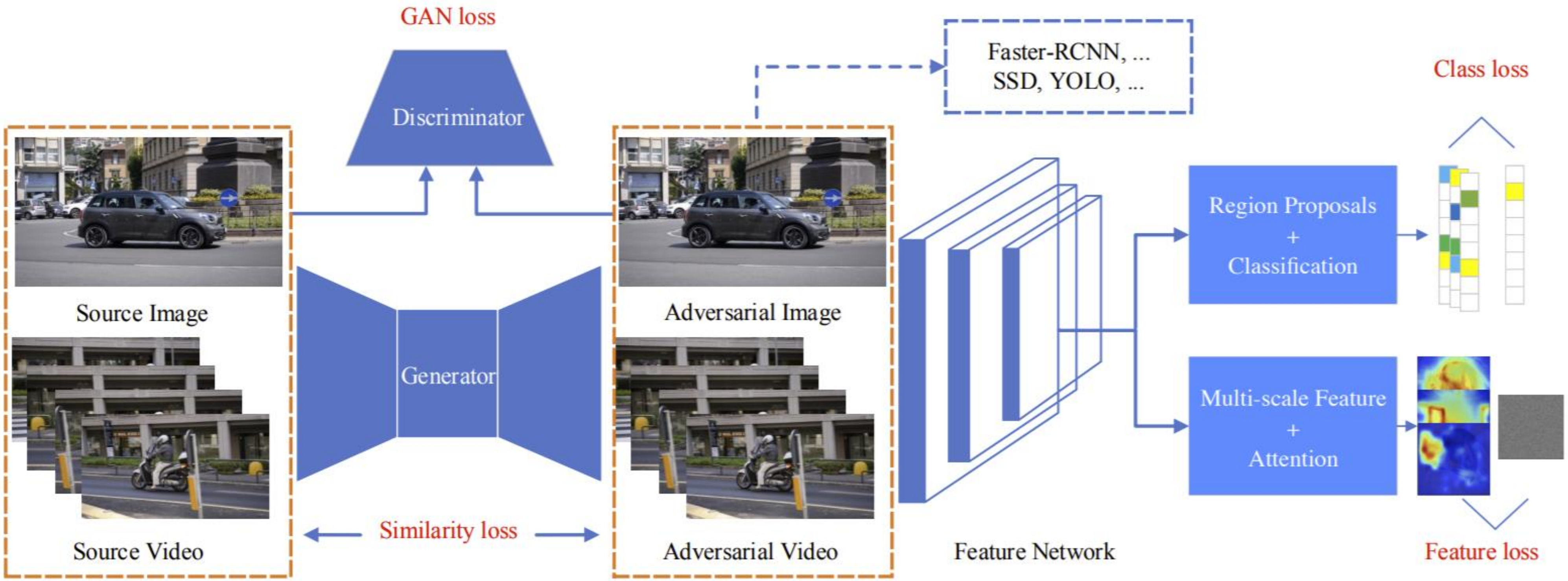
Transferable Adversarial Attacks for Image and Video Object Detection

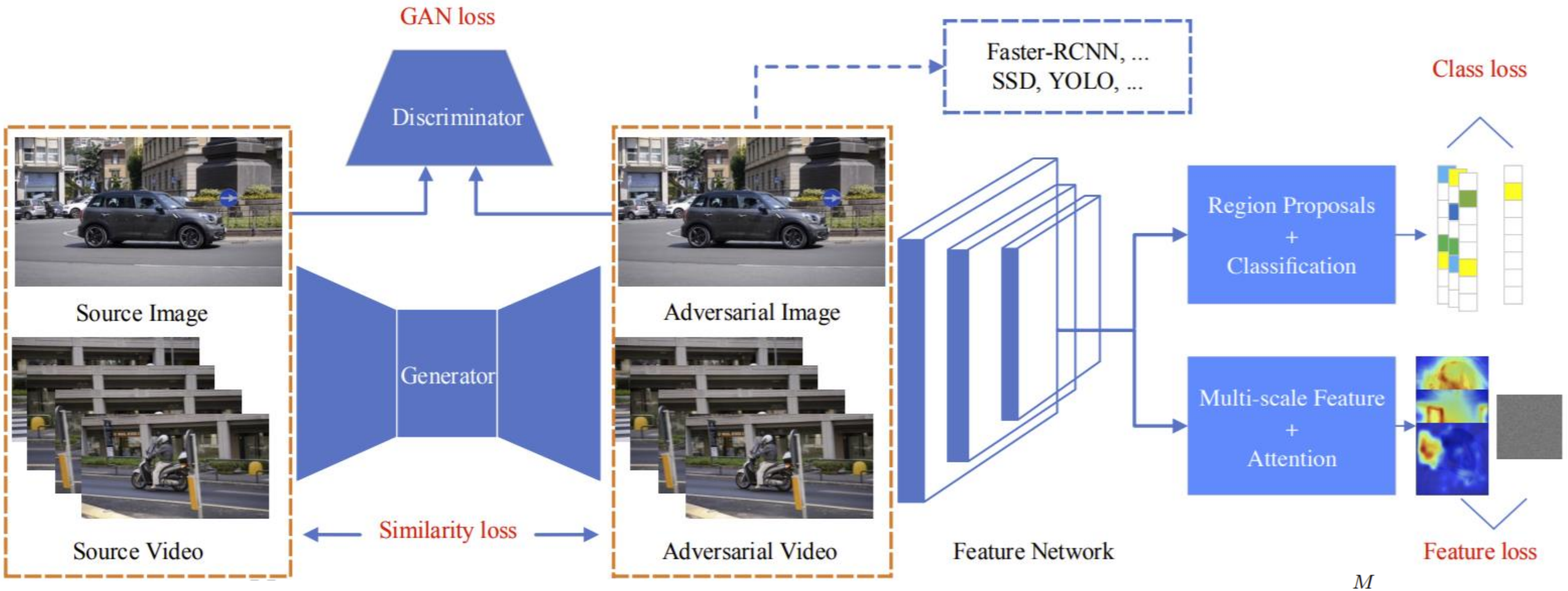
Introduction



Methodology

$$\mathcal{L}_{cGAN}(\mathcal{G},\mathcal{D}) = \mathbb{E}_I[\log\mathcal{D}(I)] + \mathbb{E}_I[\log(1-\mathcal{D}(\mathcal{G}(I)))]$$

$$\mathcal{L}_{DAG}(\mathcal{G}) = \mathbb{E}_I[\sum_{n=1}^{N}[f_{l_n}(\mathbf{X},t_n) - f_{\hat{l}_n}(\mathbf{X},t_n)]]$$



$$\mathcal{L}_{L_2}(\mathcal{G}) = \mathbb{E}_I[||I - \mathcal{G}(I)||_2]$$

$$\mathcal{L}_{Fea}(\mathcal{G}) = \mathbb{E}_I[\sum_{m=1}^{M} ||\mathbf{A}_m \circ (\mathbf{X}_m - \mathbf{R}_m)||_2]$$

Experiment

Table 2: The comparison results between DAG and UEA.

Methods	Accuracy (mAP)		Time (s)
	Faster-RCNN	SSD300	
Clean Images	0.70	0.68	\
DAG	0.05	0.64	9.3
UEA	**0.05**	**0.20**	**0.01**

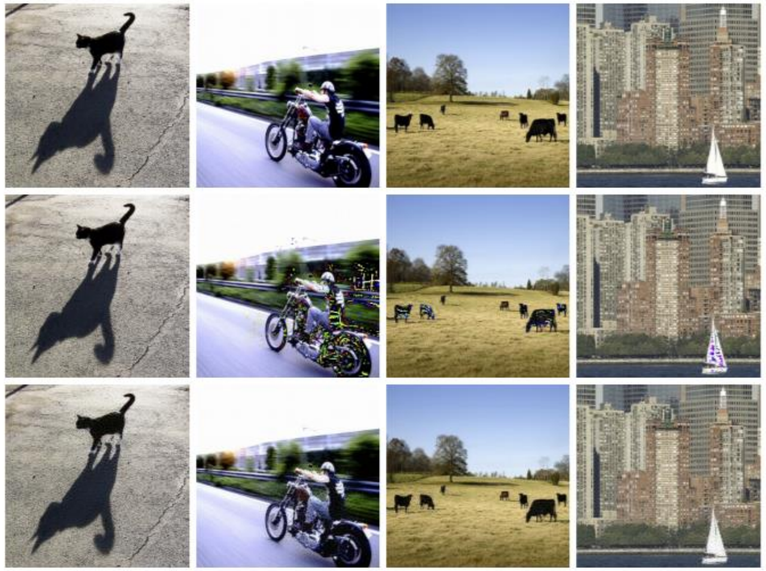


Figure 3: The perceptibility comparison of adversarial images. The first row is clean images. The second row is output by DAG (the iteration is 1, 81, 133, 41, respectively). The third row is our output.

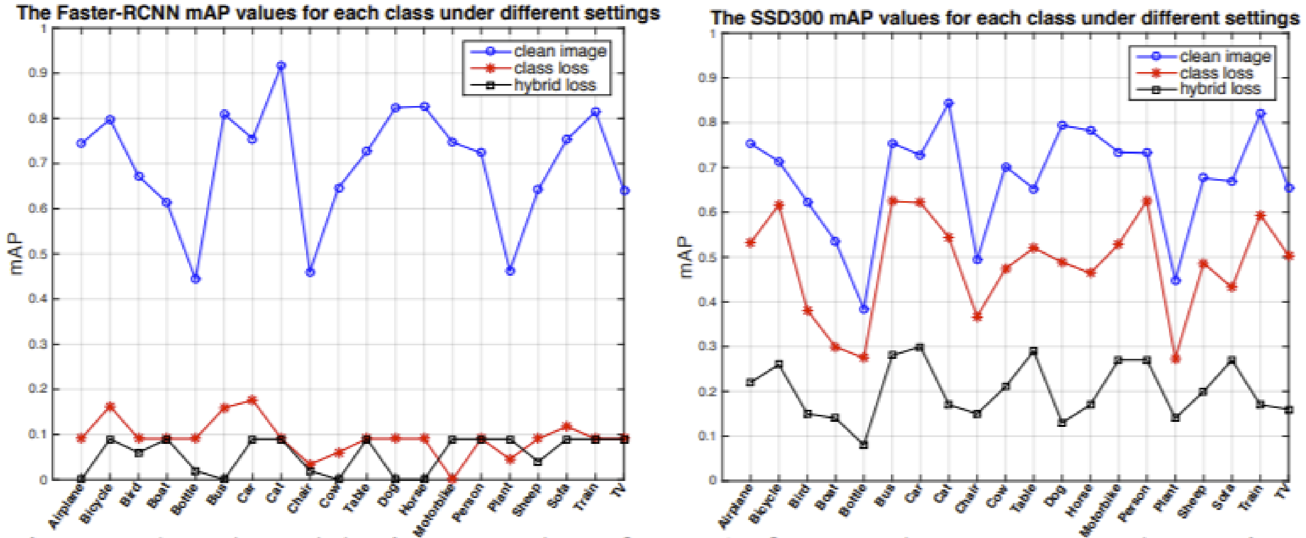


Figure 4: The ablation study of UEA for each category detection.

Experiment

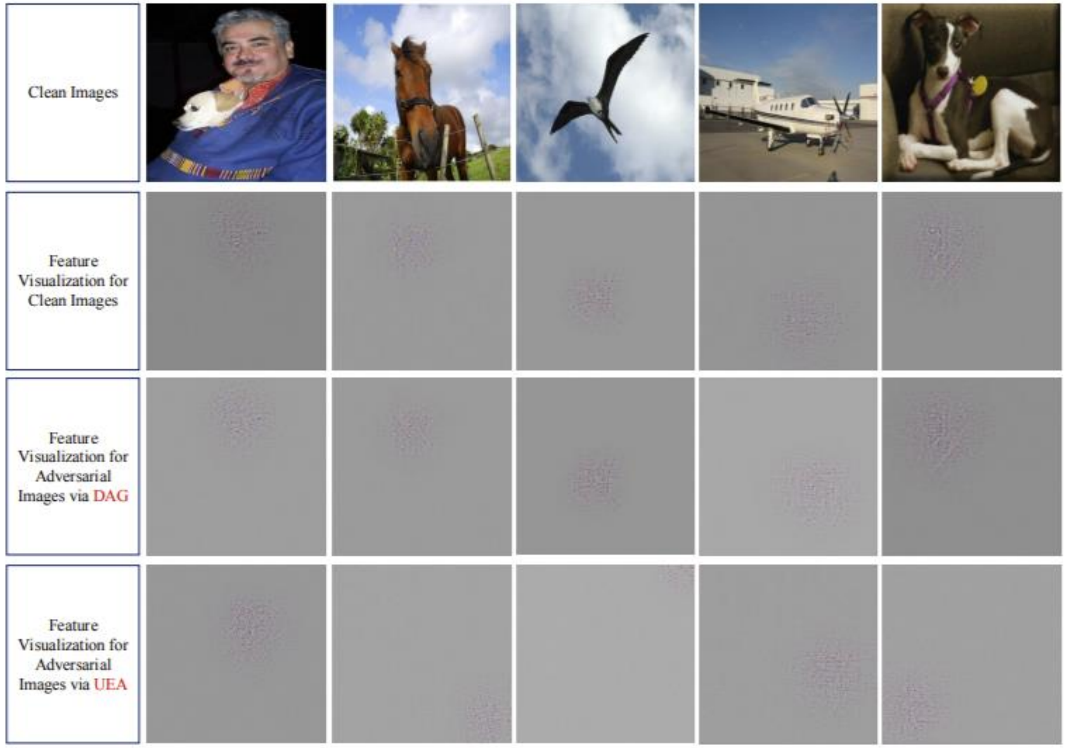
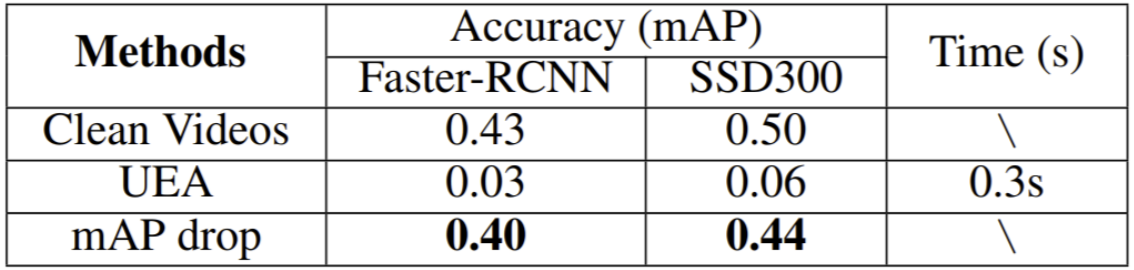

Clean Images
Feature Visualization for Clean Images
Feature Visualization for Adversarial Images via DAG
Feature Visualization for Adversarial Images via UEA

Figure 6: The feature visualization of adversarial examples via DAG and UEA, respectively. Please see the texts for details.

Table 3: The attacking performance of UEA on video detection.



Methods	Accuracy (mAP)		Time (s)
	Faster-RCNN	SSD300	
Clean Videos	0.43	0.50	\
UEA	0.03	0.06	0.3s
mAP drop	**0.40**	**0.44**	\

Conclusion

Advantages:

Efficiency
Transferability
Imperceptiblity

Disadvantages:

Targeted Attack