TransPatch: Learning a Universal Adversarial Patch for ViT–CNN Cross-Architecture Transfer in Semantic Segmentation (Student Abstract)

Sargam Goyal[1], Agam Pandey[1], Aarush Aggarwal[1*], Akshat Tomar[1*], Amritanshu Tiwari[1]

[1]Data Science Group, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India - 247667
sargam_g@mfs.iitr.ac.in, agam_p@ce.iitr.ac.in, aarush_a@ma.iitr.ac.in, akshat_t@mfs.iitr.ac.in, amritanshu_t@mfs.iitr.ac.in

Abstract

Robust semantic segmentation is crucial for safe autonomous driving, yet deployed models remain vulnerable to black-box adversarial attacks when target weights are unknown. Most existing approaches either craft image-wide perturbations or optimize patches for a single architecture, which limits their practicality and transferability. We introduce **TransPatch**, a training framework for learning a *universal adversarial patch* that generalizes across images and architectures without requiring access to target model parameters.

Code — https://github.com/dsgiitr/transpatch_ViTCNN

Introduction

Semantic segmentation is a core perception task in autonomous driving, providing dense scene understanding for planning and control. Recent work has examined adversarial attacks on segmentation models to probe vulnerabilities and improve cross-architecture transferability (Gu et al. 2021; Jia et al. 2023), but two key limitations remain. Most methods generate image-wide perturbations, which are hard to deploy physically and require training multiple examples across scenes, and they typically optimize for a single model, offering limited insight into cross-architecture transfer.

Adversarial transferability in semantic segmentation is especially underexplored across fundamentally different architectures. CNNs, with strong locality and edge-based inductive biases, are often harder to break, whereas ViTs have been shown to be more vulnerable due to their global attention mechanisms (Mahmood et al. 2021; Fu et al. 2025; Islam et al. 2025). However, most prior work still focuses on classification or within-architecture transfer, leaving the ViT→CNN (and CNN→ViT) case in segmentation far less studied.

We address this gap with **TransPatch**, a patch-based universal adversarial attack that learns a single physically deployable patch effective across images and both ViT and CNN architectures. TransPatch first exploits the ease of destabilizing ViTs, then guides transfer to CNNs through ensemble training with gradient alignment. To further boost transfer, it identifies the most vulnerable semantic class using class-wise predictive entropy from a CNN surrogate and biases patch placement toward these high-risk regions.

Method

We propose **TransPatch**, a framework that combines (i) semantic placement on empirically sensitive regions, (ii) a two-stage ViT→CNN training curriculum with gradient alignment, and (iii) lightweight priors (attention hijack, boundary, frequency and smoothness) to maximize cross-architecture transferability.

Sensitive Region Placement. Using a ViT surrogate (Segformer) we compute class-wise predictive entropy on clean images and select the class $c^\star$ with highest uncertainty (pole in our experiments). We extract $c^\star$ regions from the ViT prediction, dilate them by k pixels for boundary context, and bias placement toward top-$p\%$ high-entropy pixels. If no valid region exists, we sample randomly. This yields patches positioned where ViTs attend globally and CNNs are most fragile.

Two-Stage Training. *Stage 1 (ViT-only):* Because ViTs are more vulnerable to perturbations, we first destabilize their high-confidence predictions. We use a weighted cross-entropy loss that puts more emphasis on pixels the ViT originally classified correctly (high confidence) and less on those already misclassified:

$$L_{\text{Stage1}} = \frac{(1-\gamma)\sum_{i\in\mathcal{C}} \mathcal{L}_{CE}(p_i, y_i)}{|\mathcal{C}| + |\mathcal{I}|} + \frac{\gamma \sum_{j\in\mathcal{I}} \mathcal{L}_{CE}(p_j, y_j)}{|\mathcal{C}| + |\mathcal{I}|}. \quad (1)$$

Stage 2 (ViT+CNN ensemble): After the patch becomes effective on the ViT, we extend training to an ensemble of ViT and CNN models so it does not overfit to one architecture. We focus the loss on pixels that exhibit the largest distribution shift under the patch, measured by the Jensen–Shannon divergence between clean and adversarial logits, forming a high-transfer set $\mathcal{H}$ where errors are most likely to generalize. The Stage 2 loss then prioritizes these pixels:

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

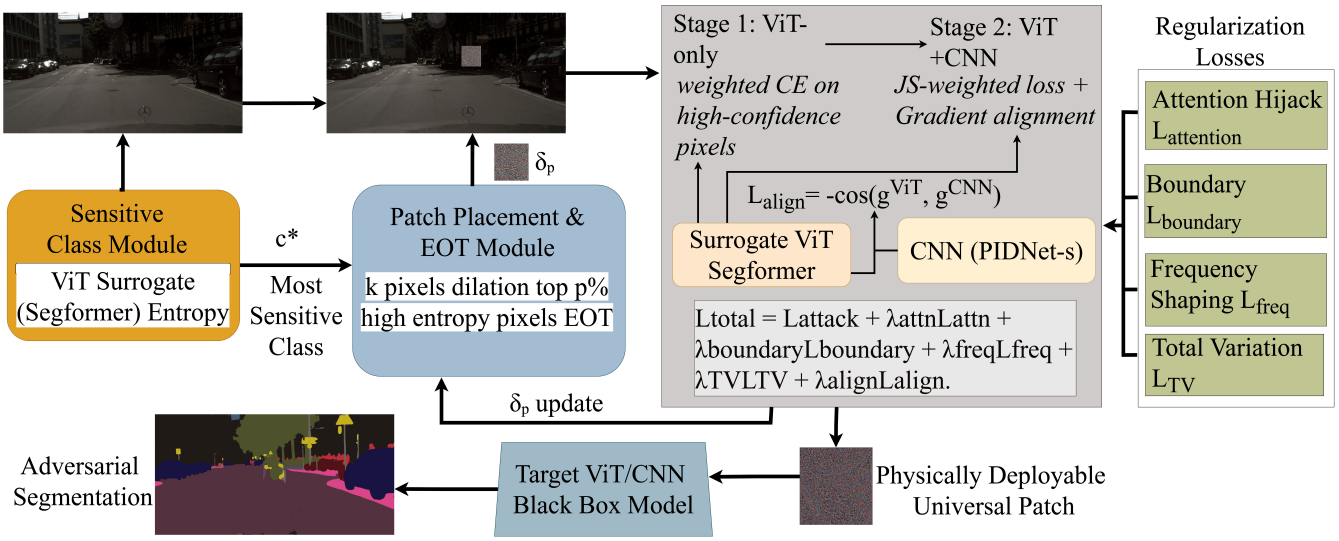

Figure 1: **TransPatch framework.** Sensitive-class EOT placement and a two-stage ViT→CNN trainer with JS loss, gradient alignment, and priors learn a universal patch θ transferable to unseen models.

$$L_{\text{Stage2}} = \frac{(1-\beta)\sum_{i\in\mathcal{H}} \mathcal{L}_{CE}(p_i, y_i)}{|\mathcal{H}| + |\mathcal{L}|} + \frac{\beta \sum_{j\in\mathcal{L}} \mathcal{L}_{CE}(p_j, y_j)}{|\mathcal{H}| + |\mathcal{L}|}. \tag{2}$$

To bridge the inductive-bias gap between global (ViT) and local (CNN) representations, we also add a gradient-alignment term with cosine similarity that encourages the patch to follow a common update direction across models:

$$L_{\text{align}} = -\cos\big(g^{\text{ViT}}, g^{\text{CNN}}\big). \tag{3}$$

Attention Hijack and Priors. Because ViTs encode information through self-attention, pulling their focus toward the patch can collapse their global context. We therefore maximize the fraction of attention allocated to patch tokens. Finally, we include boundary, frequency-shaping, and total-variation regularizers (details in the supplement) to make the patch both physically realizable and less conspicuous while maintaining transferability.

Final Objective. During training we alternate L_{Stage1} and L_{Stage2} while adding the priors:

$$L_{\text{total}} = L_{\text{attack}} + \lambda_{\text{attn}} L_{\text{attn}} + \lambda_{\text{boundary}} L_{\text{boundary}} + \lambda_{\text{freq}} L_{\text{freq}} + \lambda_{\text{TV}} L_{\text{TV}} + \lambda_{\text{align}} L_{\text{align}}. \tag{4}$$

Expectation-over-Transformation (random scale/rotation/translation) is applied each iteration to ensure physical robustness.

Experiments

Table1 reports mean mIOU on the Cityscapes validation dataset for different segmentation models and the relative percentage drop in mIOU to quantify the attacks strength.

Model	Random Patch mIOU	TransPatch mIOU (Ours)	mIOU Drop (%)
PIDNet-S	0.865101	0.814804	5.81
PIDNet-M	0.861862	0.812683	5.71
PIDNet-L	0.899625	0.844479	6.09
BiSeNetv1	0.705789	0.678369	3.88
BiSeNetv2	0.684500	0.652980	4.60
Segformer	0.767412	0.722656	5.82

Table 1: Mean mIoU on Cityscapes with patch applied to most sensitive class.

The mIOU drop in unseen models is comparable to that in source models proving its cross-model transferability. Notably, the same learnt patch is applied to the area containing "pole" (most sensitive class in Cityscapes training data) of all validation images, demonstrating its capability to generalise across multiple scenes.

Conclusion

We presented **TransPatch**, a framework for learning universal adversarial patches that transfer across Vision Transformer and CNN segmentation models. By combining entropy-driven placement, a two-stage ViT→CNN curriculum with gradient alignment, and lightweight priors such as attention hijacking, TransPatch yields a single physically deployable patch that generalizes across scenes and architectures without access to target weights, highlighting overlooked vulnerabilities in hybrid perception stacks for autonomous driving.

References

Fu, Y.; et al. 2025. Patch-Fool: Are Vision Transformers Always Robust Against Adversarial Perturbations? *arXiv*

preprint.

Gu, J.; Zhao, H.; Tresp, V.; and Torr, P. H. S. 2021. Adversarial Examples on Segmentation Models Can be Easy to Transfer. *CoRR*, abs/2111.11368.

Islam, C. M.; et al. 2025. Mechanistic Understandings of Representation Vulnerabilities and Engineering Robust Vision Transformers. *arXiv preprint*.

Jia, X.; Gu, J.; Huang, Y.; Qin, S.; Guo, Q.; Liu, Y.; and Cao, X. 2023. TranSegPGD: Improving Transferability of Adversarial Examples on Semantic Segmentation. arXiv:2312.02207.

Mahmood, K.; et al. 2021. On the Robustness of Vision Transformers to Adversarial Examples. In *ICCV*.