Imperceptible adversarial attack with entropy feature and segmentation-based constraint

RONGDONG LI*, QINLIANG LIN*, YINGLONG FU, WEICHENG XIE, and LINLIN SHEN, Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, China

Methods of adversarial attack and defense have attracting increasing attention in the fields of security and protection related applications. However, current algorithms carry out perturbations on entire images and mostly consider their imperceptibility to machines, while does not take their human imperceptibility into account. In this work, we propose a constrained adversarial attack algorithm with both machine and human imperceptibility based on image entropy feature and accurate segmentation. The proposed algorithm has three merits. First, image entropy-based feature for quantifying the imperceptibility of a semantic region is introduced, which is simple yet efficient to implement. Second, in terms of the imperceptibility metric, accurate target regions for adversarial perturbation are obtained based on scene-aware segmentation and merging. Third, a general adversarial attack based on segmentation region constraint is proposed to induce both machine and visual imperceptibility. Experimental results in terms of qualitative and quantitative analysis reflect the effectiveness of the proposed algorithm compared with the state of the arts.

CCS Concepts: • **Security and privacy**; • **Computing methodologies** → **Computer vision problems**;

Additional Key Words and Phrases: Adversarial attack, imperceptibility metric, semantic segmentation, constrained attack algorithm, robust object classification

ACM Reference Format:
Rongdong Li, Qinliang Lin, Yinglong Fu, Weicheng Xie, and Linlin Shen. 2021. Imperceptible adversarial attack with entropy feature and segmentation-based constraint. In *2021 10th International Conference on Computing and Pattern Recognition (ICCPR 2021), October 15-17, 2021/Shanghai, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Nowadays, with the increasing awareness of security in the applications of deep learning in the filed of computer vision, network adversarial attack and defense have attracted increasing attention in the security and protection related tasks.

Szegedy et al. [16] revealed that adding tiny pixel perturbations that are imperceptible to machines can result in incorrect inference to classification networks, and proposed an adversarial attack method, namely as L-BFGS. GoodFellow et al. [7] proposed to generate adversarial disturbances based on gradient solving, namely as Fast Gradient Sign Method (FGSM). In order to better solve the problem of large disturbance amplitude resulted from FGSM, Alexey Kurakin et al. [9] proposed the Basic Iterative Method (BIM) method. Different from the above-mentioned gradient-based iterative attack method, Nicholas Carlini and David Wagner [2] proposed an optimization-based attack method, namely as C&W, which took the high attack success rate and low anti-adversarial effect into account.

*Both authors contributed equally to this research.

Manuscript submitted to ACM

However, traditional algorithms most carry out the perturbation attacks on the entire image. Actually, different regions of an image may have different degrees of scene complexity, the contributions of these perturbations may differ much in a successful attack. Sharma et al. [15] proposed to attack the attention mechanism by guiding the adversarial perturbations toward the attention maps. Dong et al. [5] observed that adding perturbations to the background is less useful than that on the salient object. Thus, the adversarial perturbations are suggested to imposed on only the salient regions.

While the corruptions caused by current attack methods are imperceptible to the classification network, the imperceptibility to human vision is not well addressed. With the consideration of the human visual system (HVS), multi factor metric for measuring the perceptual loss between benign examples and adversarial ones, i.e. MulFactorLoss [11], was introduced for adversarial attack. Deng and Karam [4] proposed the frequency-tuned adversarial perturbation, i.e. universal adversarial perturbations are generated based on frequency-domain perturbation components, which are adapted to the local characteristics of discrete cosine transform frequency bands. To make adversarial examples imperceptible, Luo et al. [12] suggested to perturb pixels at high variance zones. Chhabra et al. [3] defined a visual imperceptible bound to preserve the visual appearance of an image while performing adversarial manipulation. Laidlaw and Feizi [10] proposed functional threat models to perturb similar features in the same direction to make potential adversarial examples more imperceptible to humans. Duan et al. [6] proposed to craft and camouflage physical world adversarial examples into natural styles that appear legitimate to human observers, while the regions for perturbation need be captured manually.

To adaptive distribute the perturbation according to human sensitivity to a local stimulus in the benign image, Wang et al. [18] introduced several features for measuring the region complexity to locate the regions to make adversarial attack less visible and perceptible. However, multiple features maybe entangled, and the setting of their regularization hyper-parameters maybe not adaptive to different circumstances. In this work, we introduce an entropy-based feature to evaluate the imperceptibility of the regions for perturbation, this single feature performs similar as the multiple features introduced by Wang et al. [18], while is numerically tested to be simpler and easier to implement. From another aspect, the perturbed pixels could distribute on object boundary regions, which make the attacked noises easily perceptible on the border regions. To make the perturbation regions more robust and the noises in border regions more imperceptible, we propose to use image segmentation and heuristic merging to obtain accurate semantic regions for adversarial attack. The motivation of the proposed algorithm is summarized in Fig. 1.

As shown in Fig. 1, for the backgrounds such as sky, the imposed perturbations are easy for human eye to perceive, while the perturbations on the bird object region are imperceptible, which motivates us to locate the imperceptible regions and use them to constrain the perturbations. From another aspect, the traditional algorithms obtained the region for perturbation based on global image perturbation or equal division of image and merging, which may introduce obvious perturbation traces on the object boundary regions that are perceptible to human vision. This observation motivates us to obtain more imperceptible regions for perturbation based on accurate semantic segmentation.

In this work, an adversarial attack with both network and human imperceptibility is proposed, where image entropy features are introduced for measuring the region imperceptibility. In terms of the imperceptibility metric, semantic segmentation and merging is used to obtain the accurate region for adversarial corruption. The main contributions of the work are summarized as follows

- A simple yet efficient image entropy is proposed to quantify the region imperceptibility.

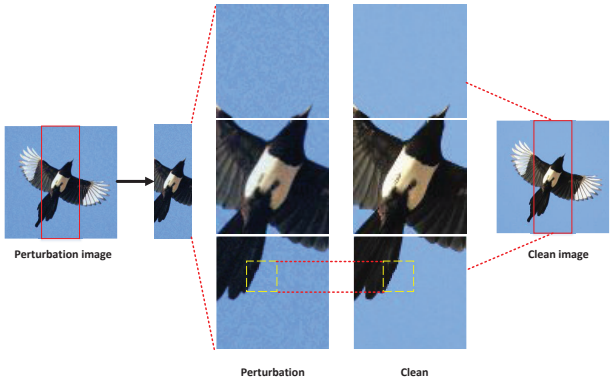

Fig. 1. The motivation of the proposed work. The adversarial perturbations imposed on the texture-complex objects by FGSM [7] present more imperceptible than those on the texture-plain regions, e.g. the yellow dotted rectangle, the perturbations on the bird's tail are more imperceptible than those in the background region.

- Semantic segmentation and region merging is introduced to locate the accurate imperceptible regions for constrained attack.
- A general framework of constrained attack is proposed, which achieved more imperceptible perturbations compared with the state of the arts without largely impairing the adversarial effectiveness, in terms of two publicly-employed metrics.

This work is structured as the following sections. The proposed algorithm is demonstrated in Section 2. Then the experimental results and the corresponding illustrations are demonstrated in Section 3. Finally, the conclusion and some discussions are presented in Section 4.

2 THE PROPOSED ALGORITHM

The flowchart of the proposed algorithm is presented in Fig. 2. First, image entropy features are introduced to measure the imperceptibility degree of each divided region, and used to form a binary mask for accurate region acquisition. Second, the image semantic segmentation of watershed algorithm [1] is employed to extract the accurate candidate regions for constrained attack. Third, a general framework of constrained adversarial attack is proposed based on the imperceptible region, which is equipped with FGSM for evaluation.

2.1 Imperceptibility Metric and Perturbation Region Generation

In this section, the image entropy is introduced for measuring the imperceptibility degree of a given region, which is formulated as follows

$$H(R_k) = -\sum_{i=0}^{255} p_{i,k} \cdot log(p_{i,k}). \tag{1}$$

where $p_{i,k}$ represents the frequency of pixels whose gray values are i in the region of R_k. We argue that the distribution range of gray values will be larger for areas with richer textures or contours, and are prone to be more imperceptible and more likely to be candidate regions for adversarial perturbation. Then the matrix of *mask* with the values of 0 or 1 is generated, where the pixels with $H(R_k)$ being larger than 0.5 are assigned with 1.

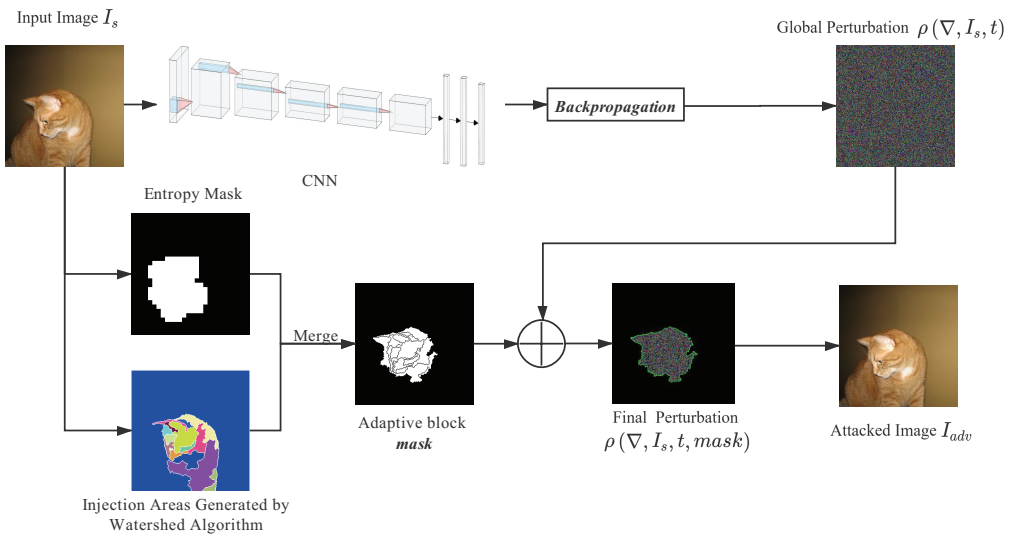

Fig. 2. The framework of the proposed imperceptible attack.

Based on the rough regions encoded in *mask*, the generation of the accurate regions for perturbation is presented in Algorithm 1. First, multiple fragment regions are obtained with the watershed segmentation algorithm [1], where this algorithm is often used for image segmentation based on the analysis of geographic morphology. Then, based on the binary mask with image entropy feature in Section 2.1, these segmented regions are gradually merged to form the final candidate region for perturbation. Formally, if more than a threshold of region overlap ratio, i.e. τ of the pixels in the segmented region have a mask value of 1, then this region is considered to be an imperceptible region and is retained, while the watershed regions that do not meet the above conditions are discarded. The candidate regions for perturbation are formed with the merging of the retained watershed regions.

Algorithm 1 Semantic segmentation and entropy feature-based region merging.

Require: The image segmented by watershed algorithm; The mask image, i.e. *mask* obtained based on image entropy.
Ensure: The merging of the retained imperceptible regions for perturbation, i.e. *RetReg*.

```
Initialize the set SegReg containing all water injection regions with watershed;
while SegReg ≠ ∅ do
    Select a block of injection region b from SegReg;
    if area(b ∩ mask) ≥ τ · area(b) then
        RetReg ← RetReg ∪ b;
        SegReg ← SegReg − b;
    end if
end while
```

Compared with the algorithms based on equal division [18], the proposed approach based on segmentation can capture the accurate boundaries of various semantic regions. Hence, the adversarial perturbations can be avoided across different objects, which are more imperceptible to human vision.

2.2 Constrained Adversarial Attack

The general attack method is formulated as follows

$$\begin{aligned} I_{adv} &\leftarrow I_s + \rho(\nabla, I_s, t), \\ s.t.\ C(I_{adv}) &\neq C(I_s)\ or\ C(I_{adv}) = t \end{aligned} \tag{2}$$

where I_{adv} is the perturbed image after adding the attack disturbance, ∇ denotes the Jacobian vector, and $\rho(\nabla, I_s, t)$ is the attack disturbance that is dependent on the gradient information, input image, i.e. I_s and the target label, i.e. t. $C(\cdot)$ represents the output label of the classification model. Targeted attacks require $C(I_{adv}) = t$; non-targeted attacks require $C(I_{adv}) \neq C(I_s)$. For the generation of $\rho(\nabla, I_s, t)$, FGSM [7] directly uses the gradient information based on network back propagation as the attack disturbance.

We argue that an imperceptible attack performs perturbations on the pixels in the imperceptible or high complexity regions of an image, and can control the noise disturbance in terms of the imperceptibility to combat the major challenge of obvious attack traces. We use *mask* to encode the imperceptible regions and use them to constrain the attack perturbations as follows

$$\rho(\nabla, I_s, t, mask) \leftarrow \rho(\nabla, I_s, t) \oplus mask \tag{3}$$

where $\oplus$ represents the AND operation of two matrixes. If the learning of $\rho(\cdot)$ is formulated as an iterative process, then each iteration needs to use *mask* to constrain the generation of the attack perturbations.

Consequently, the final adversarial attack is formulated as follows

$$\begin{aligned} I_{adv} &\leftarrow I_s + \rho(\nabla, I_s, t, mask), \\ s.t.\ C(I_{adv}) &\neq C(I_s)\ or\ C(I_{adv}) = t \end{aligned} \tag{4}$$

3 EXPERIMENTAL RESULTS

We test our algorithm using a four-kernel Nvidia TITAN GPU Card and Pytorch platform. The dataset used for evaluation is selected from a subset of the ImageNet [14] dataset. As the official dataset of the ILSVRC competition, the ImageNet dataset has the advantages of high accuracy of classification labels, rich types, and greater diversity of various objects in the same class, which has become a pre-training dataset for many classification models. We selected images of 50 categories with a total of 900 images for attack from the dataset. Images with various spatial structures are selected, which include animals and plants, general objects, etc., to test the algorithm robustness on different categories of objects for the location of perturbation regions. The hyper-parameter of τ in Algorithm 1 is set as 0.5.

3.1 Algorithm Analysis

3.1.1 Imperceptibility Features. In order to evaluate the difference between image entropy and multiple features in SpaAdv (FeaSpaAdv) [18], we use them to generate the the attack regions for comparison. Fig. 3 presents the performances of the defined imperceptibility metric compared with FeaSpaAdv.

Figs. 3(a)-(d) show that the red areas occupy large ratio of the entire image for various objects, i.e. FeaSpaAdv and the proposed entropy feature have a high overlap rate in the attack area of these images. Meanwhile, the attack areas generated with image entropy in (e) and (f) concentrate on the object edges compared with those with FeaSpaAdv, indicating that the image entropy method is more sensitive to the contour features and can find the edge area to attack.

Furthermore, we test the average runtime cost of the entropy feature and FeaSpaAdv, which are 24.73 and 760.52, respectively. Meanwhile, the overlap ratio (OR), i.e. 2x The (intersection area/union area) of the attackable regions by

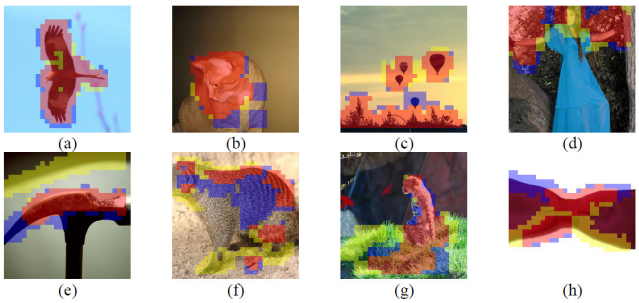

Fig. 3. Comparison of imperceptible regions based on FeaSpaAdv [18] and the proposed entropy feature. Red labels the common area generated by the two methods, blue and yellow represent the attack regions generated by SpaAdv and our method.

FeaSpaAdv and entropy feature×100%, between the resulted regions with the two kinds of features is 71.66%. Thus, the proposed entropy feature obtains the similar perturbation regions as FeaSpaAdv in terms of the overlap ratio, while requiring much less runtime cost.

3.1.2 Perturbation Regions. To compare different algorithms for obtaining the perturbation regions, the performances of the quadtree algorithm [13], end-to-end training [8], FeaSpaAdv and the watershed segmentation algorithm 1 are presented in Fig. 4.

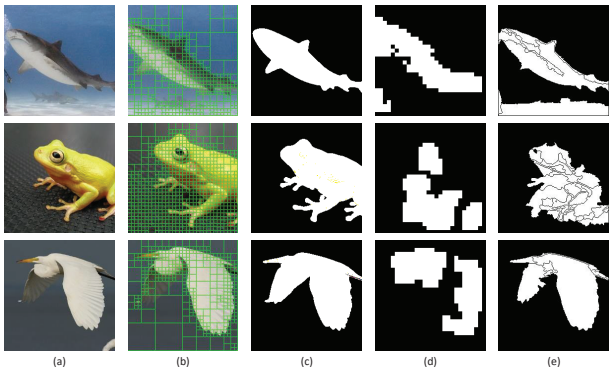

Fig. 4. The performances of different segmentation algorithms. (a) original image; (b) the candidate region with the quadtree algorithm [13]; (c) region segmentation with deep learning [8]; (d) the imperceptible reigons with FeaSpaAdv; (e) the regions with the proposed algorithm.

Fig. 4 shows that the quadtree algorithm [13] can also generate segmentation regions of different sizes, while it is sensitive to the size of the divided regions. Moreover, quadtree is based on the equally division of the entire image, which cannot well accommodate to the regions with complex contours. Semantic segmentation with an end-to-end training [8] can locate the object boundaries, while the detailed semantic regions can not be well explored. SpaAdv can locate the imperceptible regions, while these regions are also dependent on the block size, which may result in imperfect imperceptibility around object boundaries. By equipping semantic segmentation with entropy-based mask, the proposed algorithm with 'entropy+segmentation' (ENaSEG) can produce detailed semantic regions that are free to

the region size. Meanwhile, the accurate object boundaries based on the proposed algorithm allow the attack to well retain the edge information of different regions.

3.1.3 The Visualization Performance. In order to test the performance of the proposed constraint region obtained with 'entropy+segmentation' (ENaSEG), the attacked images with different features for determining the constraint regions are shown in Fig. 5. For this performance visualization, FGSM with global perturbations is used as the baseline attack for the visual performance evaluation of different region constraints. For multiple features in SpaAdv (FeaSpaAdv), we first obtain the mask region and use it to construct the corresponding constrained attack.

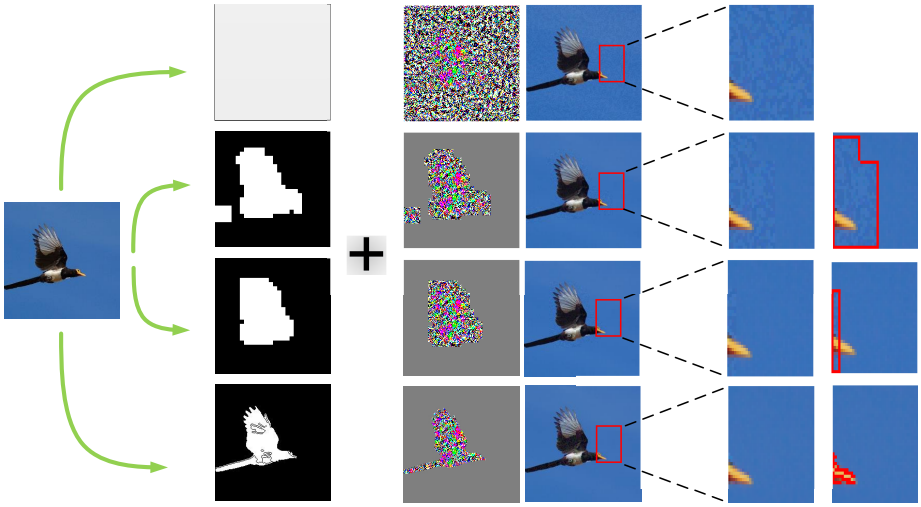

Fig. 5. Comparison of the attacked images with global attack (1st row), FeaSpaAdv (2nd row), only entropy feature (3rd row) and the proposed ENaSEG (4th row) based on FGSM attack. The red rectangles label the difference of the resulted constraint regions.

Fig. 5 shows that the perturbation noises obtained with the original FGSM distribute on the entire image and are easy to be perceive. The adversarial perturbations generated by our method fall in complex areas that are difficult for human eyes to perceive. As shown in the 2nd row, SpaAdv fails to distinguish the accurate imperceptible regions and produce relatively coarse region around the object boundary, while it performs similar as the only entropy feature in the 3rd row. Hence, the proposed ENaSEG produces perturbation noises that are more imperceptible than SpaAdv near object boundaries.

3.2 Quantitative Performance

In this section, we use two existing metrics, i.e. attack success rate and structural similarity [17] are employed to quantitatively evaluate the attack performances.

3.2.1 Attack Success Rates. Table 1 shows the success rates of the four attack methods with different region constraints and hyperparameter settings of τ. For each attack method, global noise (nomask), image entropy, multiple features in SpaAdv (FeaSpaAdv) and 'entropy+segmentation' (ENaSEG)-based attacks are used for comparison.

For FGSM that only obtains perturbation in single iteration, attacks based on image entropy or ENaSEG will lead to a significant decrease in terms of the success rate, when $\tau = 0.5$. Compared with SpaAdv, the attack based on image

Table 1. The attack success rates (%) of FGSM with different perturbation region constraints.

Global noise	Entropy	FeaSpaAdv	ENaSEG ($\tau = 0.5$)	($\tau = 0.4$)	($\tau = 0.3$)	($\tau = 0.2$)	($\tau = 0.1$)
98.17	88.63	84.84	79.35	83.40	86.93	91.77	94.90

entropy achieves a higher success rate. Due to the more strict limitation of the perturbed region, attack based on ENaSEG performs less well than that based on entropy feature. However, it is expected to improve the attack accuracy of ENaSEG-based by adjusting the perturbation regions dynamically.

Since the increase of the attack imperceptibility may in turn limit the attack success rate, the region ratio threshold, i.e. τ, is introduced to trade off the attack success rate and the imperceptibility degree. The attack success rates with different settings of τ in Table 1 show that, by gradually increasing the regions for perturbation, i.e. decreasing τ, the attack success rate of the constrained attack can be largely improved and approximate to that with global noise.

3.2.2 Results on SSIM. The structural similarity index measure (SSIM) is employed to judge the similarity of the benign and attacked images from the object structure information, which measures image quality toward the human eye's judgment. SSIM ranges in [0, 1], while the larger SSIM means the higher the similarity between two images. To calculate SSIM, an image is equally divided into small blocks, the average of SSIM values of all the blocks is used. Table 2 shows the results of SSIM with four attackers.

Table 2. The SSIM results based on FGSM.

Original FGSM	Entropy	SpaAdv	ENaSEG
0.59	0.91	0.92	**0.94**

Table 2 shows that the proposed algorithm greatly outperforms the original FGSM attack algorithm in terms of SSIM, i.e. the attacked images by our method behave higher structure similarity to the benign image than those with other algorithms, hence the proposed method can significantly improve the noise imperceptibility to human vision system. Compared to SpaAdv, the proposed algorithm also performs competitive in terms of SSIM, yet largely decreases the runtime cost as evaluated in Section 3.1.1.

4 DISCUSSION AND CONCLUSION

In this work, a new algorithm of imperceptible adversarial attack is proposed, which can well leverage the object semantic information for distributing the perturbation noises, so that the added perturbation is imperceptible to human vision system besides network imperceptibility. Experimental results based on FGSM attack reveal the both conciseness and effectiveness of the proposed algorithm.

Although competitive robustness is achieved by the proposed algorithm, there is still room for further improvement. First, the imperceptibility with style transfer [6] and the visibility of human vision system [18] can be considered to further enhance the imperceptibility for perturbation. Second, several hyper-parameters, e.g. τ in Algorithm 1 are introduced, which can be further made adaptive. Lastly, the proposed algorithm is expected to apply on more attack algorithms to evaluate its generalization performance.

ACKNOWLEDGMENT

The work was supported by Natural Science Foundation of China under grants no. 61602315, 91959108 and U1713214, the Science and Technology Project of Guangdong Province under grant no. 2020A1515010707, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20190808165203670.

REFERENCES

[1] Richard Barnes, Clarence Lehman, and David Mulla. 2014. Priority-flood: An optimal depression-filling and watershed-labeling algorithm for digital elevation models. *Computers & Geosciences* 62 (2014), 117–127.

[2] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP)* (2017), 39–57.

[3] Saheb Chhabra, Akshay Agarwal, Richa Singh, and Mayank Vatsa. 2021. Attack Agnostic Adversarial Defense via Visual Imperceptible Bound. In *International Conference on Pattern Recognition (ICPR)*. 5302–5309. https://doi.org/10.1109/ICPR48806.2021.9412663

[4] Yingpeng Deng and Lina J. Karam. 2020. Towards Imperceptible Universal Attacks on Texture Recognition. *CoRR* abs/2011.11957 (2020).

[5] X. Dong, J. Han, D. Chen, J. Liu, and N. Yu. 2020. Robust Superpixel-Guided Attentional Adversarial Attack. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] R. Duan, X. Ma, Y. Wang, J. Bailey, and Y. Yang. 2020. Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *CoRR* abs/1412.6572 (2015).

[8] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. 2018. Weakly-Supervised Semantic Segmentation Network With Deep Seeded Region Growing. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Alexey Kurakin, Ian J. Goodfellow, and S. Bengio. 2017. Adversarial examples in the physical world. *ArXiv* abs/1607.02533 (2017).

[10] Cassidy Laidlaw and Soheil Feizi. 2019. Functional Adversarial Attacks. In *Conference on Neural Information Processing Systems*.

[11] Hui Liu, Bo Zhao, Jiabao Guo, Yang An, and Peng Liu. 2020. GreedyFool: An Imperceptible Black-box Adversarial Example Attack against Neural Networks. *CoRR* abs/2010.06855 (2020).

[12] Bo Luo, Y. Liu, Lingxiao Wei, and Q. Xu. 2018. Towards Imperceptible and Robust Adversarial Example Attacks against Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[13] Z. F. Muhsin, A. Rehman, A. Altameem, T. Saba, and M. Uddin. 2014. Improved quadtree image segmentation approach to region information. *The Imaging Science Journal* (2014).

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[15] Vasu Sharma, A. Kalra, Vaibhav, Sumedha Chaudhary, L. Patel, and Louis-Phillippe Morency. 2018. Attend and Attack : Attention Guided Adversarial Attacks on Visual Question Answering Models. In *32nd Conference on Neural Information Processing Systems*.

[16] Christian Szegedy, W. Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and R. Fergus. 2014. Intriguing properties of neural networks. *CoRR* abs/1312.6199 (2014).

[17] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[18] Z. Wang, M. Song, S. Zheng, Z. Zhang, Y. Song, and Q. Wang. 2019. Invisible Adversarial Attack against Deep Neural Networks: An Adaptive Penalization Approach. *IEEE Transactions on Dependable and Secure Computing* PP, 99 (2019), 1–1.