time. As visualized in Figs. 4 and 5, both variants lie near the upper-left region and improve ASR over the corresponding competitive baselines by 34.73% and 36.14%, respectively, demonstrating a favorable balance among attack effectiveness, efficiency, and perceptual similarity.

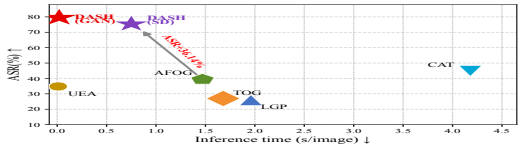


Figure 4: ASR versus inference time; the upper-left region indicates a better effectiveness–efficiency trade-off.

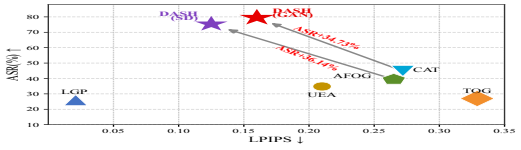


Figure 5: ASR versus LPIPS; the upper-left region indicates a better effectiveness–naturalness trade-off.

Table 6: Efficiency and perceptual quality comparison. "Iterative opt." indicates whether per-image gradient-based optimization is required at test time. Lower inference time and LPIPS indicate better efficiency and perceptual similarity, respectively. The best, second-best, and third-best results are shown in bold, underlined, and italic.

Method	Iterative opt.	Time (s/image) ↓	LPIPS ↓			
			VOC2012	INRIA	COCO-500	Avg.
DAG	✓	7.99	0.090	0.102	0.023	0.072
UEA	✗	**0.01**	0.252	0.158	0.220	0.210
TOG	✓	1.68	0.280	0.300	0.407	0.329
CAT	✓	4.18	0.271	0.283	0.264	0.272
LGP	✓	1.96	0.025	0.017	0.023	**0.021**
AFOG	✓	1.47	0.260	0.265	0.272	0.265
DASH-SD	✗	*0.75*	0.168	0.117	0.125	*0.136*
DASH-GAN	✗	0.02	0.170	0.180	0.130	0.160

4.6 Ablation Studies

Loss ablation. Table 7 shows that the full objective performs best on all six detectors, averaging 68.38% ASR and 0.077 mAP. Removing the class-agnostic or CLIP loss lowers ASR to 43.42% and 55.17%, respectively, showing that the former supplies the main suppression signal while the latter improves transferability. Replacing CLIP with YOLOv5-specific feature disruption reduces ASR to 57.62%, confirming better cross-architecture generalization from CLIP. Replacing the class-agnostic objective with cross-entropy further lowers ASR to 15.28% and raises mAP to 0.414. The two losses are therefore complementary.

Table 7: Ablation study of different loss variants on VOC2012 using YOLOv5 as the surrogate detector. Attack performance is evaluated using image-level ASR and mAP, where higher ASR and lower mAP indicate stronger attacks. The best results are shown in bold.

Variant	YOLOv5		Mask R-CNN		DETR		TOOD		VFNet		YOLOv10s		Average	
	ASR↑	mAP↓	ASR↑	mAP↓	ASR↑	mAP↓	ASR↑	mAP↓	ASR↑	mAP↓	ASR↑	mAP↓	ASR↑	mAP↓
w/o $\mathcal{L}_{\text{det}}$	55.52	0.175	35.25	0.186	34.60	0.126	50.36	0.213	35.08	0.246	49.70	0.258	43.42	0.201
w/o $\mathcal{L}_{\text{clip}}$	72.09	0.101	43.99	0.157	41.17	0.097	58.99	0.177	49.18	0.187	65.62	0.176	55.17	0.149
CE + $\mathcal{L}_{\text{clip}}$	14.20	0.379	14.99	0.459	20.68	0.355	13.65	0.455	13.98	0.423	14.19	0.413	15.28	0.414
$\mathcal{L}_{\text{det}} + \mathcal{L}_{\text{feat}}^{\text{Y5}}$	71.54	0.090	46.94	0.143	44.57	0.086	61.82	0.162	52.29	0.170	68.57	0.161	57.62	0.135
$\mathcal{L}_{\text{det}} + \mathcal{L}_{\text{clip}}$ (Ours)	**86.84**	**0.034**	**55.71**	**0.117**	**54.90**	**0.050**	**69.00**	**0.099**	**61.05**	**0.103**	**82.75**	**0.060**	**68.38**	**0.077**

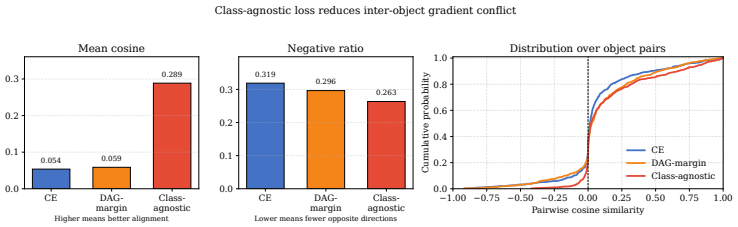


Figure 6: Inter-Object Gradient Consistency under Different Attack Losses

Gradient conflict analysis. Figure 6 compares our loss with two category-dependent objectives: cross-entropy (CE) classification supervision and the detection-specific DAG margin. Our class-agnostic loss raises mean inter-object gradient similarity from 0.054 and 0.059 to 0.289, while reducing negatively aligned object pairs from 31.9% and 29.6% to 26.3%. Removing category-dependent optimization therefore yields more consistent updates and alleviates inter-object gradient conflict.

5 Conclusion

We presented DASH, a distribution-level framework that learns reusable input-conditioned transformations for transferable black-box attacks on object detectors. Under the same formulation, DASH-GAN enables one-pass generation, while DASH-SD exploits the Stable Diffusion prior. Class-agnostic confidence suppression and CLIP feature disruption alleviate multi-object conflicts and surrogate-specific dependence. Experiments on three datasets and thirteen detectors demonstrate strong cross-model and cross-dataset transferability, efficient generation, and sustained effectiveness against adversarially defended detectors.

AI USE STATEMENT

Generative AI tools were used in this work to assist with refining the methodological presentation, providing feedback on experimental design, generating and debugging portions of the research code, drafting, editing, and translating parts of the manuscript, surveying and summarizing related literature, and preparing the manuscript layout. We did not use generative AI to generate or clean datasets, nor did we rely on AI to independently determine, select, or modify the numerical results reported in this paper. Tasks involving surveys, interviews, transcription, or qualitative thematic analysis were not applicable to this study. All AI-assisted code, references, and manuscript text were manually reviewed by the authors. Specifically, the authors verified all citations against the original sources, tested AI-generated code for correctness, and revised the final manuscript to ensure technical accuracy and originality.

ETHICS STATEMENT

DASH can generate natural-looking adversarial examples that transfer across heterogeneous object detectors. We acknowledge that DASH may pose practical threats to object detection systems, particularly those deployed in safety-critical applications. However, investigating these vulnerabilities is essential for developing reliable and robust detectors. This work can raise awareness of the security risks faced by existing object detection models, and its defensive value outweighs the potential risk of misuse. DASH can be used to evaluate the robustness of detectors in real-world applications and to improve their resilience through adversarial training.

REPRODUCIBILITY STATEMENT

To support the reproducibility of this work, we use publicly available datasets and object detectors, and describe the dataset splits, preprocessing procedures, evaluation metrics, perturbation constraints, and training configurations in the main paper. Additional implementation details, complete experimental results, and supplementary analyses are provided in the appendix. Upon acceptance, we will publicly release the source code, trained checkpoints, data splits, and evaluation scripts. These resources are intended to facilitate independent reproduction and verification of our results.

REFERENCES

Hafiz Mughees Ahmad and Afshin Rahimi. Deep learning methods for object detection in smart manufacturing: A survey. *Journal of Manufacturing Systems*, 64:181–196, 2022.

Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019.

Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and M Salman Asif. Context-aware transfer attacks for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 149–157, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Ka-Ho Chow, Ling Liu, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu. Tog: targeted adversarial objectness gradient attacks on real-time object detection systems. *arXiv preprint arXiv:2004.04320*, 2020.

CompVis and RunwayML. Stable diffusion v1-4 model card. `https://huggingface.co/CompVis/stable-diffusion-v1-4`, 2022.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886–893, 2005.

Xinlong Ding, Jiansheng Chen, Hongwei Yu, Yu Shang, Yining Qin, and Huimin Ma. Transferable adversarial attacks for object detection using object-aware significant feature distortion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1546–1554, 2024.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193. IEEE, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4307–4316. IEEE, 2019.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3490–3499. IEEE Computer Society, 2021.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Guopeng Li, Yue Xu, Jian Ding, and Gui-Song Xia. Toward generic and controllable attacks against object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

Xiao Li, Hang Chen, and Xiaolin Hu. On the importance of backbone to the adversarial robustness of object detectors. *IEEE Transactions on Information Forensics and Security*, 20:2387–2398, 2025.

Yuezun Li, Daniel Tian, Ming-Ching Chang, Xiao Bian, and Siwei Lyu. Robust adversarial perturbation on deep proposal-based models. *arXiv preprint arXiv:1809.05962*, 2018.

Quanyu Liao, Xin Wang, Bin Kong, Siwei Lyu, Bin Zhu, Youbing Yin, Qi Song, and Xi Wu. Transferable adversarial examples for anchor free object detection. *arXiv preprint arXiv:2106.01618*, 2021.

Shuo-Yen Lin, Ernie Chu, Po-Hung Yeh, Jun-Cheng Chen, and Jia-Ching Wang. Diffusion to confusion: Naturalistic adversarial patch generation based on diffusion model for object detector. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 2378–2383. IEEE, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pp. 21–37. Springer, 2016.

Boming Miao, Chunxiao Li, Yao Zhu, Weixiang Sun, Zizhe Wang, Xiaoyi Wang, and Chuanlong Xie. Advlogo: Adversarial patch attack against object detectors based on diffusion models. *arXiv preprint arXiv:2409.07002*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

GSDMA Sreenu and Saleem Durai. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*, 6(1):1–27, 2019.

Ultralytics. Yolov5 by ultralytics. `https://github.com/ultralytics/yolov5`, 2020.

Yang Wang, Yunfei Zheng, Lei Chen, Zhen Yang, and Tieyong Cao. Dib-uap: enhancing the transferability of universal adversarial perturbation via deep information bottleneck. *Complex & Intelligent Systems*, 10(5):6825–6837, 2024.

Zhixiang Wang, Xingjun Ma, and Yu-Gang Jiang. Badpatch: Diffusion-based generation of physical adversarial patches. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6244–6254, 2025.

Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 954–960, 2019.

Xing Wu, Lifeng Huang, Chengying Gao, WS Lee, and T Suzuki. G-uap: Generic universal adversarial perturbation that fools rpn-based detectors. In *ACML*, pp. 1204–1217, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1369–1378, 2017.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2725–2734. IEEE, 2019.

Zachary Yahn, Selim Furkan Tekin, Fatih Ilhan, Sihao Hu, Tiansheng Huang, Yichang Xu, Margaret Loper, and Ling Liu. Adversarial attention perturbations for large object detection transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3184–3193, 2025.

Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8514–8523, 2021.

Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *European conference on computer vision*, pp. 471–486. Springer, 2018.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A TRAINING PSEUDOCODE

Algorithms 1 and 2 summarize the training procedures of the two DASH instantiations. Both optimize the generator using class-agnostic confidence suppression and CLIP feature disruption under the same ℓ_∞ constraint, while keeping the surrogate detector and CLIP encoder frozen. DASH-GAN alternately trains a conditional residual generator and a discriminator, whereas DASH-SD updates only the LoRA parameters during adversarial denoising. Thus, the two variants share the same distribution-level attack objective while realizing it through different generative mechanisms.

Algorithm 1: Training DASH-SD

Input: Training set $\mathcal{S}$; frozen proxy detector F; frozen CLIP encoder Ψ; pretrained SD encoder E, decoder D, and U-Net ϵ_{θ_0}; perturbation budget ε; CLIP weight λ_{clip}; inversion and generation steps $s_{\text{for}}, s_{\text{gen}}$.
Output: Trained LoRA parameters $\Delta\theta$.
Init: Insert LoRA modules $\Delta\theta$ into the attention layers of ϵ_{θ_0} and freeze all remaining parameters.

$c \leftarrow E_{\text{text}}(\text{“”})$; initialize AdamW for $\Delta\theta$;
; // Adversarial LoRA fine-tuning
for (x, y) *in* S **do**
  $z_{s_{\text{for}}} \leftarrow \text{DDIMInv}(E(x), \epsilon_{\theta_0}, c, s_{\text{for}})$;
  $h_x \leftarrow \text{CLIP}(\Psi, x, y)$ without gradients;
  for t *in* s_{gen} **do**
    $(z_{t-1}, \hat{z}_0^{(t)}) \leftarrow \text{DDIMStep}(z_t, t, \epsilon_{\theta_0+\Delta\theta}, c)$;
    $\tilde{x}_{\text{adv}} \leftarrow D(\hat{z}_0^{(t)})$;
    $x_{\text{adv}} \leftarrow \Pi_{\mathcal{B}_\infty(x,\varepsilon)}(\tilde{x}_{\text{adv}})$;
    $\mathcal{L}_{\text{det}} \leftarrow \mathcal{L}_{\text{CA}}(F(x_{\text{adv}}), y)$;
    $d_{\text{clip}} \leftarrow d(\text{CLIP}(\Psi, x_{\text{adv}}, y), h_x)$;
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{det}} - \lambda_{\text{clip}} d_{\text{clip}}$;
    Update $\Delta\theta$ using $\text{AdamW}(\nabla_{\Delta\theta}\mathcal{L})$;
return $\Delta\theta$;

Algorithm 2: Training DASH-GAN

Input: Training set $\mathcal{S}$; frozen proxy detector F; frozen CLIP encoder Ψ; perturbation budget ε; weights λ_{clip} and λ_{gan}.
Output: Trained conditional residual generator G_ϕ.
Init: Initialize generator G_ϕ, discriminator D_ω, and their Adam optimizers; freeze F and Ψ.

for $e = 1$ **to** N_{epoch} **do**
  for (x, y) *in* S **do**
    $r_\phi(x) \leftarrow \tanh(G_\phi(x))$;
    $x_{\text{adv}} \leftarrow \text{clip}(x + \varepsilon r_\phi(x), 0, 1)$;
    ; // Update discriminator
    $\mathcal{L}_D \leftarrow \mathbb{E}[\max(0, 1 - D_\omega(x))] + \mathbb{E}[\max(0, 1 + D_\omega((x_{\text{adv}})))]$;
    Update ω using $\text{Adam}(\nabla_\omega \mathcal{L}_D)$;
    ; // Update the adversarial generator
    $h_x \leftarrow \text{CLIP}(\Psi, x, y)$ without gradients;
    $\mathcal{L}_{\text{det}} \leftarrow \mathcal{L}_{\text{CA}}(F(x_{\text{adv}}), y)$;
    $d_{\text{clip}} \leftarrow d(\text{ObjectCLIP}(\Psi, x_{\text{adv}}, y), h_x)$;
    $\mathcal{L}_{\text{gan}} \leftarrow -\mathbb{E}[D_\omega(x_{\text{adv}})]$;
    $\mathcal{L}_G \leftarrow \mathcal{L}_{\text{det}} - \lambda_{\text{clip}} d_{\text{clip}} + \lambda_{\text{gan}} \mathcal{L}_{\text{gan}}$;
    Update ϕ using $\text{Adam}(\nabla_\phi \mathcal{L}_G)$;
return G_ϕ;

B SENSITIVITY TO THE CONFIDENCE THRESHOLD

Figure 7 reports ASR under confidence thresholds ranging from 0.3 to 0.7 across six representative victim detectors. As expected, ASR generally increases with the threshold because low-confidence detections are more readily suppressed. Nevertheless, the relative performance remains consistent: DASH-GAN achieves the highest ASR across nearly all detector–threshold combinations, while DASH-SD also consistently outperforms the competing methods. In particular, DASH maintains

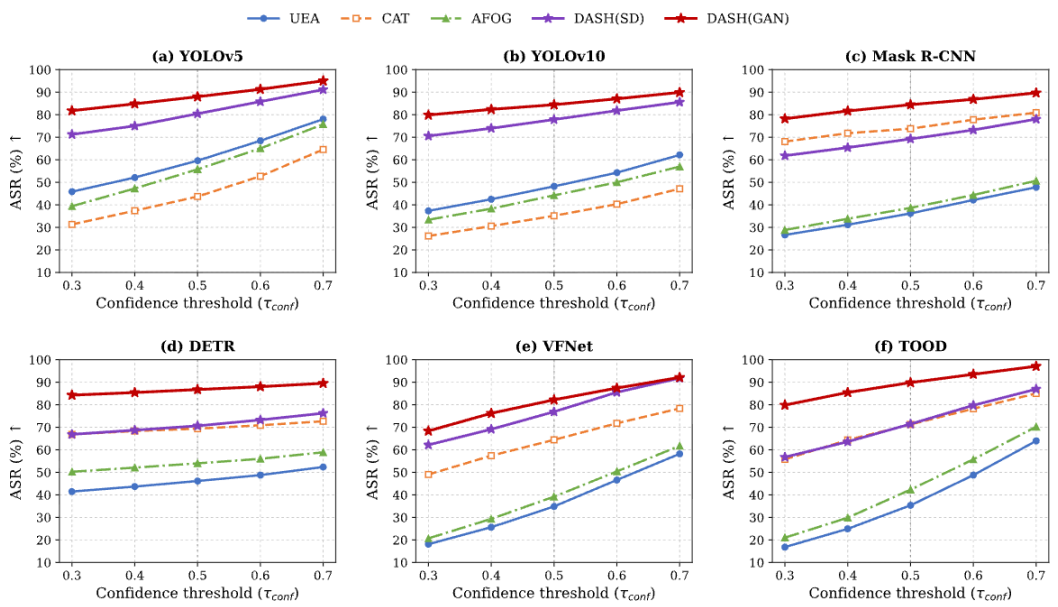


Figure 7: ASR under different detection confidence thresholds.

strong attack performance even at the strict threshold of 0.3. These results show that the advantage of DASH is robust to the confidence-threshold selection rather than being tied to a specific evaluation setting.

C Breakdown of Attack Success Types

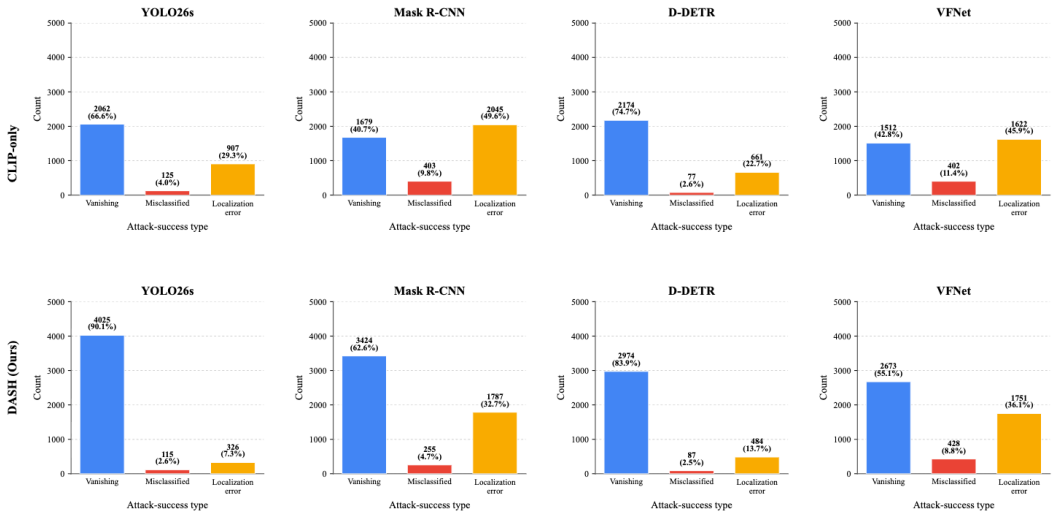


Figure 8: Breakdown of successful attack types for CLIP-only supervision and the complete DASH objective across representative victim detectors. Numbers above the bars report counts and their corresponding proportions.

Figure 8 categorizes successful attacks into object vanishing, misclassification, and localization error. Compared with CLIP-only supervision, the complete DASH objective substantially increases both the number and proportion of vanishing objects across all four victim detectors, while reducing the relative contribution of localization errors. This shift is particularly evident on YOLO26s, where the proportion of vanishing attacks increases from 66.6% to 90.7%. These results show that the class-agnostic confidence suppression loss complements CLIP supervision by directly weakening object detectability, making complete object disappearance the dominant failure mode.

D Additional Experimental Results

Table 8: Object-level ASR (%) of DASH-SD trained with YOLOv5 as the surrogate detector under different ℓ_∞ perturbation budgets. The shaded columns correspond to white-box evaluations, while the remaining columns report black-box transfer results.

ϵ	VOC2012						COCO-500					
	YOLOv5	**YOLOv10**	**F-RCNN**	**DETR**	**VFNet**	**TOOD**	**YOLOv5**	**YOLOv10**	**F-RCNN**	**DETR**	**VFNet**	**TOOD**
$10/255$	67.84	62.06	50.63	52.58	50.30	56.30	54.33	51.57	41.79	42.10	40.19	43.45
$12/255$	77.50	73.30	59.97	61.76	59.87	67.18	65.11	63.32	49.05	50.12	49.29	54.70
$16/255$	92.37	90.79	72.73	74.66	77.78	81.45	73.15	70.13	54.28	57.28	56.90	62.60

Effect of the Perturbation Budget

Analysis. Increasing the perturbation budget consistently improves both white-box performance on YOLOv5 and black-box transferability to the other detectors. The same trend is observed on COCO-500 without additional training, indicating that a larger feasible perturbation region strengthens not only in-domain attacks but also cross-dataset transfer. We use $\epsilon = 16/255$ in the main experiments because it provides the strongest transferable attack performance under the adopted perturbation constraint.

Table 9: Object-level ASR (%) comparison of image-wise PGD-CA+CLIP, DASH-SD, and DASH-GAN under three surrogate detectors. PGD-CA+CLIP uses the same class-agnostic and CLIP objectives as DASH but optimizes each input independently. Gray cells indicate white-box evaluations in which the victim and surrogate detectors are identical. The averages are computed over all 13 victim detectors.

Victim	**YOLOv5 surrogate**			**YOLO11s surrogate**			**F-RCNN surrogate**		
	DASH-SD	**PGD**	**DASH-GAN**	**DASH-SD**	**PGD**	**DASH-GAN**	**DASH-SD**	**PGD**	**DASH-GAN**
YOLOv5	92.37	98.60	95.83	90.48	51.43	95.62	80.45	43.19	87.99
YOLOv2	83.75	41.92	89.35	84.22	33.73	95.45	74.01	27.21	82.34
YOLOv8	90.95	66.06	92.66	88.55	55.20	94.13	78.32	34.35	86.63
YOLOv10	90.79	59.26	89.94	88.58	53.34	94.75	77.86	31.44	84.45
YOLO11s	93.15	57.57	90.34	86.08	99.30	96.57	74.46	29.54	85.34
YOLO26s	92.64	51.40	88.36	86.12	59.74	96.02	72.89	30.56	85.79
F-RCNN	72.73	40.83	79.76	81.13	31.08	85.44	70.00	97.74	84.53
Mask R-CNN	74.83	38.47	78.48	81.57	28.87	86.08	69.25	79.00	84.46
VFNet	77.78	38.87	80.06	80.84	28.90	86.93	71.45	81.28	82.19
DETR	74.66	56.14	88.60	81.80	50.92	91.96	70.63	80.39	86.74
D-DETR	83.61	42.78	89.74	84.76	30.92	93.49	78.96	81.73	90.87
SSD	69.43	14.67	23.85	67.43	11.57	26.30	64.61	11.49	20.90
TOOD	81.45	42.59	87.70	83.89	31.51	91.25	76.89	73.28	89.83
Average	**82.93**	49.93	82.66	83.49	43.58	**87.23**	73.82	53.93	**80.92**

Full Comparison under Different Surrogate Detectors

Analysis. PGD-CA+CLIP achieves strong white-box ASR when the victim matches its surrogate, but its average performance drops substantially on heterogeneous detectors. In contrast, both DASH variants maintain considerably higher averages under all three surrogates. Since the three methods use the same attack objectives, this comparison isolates the role of the shared generator: encoding supervision from multiple training images into common parameters reduces image-specific surrogate overfitting and produces more reusable black-box attack knowledge.

E VISUALIZATION RESULTS

Figures 9–11 present attack results on VOC2012 and two unseen datasets, INRIA Person and COCO-500. AFOG and CAT frequently retain correct, high-confidence detections across the YOLO victims, indicating inconsistent cross-architecture transfer. In comparison, the two DASH variants suppress more target objects or induce incorrect localization across heterogeneous detectors. Notably, DASH maintains this behavior on INRIA Person and COCO-500 without additional training or fine-tuning, qualitatively demonstrating that the generator encodes reusable attack knowledge that generalizes beyond the training distribution. Among the selected examples, DASH-GAN exhibits more consistent suppression, while DASH-SD produces structured content-dependent changes with occasional residual detections.



Figure 9: Qualitative cross-detector attack results on VOC2012.

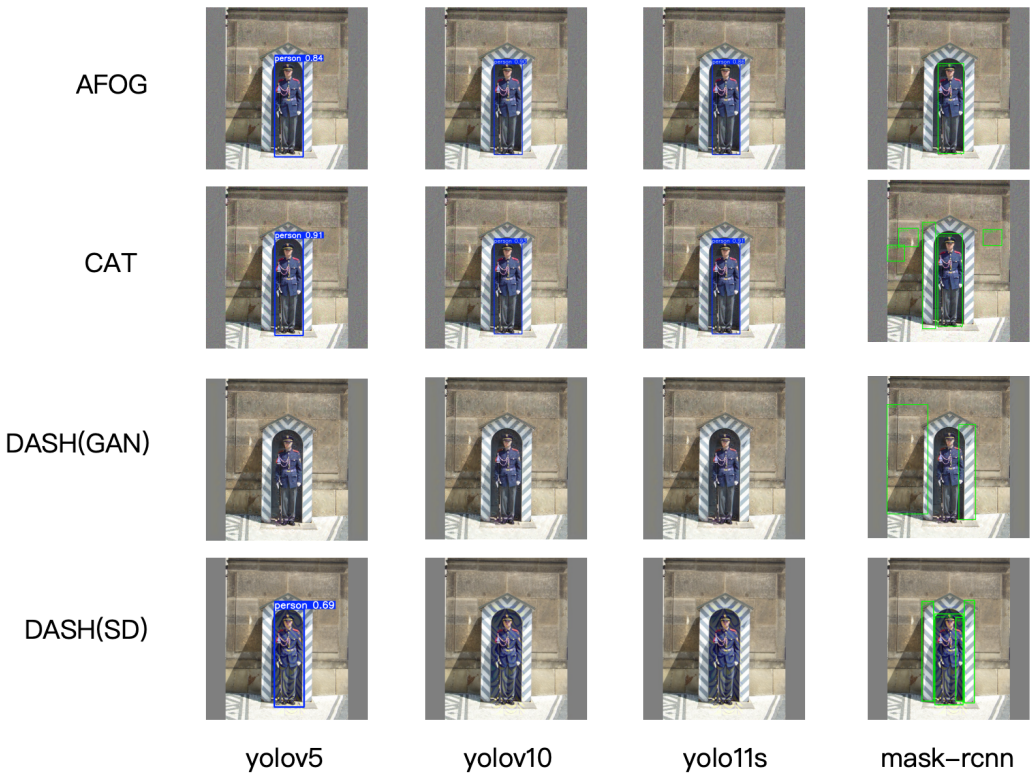


Figure 10: Zero-shot cross-dataset attack results on INRIA Person.

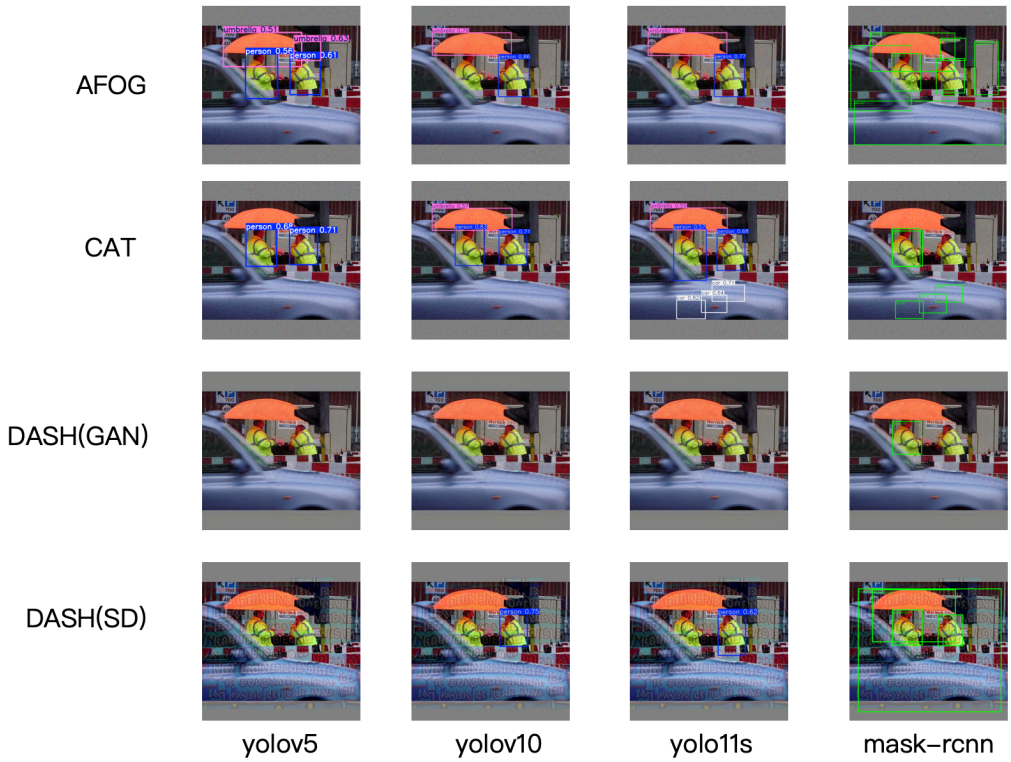


Figure 11: Zero-shot cross-dataset attack results on COCO-500.