Training-Free Traffic-Sign Detection Recovery via Open-Vocabulary Adversarial-Patch Localization

Kangmin Bae, Hyunwoo Jin, and Sungwon Yi

ETRI, South Korea
{kmbae,maizer,sungyi}@etri.re.kr

Abstract. Removing an adversarial patch does not necessarily restore the predictions of an object detector. In traffic-sign images, a patch can occlude the class-defining symbol; a perceptually plausible completion may therefore depict the wrong sign, while unconstrained recovery may overwrite predictions that survived the attack. We study this distinction on REAP and propose a training-free recovery framework for settings in which detector retraining is unavailable. A frozen open-vocabulary detector first proposes traffic-sign regions and then localizes patch-like subregions within them, without access to patch size, count, or placement. The resulting boxes define masks for native-resolution local inpainting. Detections from the recovered image and contextual crops are fused asymmetrically with the immutable attacked-image output, while a frozen visual encoder supplies ranked class hypotheses from clean references and surviving source content. No component is updated. The framework thus separates patch localization, visual restoration, semantic recovery, and prediction preservation without relying on privileged benchmark geometry.

Keywords: Adversarial patches · Traffic sign detection · Training-free recovery · Image inpainting

1 Introduction

Adversarial patches translate the vulnerability of neural networks into a physically realizable threat. A printed pattern attached to an object can remain effective across changes in viewpoint, scale, and illumination [2, 4]. For traffic-sign perception, the consequence is not limited to an incorrect class label: a patch may suppress the sign proposal altogether, thereby removing both the bounding box and class prediction from a downstream driving system. REAP [5] makes this failure measurable at scale by rendering physically parameterized patches onto signs in diverse Mapillary scenes [3] and evaluating complete object detectors.

Adversarial training is the strongest baseline reported by REAP, but it presupposes access to the detector's training procedure, parameters, and sufficient attack data. These requirements are incompatible with deployed or externally supplied detectors whose weights cannot be modified. Input-space recovery is attractive in this regime because it can be inserted before a frozen detector.

The relevant objective, however, is easy to misstate. Conventional inpainting is designed to produce a plausible image, whereas detector recovery must restore class-defining visual content, preserve uncorrupted pixels, and avoid deleting predictions that already survived the attack. A generated crop can look like a traffic sign while changing its speed value or pictogram; even resizing a full frame through a generative pipeline can alter small-object features outside the mask.

We therefore formulate the problem as training-free traffic-sign detection recovery through open-vocabulary adversarial-patch localization. Given only an attacked image and the immutable output of the target model, the proposed framework first identifies candidate traffic signs and then localizes patch-like regions within them. Only the predicted patch regions are modified, while detections that survive the attack are retained as an immutable base set. A fixed visual encoder and a small clean reference gallery provide ranked class hypotheses when image recovery cannot reconstruct the occluded sign semantics. During recovery, the method accesses neither test annotations nor ground-truth sign classes, patch dimensions or placement, oracle patch masks, or renderer transforms. We evaluate the framework in a post-attack setting, applying recovery after the adversarial input has been generated and measuring its effect on the target model without modifying detector parameters. Overall, this work presents a training-free pipeline that connects open-vocabulary patch localization, localized image recovery, and prediction fusion, and analyzes how each stage affects downstream detection under the REAP protocol.

2 Related Work

2.1 Adversarial Attacks and Defenses

Printable patch attacks progressed from universal image-classification attacks [2] to physically robust traffic-sign perturbations [4] and detector attacks that disrupt both localization and classification [10, 24]. Their evaluation likewise moved from isolated demonstrations to natural and reproducible benchmarks: APRICOT captures printed adversarial targets under real capture variation [1], while TT100K [27] and Mapillary [3] provide large-scale in-the-wild traffic-sign data. REAP augments Mapillary annotations with sign-aware perspective, placement, and relighting models for differentiable patch rendering [5]. We adopt its threat specification and evaluator but study inference-time recovery rather than attack construction.

Patch defenses broadly pursue empirical localization and removal or certified robustness. Segment and Complete couples a trained patch segmenter with shape completion [8]; PatchZero learns adversarial-pixel localization before replacement [25]; Jedi combines entropy proposals with learned refinement [18]; and PAD localizes and removes patches without additional training or appearance priors [6]. Certified methods instead restrict or mask patch influence: PatchGuard uses small receptive fields and robust feature masking [20], PatchCleanser applies two-round input masking [21], ObjectSeeker extends masking guarantees to object detection [22], and PatchCURE studies the resulting robustness–

utility–cost trade-off [23]. Our method instead targets empirical inference-time recovery, using frozen open-vocabulary localization and fusing recovered predictions without deleting detections that survive the attack.

2.2 Image Inpainting and Open-Vocabulary Detection

Classical fast-marching inpainting propagates nearby structure into a mask without learned parameters [19]. Learned approaches use larger semantic priors: LaMa expands the effective receptive field with Fourier convolutions [17], RePaint conditions a pretrained diffusion process on observed pixels [11], and latent diffusion and ControlNet enable high-quality, structurally conditioned generation [16, 26]. These methods are generally evaluated for reconstruction or perceptual quality. Traffic-sign recovery imposes a different criterion because a visually coherent completion may not preserve a fine-grained class symbol. This mismatch motivates our conservative native-resolution completion path and a separate source of class information.

Vision-language pretraining enables recognition beyond a fixed supervised vocabulary. CLIP learns transferable image-text representations at image level [14]; GLIP unifies detection and phrase grounding [7]; and OWL-ViT transfers contrastive image-text pretraining to open-vocabulary detection [12]. Grounding DINO tightly fuses language and detector features for open-set localization [9]. We use these models only as frozen infrastructure: open-vocabulary grounding proposes sign and patch regions, while visual encoders compare the resulting crops with clean references. Neither component is adapted to REAP or optimized on adversarial patches.

3 Method

3.1 Problem Formulation

Let $(x, y) \sim \mathcal{D}$ denote a clean image and its detection annotations, and let D_f be the fixed target model. Following REAP, an attack $A \in \mathcal{A}$ constructs an attacked image and the corresponding base detections as

$$x_{\text{adv}} = A(x, y; D_f), \qquad \mathcal{P}_{\text{adv}} = D_f(x_{\text{adv}}). \tag{1}$$

Recovery receives only x_{adv} and $\mathcal{P}_{\text{adv}}$, without annotations, and retains the attacked-image predictions as an immutable base:

$$\mathcal{P}_{\text{adv}} \subseteq \mathcal{P}_{\text{rec}}. \tag{2}$$

Recovery additionally uses a frozen open-vocabulary detector D_o, a frozen visual encoder E, and a clean reference gallery $\mathcal{R} = \bigcup_{c=1}^{C} \mathcal{R}_c$ over C target classes. The detector is conditioned on queries τ_s and τ_p for sign and patch localization. The overall operator is

$$\mathcal{P}_{\text{rec}} = \mathcal{F}(x_{\text{adv}}, \mathcal{P}_{\text{adv}}; D_f, D_o, E, \tau_s, \tau_p, \mathcal{R})\,. \tag{3}$$

At test time, $\mathcal{F}$ infers patch locations directly from x_{adv}; its gallery and hyperparameters are fixed on development, and D_f, D_o, and E remain unchanged.

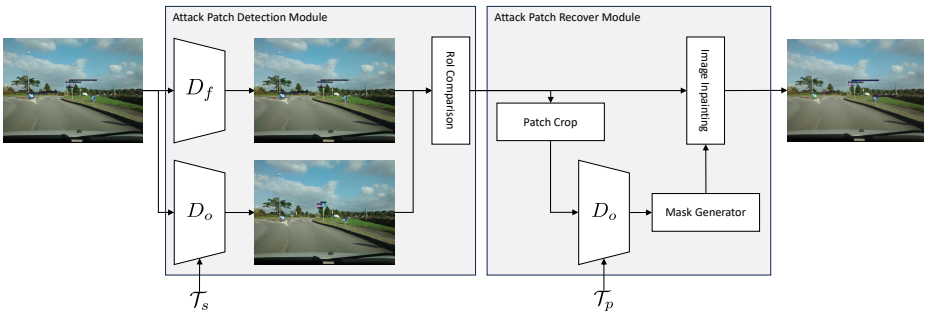

Fig. 1: Overview of the proposed training-free detection recovery framework. The target model D_f produces base detections from the attacked image, while the open-vocabulary detector D_o hierarchically localizes candidate sign regions and patch-like subregions using τ_s and τ_p. Predicted patch boxes define native-resolution recovery masks. The recovered image, contextual crops, and a fixed visual encoder provide complementary candidates, which are fused with the base detections without discarding predictions that survive the attack.

3.2 Attack Patch Detection Module

Given x_{adv}, D_o first localizes candidate traffic-sign regions using τ_s:

$$\mathcal{B}_o = D_o(x_{\text{adv}}; \tau_s), \tag{4}$$

where filtering and non-maximum suppression are implicit. A proposal enters $\mathcal{B}_s \subseteq \mathcal{B}_o$ if it lacks a confident class-specific match in $\mathcal{P}_{\text{adv}}$ or overlaps only the generic *other* class. We evaluate Eq. (4) under primary and finer support partitions; cross-partition agreement retains supported hypotheses, downweights unsupported ones, and conservatively admits support-only proposals.

Patch localization is conditioned on suspicious sign proposals rather than performed over the full image. For each $B \in \mathcal{B}_s$, a second grounding step operates in the normalized crop domain:

$$\mathcal{C}_{p,B} = D_o(x_{\text{adv}}[B]; \tau_p), \qquad \widehat{\mathcal{B}}_p = \bigcup_{B \in \mathcal{B}_s} \pi_B(\Phi_p(\mathcal{C}_{p,B}))\,. \tag{5}$$

Here Φ_p applies compatibility, confidence, compactness, cross-query agreement, and non-maximum suppression, while π_B restores image coordinates. The contextual set $\widetilde{\mathcal{B}}$ augments $\mathcal{B}_o$ with confident boxes from $\mathcal{P}_{\text{adv}}$ for recovery, but the mask depends only on Eq. (5) and may be empty.

3.3 Attack Patch Recovery Module

The localized boxes define a binary mask $\hat{M} = \text{Rasterize}(\widehat{\mathcal{B}}_p)$. Let G be a native-resolution local inpainting operator. Its output is hard-composited with the attacked image:

$$x_{\text{rec}} = \hat{M} \odot G(x_{\text{adv}}, \hat{M}) + (\mathbf{1} - \hat{M}) \odot x_{\text{adv}}. \tag{6}$$

Thus unmasked pixels are copied exactly from x_{adv} and $\hat{M} = \mathbf{0}$ leaves the image unchanged. Applying D_f to x_{rec} and crops in $\widetilde{\mathcal{B}}$ yields image-coordinate candidates $\mathcal{P}_x$.

Inpainting may remove patch texture without restoring the class symbol. For each $B \in \widetilde{\mathcal{B}}$, E compares attacked and recovered crops with $\mathcal{R}_c$ using global similarity S_g and unmasked local-token similarity S_l. Recovered-view retrieval is combined with pre-threshold crop logits from D_f, and only its strongest direct class hypothesis is retained in $\mathcal{P}_{E,\text{rec}}$. The attacked view supplies two complementary outputs: a top-ranked hypothesis is promoted only when semantic, localization, and cross-partition confidence support it, whereas a low-score top-K set remains immediately above the class thresholds as a recall fallback. Their union $\mathcal{P}_E$ separates confident recognition from ambiguity instead of assigning high scores to every plausible class.

The box-regression head of D_f further refines candidate geometry without updating the target model. For a consolidated hypothesis (b, c), let b_v be a matched proposal from partition v and let $r_{v,c}$ denote its class-conditioned regressed box. We retain only views satisfying $\text{IoU}(b, b_v) \geq \mu$ and $\text{IoU}(b_v, r_{v,c}) \geq \eta$, and update

$$b' = (1 - \alpha)b + \alpha \, \text{Mean}_{v \in \mathcal{V}(b,c)} \, r_{v,c}. \tag{7}$$

The highest-overlap views define $\mathcal{V}(b, c)$; if no view passes both gates, b is left unchanged. This conservative update targets the high-IoU localization errors penalized by mAP while preserving the candidate set and its operating-point recall.

Finally, asymmetric fusion retains the base predictions and selectively adds image- and reference-based candidates:

$$\mathcal{P}_{\text{rec}} = \mathcal{P}_{\text{adv}} \cup \{p \in \mathcal{P}_x \cup \mathcal{P}_E : g(p, \mathcal{P}_{\text{adv}}) = 1\}. \tag{8}$$

The gate g admits uncovered objects, refinements of generic detections, and hypotheses exceeding an overlapping base prediction by a margin. Added candidates are deduplicated while every element of $\mathcal{P}_{\text{adv}}$ is retained.

4 Experiments

4.1 Implementation Details

We follow REAP's 54-class Faster R-CNN protocol [5, 15], using its fixed class thresholds at 0.5 IoU for FNR and released mAP evaluator. Grounding DINO queries sign shapes and patches at box/text thresholds 0.15/0.25 and 0.20/0.25; overlapping 2×2 and 3×3 partitions are fused at 0.3 IoU. Telea inpainting [19] uses radius 3 at native resolution. Frozen EVA ViT-G/14 retrieves from ten clean references per class, with target-model crop logits weighted by 0.25. We retain one recovered-view class, promote attacked-view top-1 hypotheses above 0.1, and keep $K = 15$ fallback classes at score scale 0.001. OCR promotes an EVA top-3 speed-limit hypothesis at confidence 0.4 and scale 0.4; a non-octagonal phrase

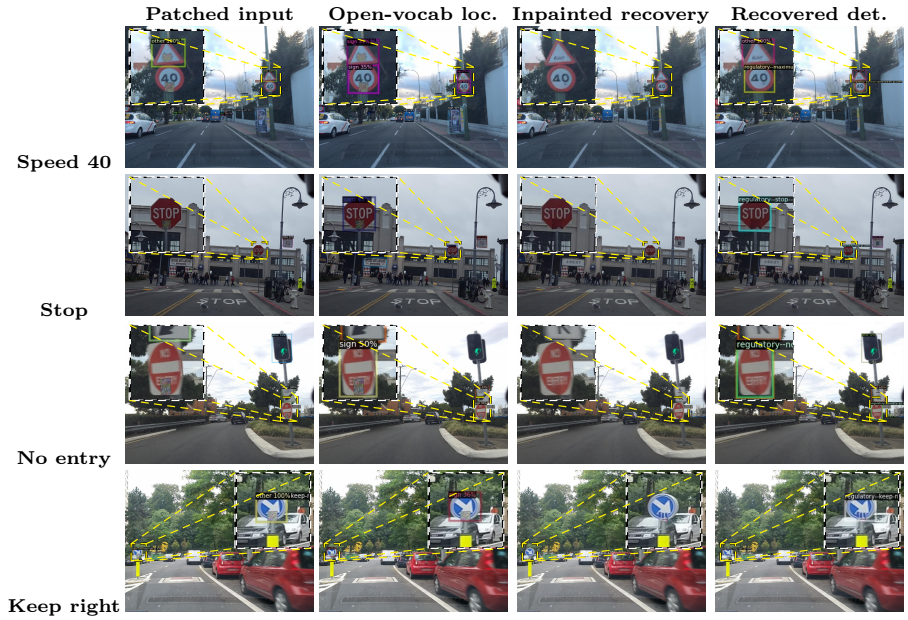

Fig. 2: Representative recovery results across sign classes, scales, and scenes. Each row shows the attacked input, open-vocabulary localization, local inpainting, and recovered detector output; the inset magnifies the dashed target-sign region. The results illustrate recovered detections and initially failed detections.

scales only the above-threshold stop-sign residual by 0.25. Equation (7) averages up to two views with $(\mu, \eta, \alpha) = (0.5, 0.7, 0.3)$. Settings are fixed on the disjoint attack split; test annotations and renderer geometry are never used.

4.2 Qualitative Results

Figure 2 presents representative results across four traffic-sign classes, varying object scales, and diverse scene conditions. In Fig. 2, the hierarchical localization stage identifies a patch-like region within the candidate sign crop. Local inpainting removes the adversarial texture, after which the recovery branches restore target-model detections that were suppressed in the patched input. Together, these examples demonstrate that the framework can recover missing detections while preserving predictions that survive the attack.

4.3 Quantitative Results and Discussion

Table 1 compares our training-free pipeline with the Standard and adversarially trained REAP baselines. Relative to the Standard detector, our method consistently improves both FNR and mAP under patch attacks. Recovery becomes more effective as attack severity increases, with the largest gains under

Table 1: Quantitative results on the REAP [5] test set. Standard and adversarially trained values are reported by REAP; our training-free method is evaluated with the same 54-class Faster R-CNN protocol.

Detector	Patch	Standard		Adv. trained		Ours	
		FNR ↓	mAP ↑	FNR ↓	mAP ↑	FNR ↓	mAP ↑
Faster R-CNN	No patch	4.3	72.9	3.1	73.3	3.3	69.1
Faster R-CNN	Small, 10 in. × 10 in.	15.4	59.4	3.8	71.8	3.8	64.4
Faster R-CNN	Medium, 10 in. × 20 in.	22.4	46.5	6.1	66.8	5.0	58.1
Faster R-CNN	Large, 2 × 10 in. × 20 in.	50.0	18.2	13.9	56.3	13.3	42.5

the two-patch condition. Our method also substantially narrows the gap between the Standard and adversarially trained detectors while leaving the target model unchanged. Its competitive operating-point recall and consistent mAP gains demonstrate that open-vocabulary localization and recovery provide an effective training-free alternative when adversarial retraining is unavailable or impractical. However, the proposed framework is not designed for latency-critical real-time deployment. It relies on computationally demanding components, including an open-vocabulary detector and a large frozen visual encoder, and repeatedly invokes detection and recovery branches for each input. In its current form, therefore, the framework is well suited to offline or post-hoc recovery and verification after a suspected attack, where immediate processing is not required.

5 Conclusion

We studied adversarial-patch removal as a detection-recovery problem rather than a perceptual-editing task. The framework keeps D_f, D_o, and all visual recognizers fixed; preserves unmasked pixels and base detections; and uses no patch size, count, placement, or rendered mask at inference. It consistently lowers FNR relative to Standard and improves both FNR and mAP under the two-patch attack. The remaining gap to adversarial training identifies precise class ranking under severe occlusion as the principal limitation of inference-time recovery. Future work will investigate lightweight visual encoders and efficient open-vocabulary detectors, such as NanoOWL-style alternatives [13], together with systematic evaluation of the accuracy–latency trade-off toward online deployment.

Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under Grant No. RS-2025-02215344 (Development of AI Technology with Robust and Flexible Resilience Against Risk Factors).

References

1. Braunegg, A., Chakraborty, A., Krumdick, M., Lape, N., Leary, S., Manville, K., Merkhofer, E., Strickhart, L., Walmer, M.: APRICOT: A dataset of physical adversarial attacks on object detection. In: ECCV (2020)
2. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
3. Ertler, C., Mislej, J., Ollmann, T., Porzi, L., Neuhold, G., Kuang, Y.: The mapillary traffic sign dataset for detection and classification on a global scale. In: ECCV (2020)
4. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: CVPR (2018)
5. Hingun, N., Sitawarin, C., Li, J., Wagner, D.: REAP: A large-scale realistic adversarial patch benchmark. In: ICCV (2023)
6. Jing, L., Wang, R., Ren, W., Dong, X., Zou, C.: PAD: Patch-agnostic defense against adversarial patch attacks. In: CVPR (2024)
7. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022)
8. Liu, J., Levine, A., Lau, C.P., Chellappa, R., Feizi, S.: Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In: CVPR (2022)
9. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: ECCV (2024)
10. Liu, X., Yang, H., Liu, Z., Song, L., Li, H., Chen, Y.: DPatch: An adversarial patch attack on object detectors. arXiv preprint arXiv:1806.02299 (2018)
11. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022)
12. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection with vision transformers. In: ECCV (2022)
13. NVIDIA-AI-IOT: NanoOWL: Real-time open-vocabulary object detection with NVIDIA TensorRT. `https://github.com/NVIDIA-AI-IOT/nanoowl`
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
17. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: WACV (2022)
18. Tarchoun, B., Ben Khalifa, A., Mahjoub, M.A., Abu-Ghazaleh, N., Alouani, I.: Jedi: Entropy-based localization and removal of adversarial patches. In: CVPR (2023)

19. Telea, A.: An image inpainting technique based on the fast marching method. J. Graph. Tools **9**(1) (2004)
20. Xiang, C., Bhagoji, A.N., Sehwag, V., Mittal, P.: PatchGuard: A provably robust defense against adversarial patches via small receptive fields and masking. In: USENIX Security. USENIX Association (2021)
21. Xiang, C., Mahloujifar, S., Mittal, P.: PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. In: USENIX Security. USENIX Association (2022)
22. Xiang, C., Valtchanov, A., Mahloujifar, S., Mittal, P.: ObjectSeeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking. In: IEEE S&P (2023). https://doi.org/10.1109/SP46215.2023.10179319
23. Xiang, C., Wu, T., Dai, S., Petit, J., Jana, S., Mittal, P.: PatchCURE: Improving certifiable robustness, model utility, and computation efficiency of adversarial patch defenses. In: USENIX Security. USENIX Association (2024)
24. Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., Chen, P.Y., Wang, Y., Lin, X.: Adversarial T-Shirt! evading person detectors in a physical world. In: ECCV (2020)
25. Xu, K., Xiao, Y., Zheng, Z., Cai, K., Nevatia, R.: PatchZero: Defending against adversarial patch attacks by detecting and zeroing the patch. In: WACV (2023)
26. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
27. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S.: Traffic-sign detection and classification in the wild. In: CVPR (2016)