REAP: A Large-Scale Realistic Adversarial Patch Benchmark

Nabeel Hingun[1] Chawin Sitawarin[1] Jerry Li[2] David Wagner[1]

[1]UC Berkeley [2]Microsoft Research

Summary

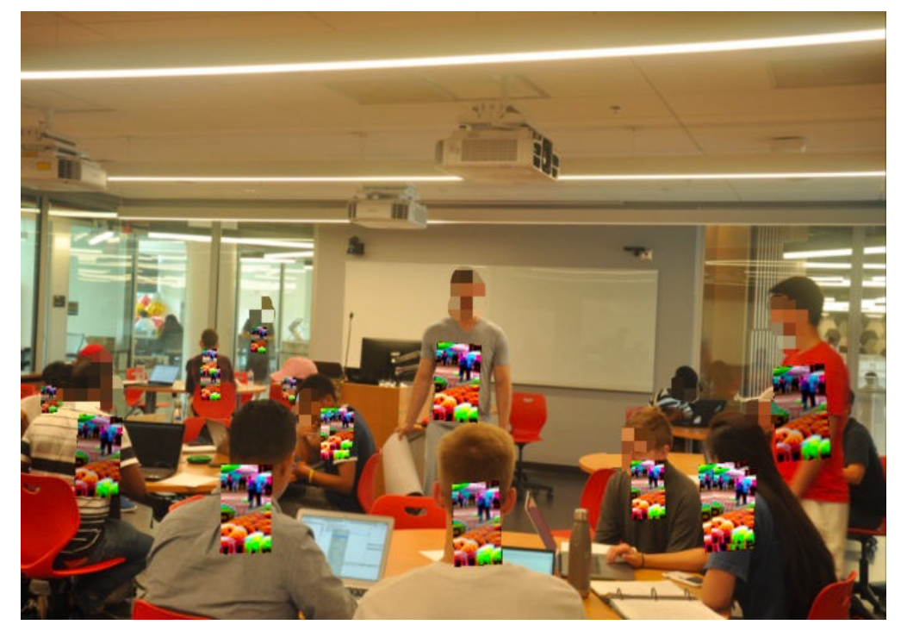
1. We propose REAP, a **realistic and large-scale** benchmark for **adversarial patches**.
2. Realistic: comes with annotated 3D geometric and brightness-contrast transformations.
3. Large-scale: 14K samples over 10K images of driving scenes from Mapillary Vistas dataset.

Evaluation in Past Literature

Unrealistic

Karmon et al. [2018]

Brown et al. [2018]

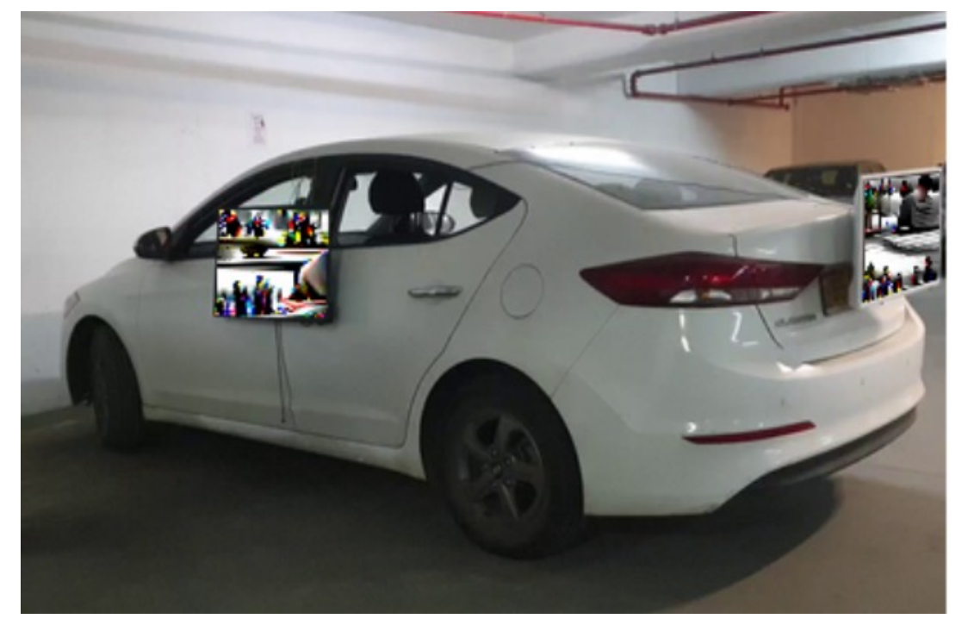

Wu et al. [2020]

Hard to reproduce

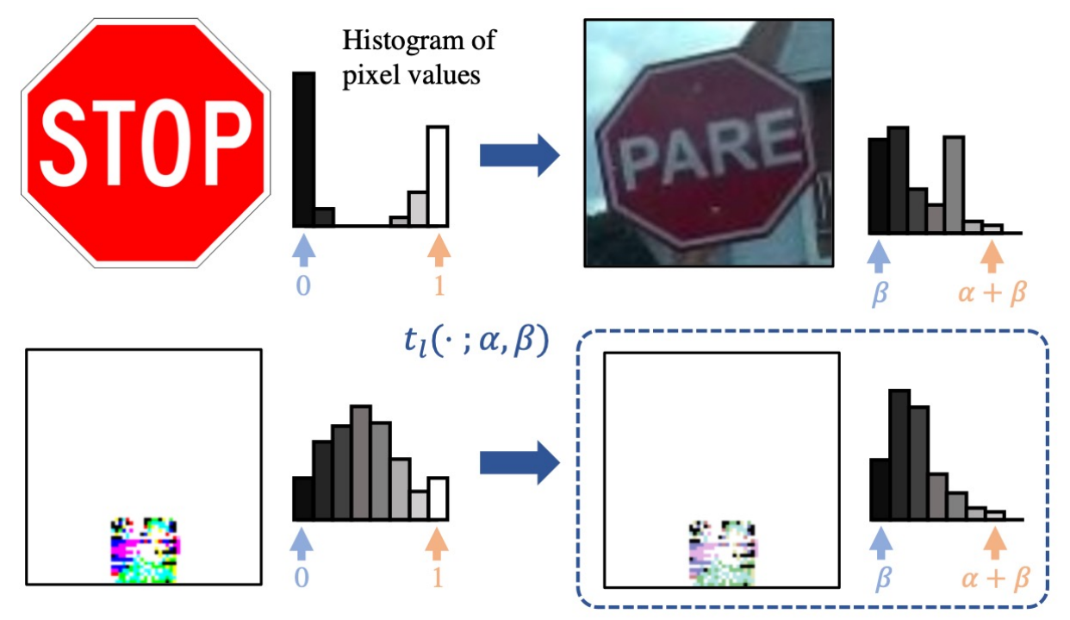
Jan et al. [2019]

More realistic but small and not diverse

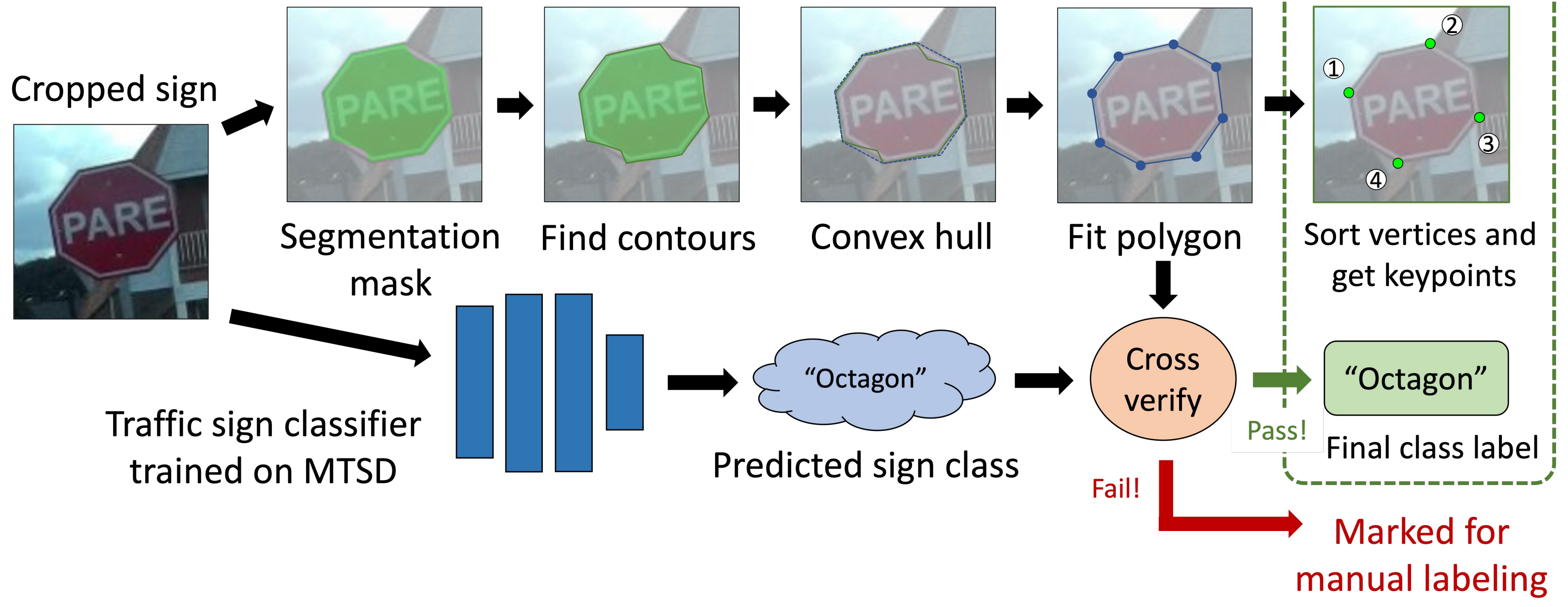
Zhao et al. [2019]


Hoory et al. [2020]

Relighting Transform



Geometric Transform


Cropped sign

Segmentation mask

Find contours

Convex hull

Fit polygon

Sort vertices and get keypoints

Traffic sign classifier trained on MTSD

"Octagon"

Predicted sign class

Cross verify

Pass!

"Octagon"

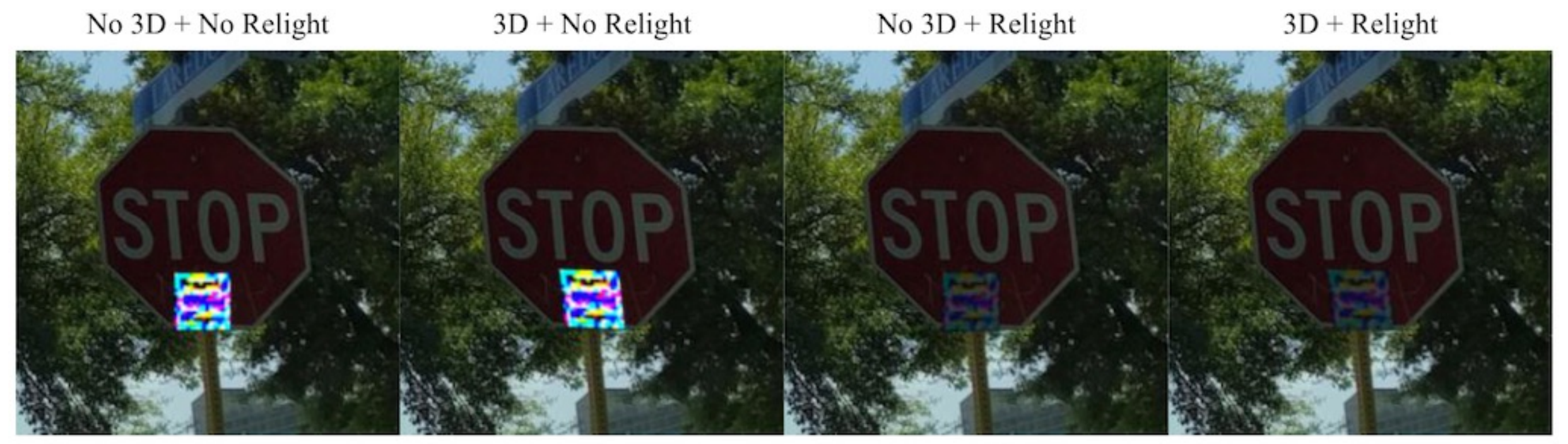
Final class label

Fail!

Marked for manual labeling

Samples From REAP Benchmark

Effects of the Transforms

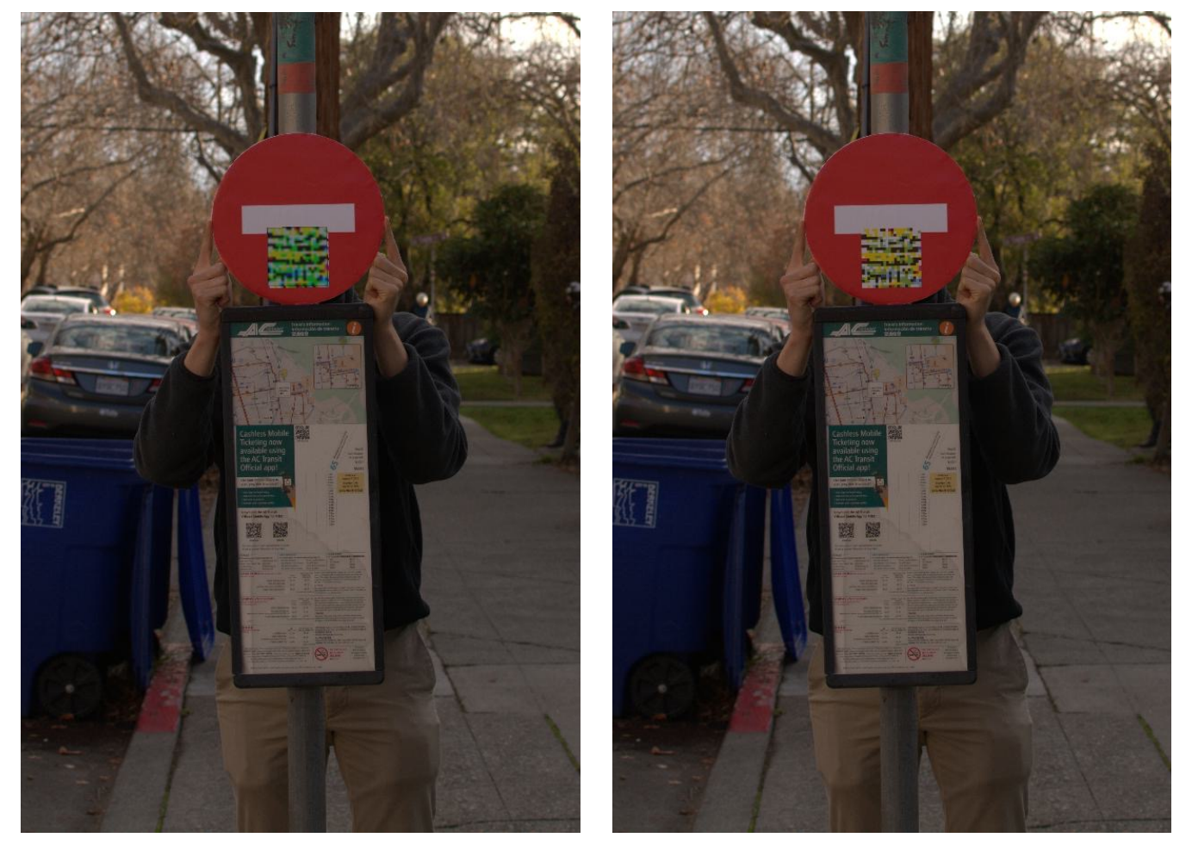

No 3D + No Relight

3D + No Relight

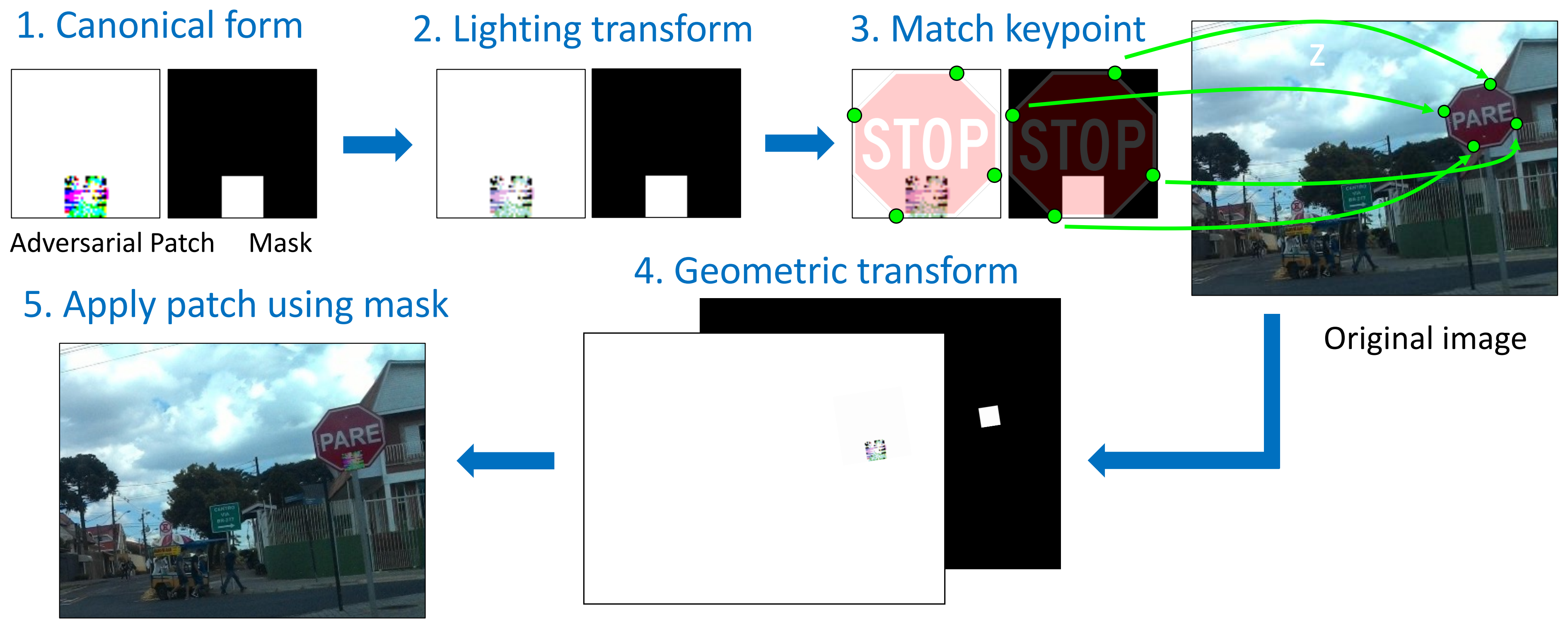
No 3D + Relight

3D + Relight

Realism Test



Adversarial Patch Rendering


1. Canonical form

Adversarial Patch

Mask

2. Lighting transform

3. Match keypoint

Original image

4. Geometric transform

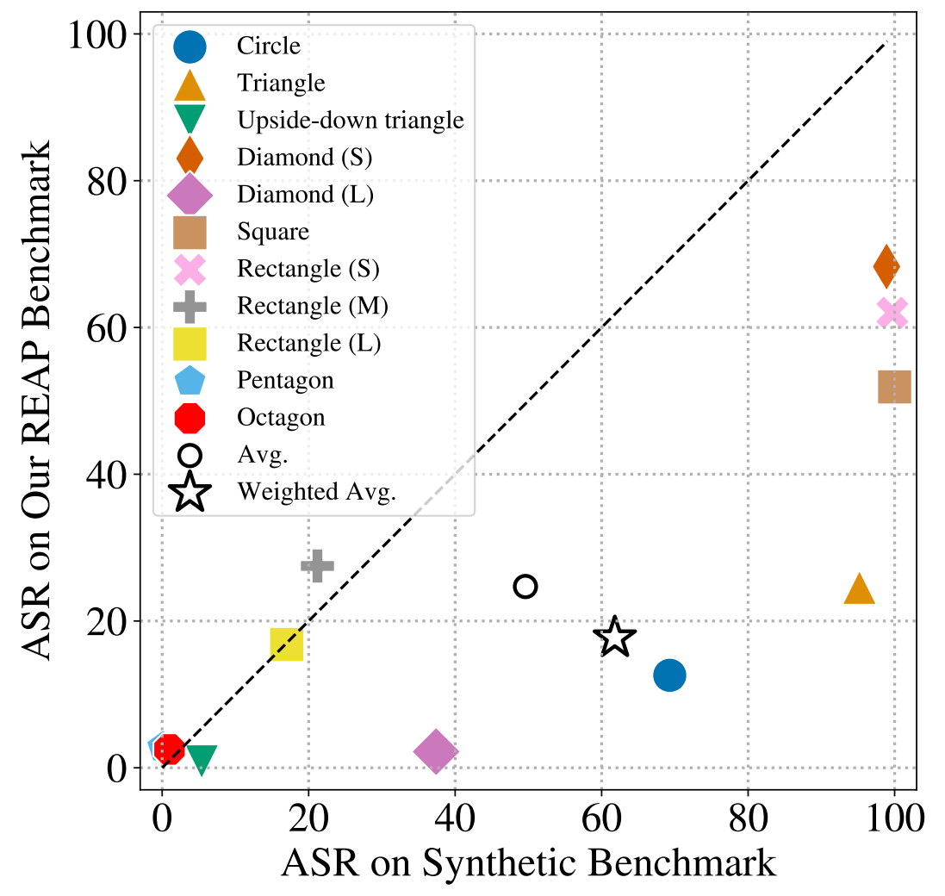
5. Apply patch using mask

Results From REAP Benchmark

Patch Size	FRCNN		YOLOF		DINO	
	FNR	mAP	FNR	mAP	FNR	mAP
No patch	4.3	72.9	18.5	54.8	14.1	68.2
Small (10"×10")	15.4	59.4	33.7	43.5	32.0	60.4
Medium (10"×20")	22.4	46.5	42.7	36.6	35.4	52.6
Large (two 10"×20")	50.0	18.2	72.8	19.4	62.8	39.5

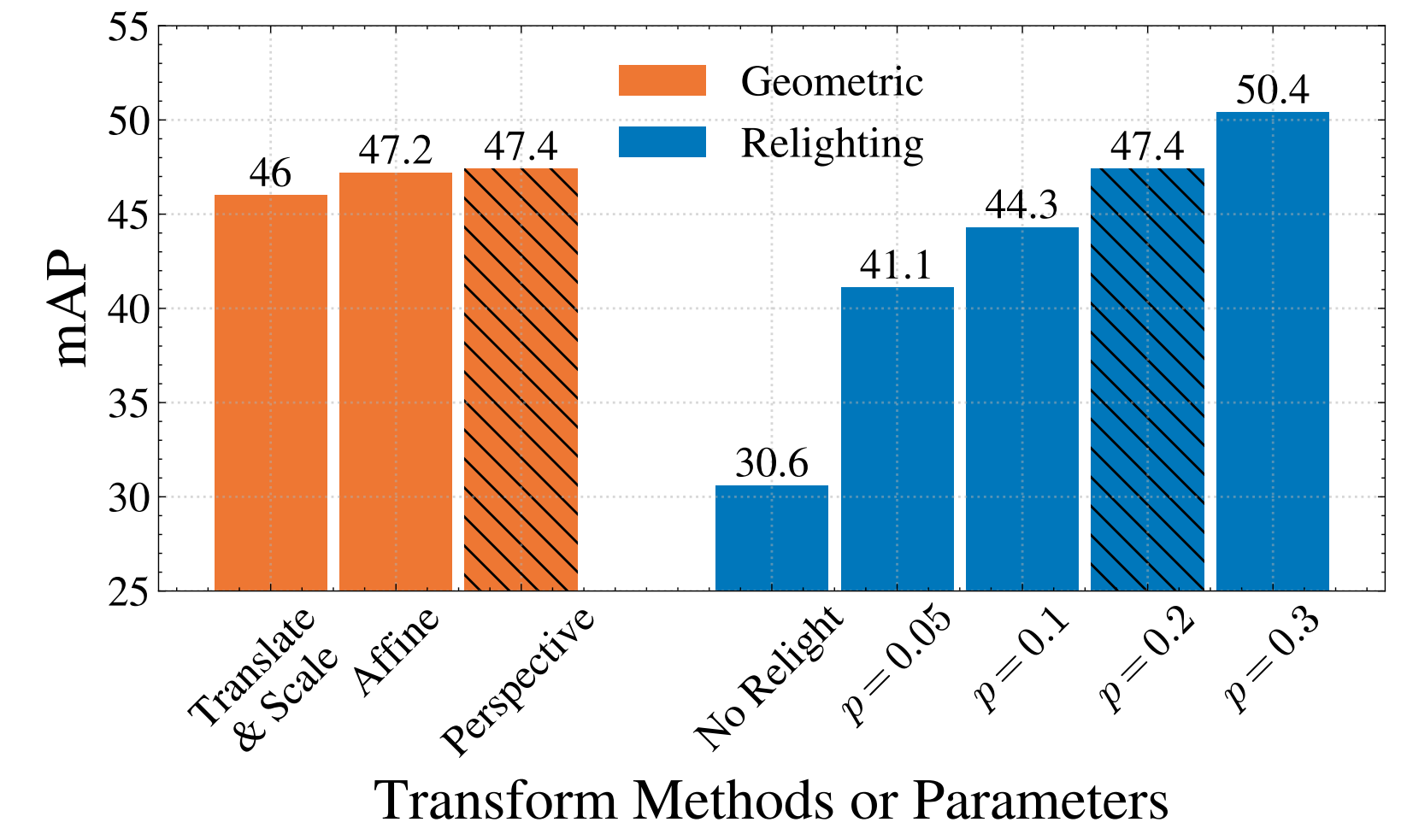
Patch Size	Adv. FRCNN		Adv. YOLOF		Adv. DINO	
	FNR	mAP	FNR	mAP	FNR	mAP
No patch	3.1	73.3	21.0	55.0	9.4	74.2
Small (10"×10")	3.8	71.8	22.5	54.7	1.8	80.6
Medium (10"×20")	6.1	66.8	27.1	51.9	1.2	80.1
Large (two 10"×20")	13.9	56.3	57.7	34.1	3.6	77.8

Small Medium Large



◁ Naïve synthetic benchmark overestimates attack success rate of the patches for all classes of the signs and for all patch sizes.

Lighting transform is important to achieve a faithful benchmark. ▷



Attacks	ASR (↑)	mAP (↓)
Adv. DINO		
No Attack	n/a	65.7
Per-Class Attack	0.1	75.1
Per-Instance Attack	**2.7**	**63.7**
Transfer from Adv. Faster R-CNN	0.1	76.5
Transfer from Adv. YOLOF	0.2	76.1
Transfer from DINO	0.0	79.6
Transfer from Synthetic	0.4	72.7

- Adversarial training seems very effective at stopping universal attacks.
- But it seems to also overfit to the attack, but no evidence of gradient obfuscation.