Supplementary Materials of *Design and Interpretation of Universal Adversarial Patches in Face Detection*

Xiao Yang[1⋆] Fangyun Wei [2⋆] Hongyang Zhang[3⋆] Jun Zhu[1‡]

Dept. of Comp. Sci. & Tech., BNRist Center, Institute for AI, Tsinghua University[1]
Microsoft Research Asia[2] TTIC[3]
{yangxiao19@mails, dcszj@}.tsinghua.edu.cn[1]
fawe@microsoft.com[2] hongyanz@ttic.edu[3]

A Generality Experiments

Generality of training dataset. We randomly split WIDER FACE training dataset into two subsets, namely, split 1 and split 2, to verify the generality of different training sources. Figure 1 shows the evolution of adversarial samples opAs illustrated in Table 2, we reduce the patch to different proportions from 90% to 60%. As the scale of patches decreases, there exists a general downward trend for performance. We also paste adversarial patches on different 5 locations from top to bottom for comparisons in Table 3. The results show that it is effective to stick on different locations.

As illustrated in Table 2, we reduce the patch to different proportions from 90% to 60%. As the scale of patches decreases, there exists a general downward trend for performance. We also paste adversarial patches on different 5 locations from top to bottom for comparisons in Table 3. The results show that it is effective to stick on different locations.

timized by *Patch-IoU* method on each split. The final patches are face-like and can be falsely detected by baseline face detector. The two patches look slightly different, partially because of the non-convex properties of optimization [1, 2] and different initialization.

Attack by part of patch. To examine the attacking performance of part of the adversarial patch that is optimized by *Patch-IoU*, we remove a half and one third area of the whole patch and test the performance of the remaining part of the patch on the WIDER FACE validation dataset. Table 1 shows the associated numerical results. We see that removing a part of the patch hurts the performance of the adversarial patch as an attacker.

B Improved Patches on Different Scales and Locations

As illustrated in Table 2, we reduce the patch to different proportions from 90% to 60%. As the scale of patches decreases, there exists a general downward trend

⋆ Equal contribution. ‡ corresponding author.

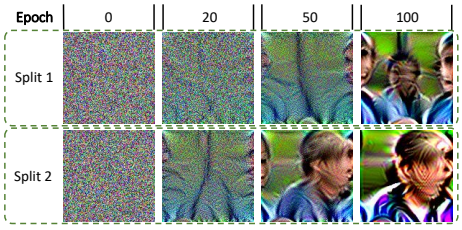

Fig. 1. Optimization results by *Patch-IoU* on WIDER FACE training subset, split 1 and split 2. *Patch-IoU* generates face-like adversarial patch which is falsely detected by baseline face detector.

Table 1. Attacking performance of parts of adversarial patches. Half (One-third)-*Part* means removing one half (one-third) area of the whole patches.

Precision/ Recall	*Easy*	*Medium*	*Hard*	*All*
Baseline-SLN	99.0/ 73.4	99.4/ 62.4	99.4/ 27.9	99.4/ 22.5
Patch-IoU-SLN	2.7/2.7	6.5/3.7	7.3/1.8	7.3/1.4
Half-*Top*	93.3/ 38.8	95.9/ 35.6	96.1/ 15.6	96.1/ 12.5
Half-*Bottom*	99.2/ 47.4	99.5/ 40.1	99.5/17.2	99.5/13.8
Half-*Left*	98.5/ 41.2	99.1/ 36.3	99.1/15.8	99.1/ 12.7
Half-*Right*	82.0/ 30.6	88.8/28.7	89.2/12.5	89.2/ 10.1
One-third-*Top*	25.3/18.7	39.1/19.1	40.6/8.5	40.6/6.8
One-third-*Bottom*	74.6/23.0	83.2/12.0	83.7/9.1	83.7/7.3
One-third-*Left*	88.1/24.1	92.9/22.8	93.2/10.1	93.2/8.1
One-third-*Right*	12.8/11.9	22.44/12.7	23.6/5.7	23.6/4.6

for performance. We also paste adversarial patches on different 5 locations from top to bottom for comparisons in Table 3. The results show that it is effective to stick on different locations.

C Transferablility on Different Models

We examine the transferability of adversarial patches between different models in Table 4, including ResNet-50, ResNet-101 and ResNext-101. ResNet-18 is a surrogate white-box model. It is also observed that proposed adversarial patches have some effects on unknown black-box models.

D Some Examples about Shortcomings of Average Precision (AP)

We illustrate some arguments about shortcomings of Average Precision (AP) as an evaluation metric in Figure 2.

Table 2. Precision and recall comparisons of *Patch-Score*, *Patch-Score-Focal* and *Patch-Combination* based on different scales with $\delta = 0.99$.

Scale	Precision/ Recall	*Easy*	*Medium*	*Hard*	*All*
90%	*Patch-Score*	98.9/40.0	99.3/30.7	99.2/13.0	99.3/10.5
	Patch-Score-Focal	98.9/37.7	99.2/29.5	99.2/12.5	99.2/10.0
	Patch-Combination	98.9/35.4	99.2/27.0	99.2/11.5	99.2/9.2
80%	*Patch-Score*	99.1/50.1	99.4/39.7	99.4/16.9	99.4/13.6
	Patch-Score-Focal	99.1/49.2	99.4/39.1	99.4/16.8	99.4/13.3
	Patch-Combination	99.1/46.7	99.4/36.1	99.4/15.4	99.4/12.4
70%	*Patch-Score*	99.2/58.7	99.5/47.0	99.5/20.0	99.5/16.1
	Patch-Score-Focal	99.2/58.2	99.5/46.6	99.5/19.8	99.5/15.9
	Patch-Combination	99.2/55.9	99.4/44.3	99.4/18.9	99.4/15.2
60%	*Patch-Score*	99.1/64.2	99.4/52.0	99.4/22.2	99.4/17.9
	Patch-Score-Focal	99.1/63.9	99.4/51.8	99.4/22.1	99.4/17.8
	Patch-Combination	99.1/62.5	99.4/50.2	99.4/21.4	99.4/17.2

Table 3. Precision and recall comparisons of *Patch-Score*, *Patch-Score-Focal* and *Patch-Combination* based on different locations with $\delta = 0.99$.

Location	Precision/ Recall	*Easy*	*Medium*	*Hard*	*All*
Top	*Patch-Score*	98.5/25.9	99.0/20.0	99.0/8.5	99.0/6.8
	Patch-Score-Focal	98.4/23.4	98.9/18.1	98.9/7.7	98.9/6.2
	Patch-Combination	98.2/20.6	98.7/15.7	98.8/6.7	98.8/5.3
Center of top and middle	*Patch-Score*	97.9/17.1	98.7/14.8	98.7/6.3	98.7/5.1
	Patch-Score-Focal	97.7/15.1	98.6/12.9	98.6/5.5	98.6/4.5
	Patch-Combination	97.7/15.3	98.6/13.0	98.6/5.6	98.6/4.5
Middle	*Patch-Score*	97.9/17.9	98.6/15.0	98.7/6.1	98.7/5.2
	Patch-Score-Focal	97.5/14.6	98.3/12.1	98.4/5.2	98.4/4.2
	Patch-Combination	97.5/14.8	98.3/11.8	98.4/5.0	98.4/4.0
Center of bottom and middle	*Patch-Score*	98.1/24.1	98.7/18.5	98.7/7.8	98.7/6.3
	Patch-Score-Focal	98.4/25.1	98.8/18.3	98.8/7.8	98.8/6.2
	Patch-Combination	98.3/23.3	98.7/15.8	98.7/6.7	98.7/5.4
Bottom	*Patch-Score*	98.9/39.8	99.3/31.1	99.3/13.3	98.3/10.7
	Patch-Score-Focal	98.8/38.8	99.2/31.2	99.2/13.3	98.8/10.7
	Patch-Combination	98.8/39.0	99.2/28.8	99.2/12.1	98.7/9.7

Table 4. Precision and recall of *Patch-Score-Focal* and *Patch-Combination* based on different models with $\delta = 0.99$.

Models	Precision/ Recall	*Easy*	*Medium*	*Hard*	*All*
ResNet-50	Baseline	99.0/81.9	99.4/75.2	99.6/40.3	99.6/32.5
	Patch-Score-Focal	98.8/46.1	99.2/36.0	99.2/15.4	99.2/12.4
	Patch-Combination	98.9/47.6	99.2/37.2	99.2/15.9	99.2/12.8
ResNet-101	Baseline	99.0/83.6	99.4/77.1	99.6/39.8	99.6/32.1
	Patch-Score-Focal	98.9/54.5	99.3/45.2	99.3/19.3	99.3/15.6
	Patch-Combination	99.0/56.0	99.3/46.7	99.4/20.0	99.4/16.1
ResNext-101	Baseline	99.0/82.7	99.4/75.6	99.6/41.0	99.6/33.0
	Patch-Score-Focal	98.9/54.6	99.2/42.3	99.3/17.9	99.3/14.2
	Patch-Combination	98.3/55.7	98.7/43.4	98.7/18.4	98.7/14.8

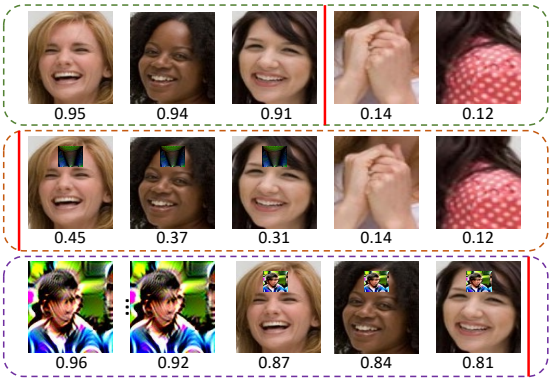

Fig. 2. The first row contains top-5 face proposals for the baseline model and the red line represents the threshold boundary. The second row represents a successful attack. However, the APs in the two rows are the same because the relative rank is the same. The third row illustrates a case where Score(AdversarialPatch) > Score(Face) > Threshold. The AP becomes smaller because of the false positives, though the attack fails according to criterion 2. Thus we cannot use AP to evaluate the performances of adversarial patches.

References

1. Zhang, H., Shao, J., Salakhutdinov, R.: Deep neural networks with multi-branch architectures are intrinsically less non-convex. In: International Conference on Artificial Intelligence and Statistics. pp. 1099–1109 (2019)
2. Zhang, H., Xu, S., Jiao, J., Xie, P., Salakhutdinov, R., Xing, E.P.: Stackelberg gan: Towards provable minimax equilibrium via multi-generator architectures. arXiv preprint arXiv:1811.08010 (2018)