DEFENDING AGAINST UNIVERSAL ATTACK VIA CURVATURE-AWARE CATEGORY ADVERSARIAL TRAINING

Peilun Du, Xiaolong Zheng, Liang Liu, Huadong Ma

{dupeilun1995, zhengxiaolong, liangliu, mhd}@bupt.edu.cn
Beijing Key Lab of Intelligent Telecommunication Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing 100876, China

ABSTRACT

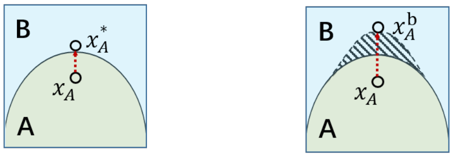
Adversarial training can defend against universal adversarial perturbation (UAP) by injecting corresponding adversarial samples during training. However, adversarial samples used by existing methods, such as UAP, inevitably include excessive perturbations related to other categories due to its inherent goal of universality. Training with them will cause more erroneous predictions with larger local positive curvature. In this paper, we propose a curvature-aware category adversarial training method to avoid excessive perturbations. We introduce the category-oriented adversarial masks that are synthesized with class distinctive momentum. Besides, we split the min-max optimization loops of adversarial training into two parallel processes to reduce the training cost. Experimental results on CIFAR-10 and ImageNet show that our method achieves better defense accuracy under UAP with less training cost than state-of-the-art baselines.

Index Terms— Adversarial defense, universal adversarial perturbation, adversarial training

1. INTRODUCTION

Existing works [1, 2, 3] have revealed that deep learning models are vulnerable to adversarial examples, in which small perturbations are added to the natural images to mislead the predictions of the target model. According to the effective scope of one perturbation, the adversarial attacks can be divided into per-instance attack [1] and universal attack [4, 5, 6, 7]. Per-instance attack generates image-specific perturbation for each image to mislead the model prediction. In contrast, universal attack generates universal adversarial perturbation (UAP) to fool the model for a broad class of images that can be crafted in advance, which is more likely to be encountered in practical vision applications [8]. Therefore, defending against UAPs is significant for improving adversarial robustness [8, 9, 10].

To conquer the UAP, recent study of adversarial training [11, 8, 9] improves the robustness of model itself under malicious UAPs. The concept of adversarial training first came from per-instance adversarial training [1], which injects the adversarial samples into conventional deep learning training, and then is boosted with many techniques [11, 12, 13, 14]. Specific to UAP, Mummadi et al. [8] propose shared adversarial training to utilize shared gradients of image heap. Shafahi et al. [9] defend against UAP by injecting FGSM-based UAP [1]. However, the methods above lack appropriate design of injected adversarial samples which induce confusing updates of the decision boundary curvature and more erroneous predictions on clean images.



(a) Desired sample x_A^* closes to the decision boundary.

(b) Excessive perturbations x_A^b cause larger positive curvature.

Fig. 1. The geometric illustration of curvature during adversarial training. Adversarial attacks (red dotted line) push x_A from image space of category A to B. The shaded part in Fig. 1 (b) is the updated decision boundary with larger positive curvature caused by x_A^b. The x_A^* in Fig. 1 (a) is desired.

Recent parallel work has proofed that the universal perturbations are more likely to fool the classifier of positive curvature [15, 16]. Though crucial, existing adversarial training methods lack efficient means to find desirable adversarial samples to approximate the curvature along with each category. We argue that injected adversarial samples include too many excessive perturbations into the adversarial samples that mislead the decision boundary curvature of the training model, then cause a decline in defending against UAP. As shown in Fig. 1, the geometric illustration of curvature during adversarial training demonstrates that training with adversarial samples close to the decision boundary (x_A^*) can prevent the model's decision boundary from becoming larger positive curvature. The existing injected adversarial samples x_A^b in Fig. 1 (b) will confuse the category decision boundary and cause larger positive curvature updated decision boundary, then decline the defending performance. Therefore, we need to find the training samples closer to the decision boundary to improve robustness under UAPs [15, 16].

In this paper, we propose a curvature-aware category adversarial training method that can learn a robust model by category-oriented UAP, which is called category-oriented

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

adversarial mask (CoAM). CoAM can obtain curvature information by approaching the actual decision boundary. To avoid excessive perturbations that confuse the category decision boundary, we first study and illustrate the geometric argument of adversaries used in adversarial training. Then, we design the CoAM which is synthesized perturbations by the images within the same category to reduce the excessive perturbations synthesized from other categories. Training with CoAM can avoid steeper updates of the decision boundary while retaining the image-agonistic feature in terms of universality. To craft CoAM, we introduce the class distinctive momentum on the approximate orthogonal direction towards the decision boundary. This momentum will force the direction of CoAM along with the adjacent category and leverage inter-sample relationships of the same category to find the common positive curvature parts of the adjacent decision boundary and flatten them. To reduce the computation cost of category adversarial training, we perform image-agnostic training by splitting the min-max optimization loops of adversarial training into CoAM generation and adversarial sample injection process. Experimental results on widely used datasets show that our method achieves impressive defense accuracy under universal attacks.

2. RELATED WORK

2.1. Adversarial Attack

Per-instance attacks. Fast Gradient Sign Method (FGSM) is a classical one-step attack [1] for a specific image. It finds the perturbation by maximizing gradient direction of cross-entropy loss by sign function. Iterative FGSM (I-FGSM) [17] and PGD [11] find the perturbation by splitting disturbance budget into small step size and adding perturbation step by step with clip operation. *Deepfool* performs adversarial attack with variable step size [18] to find optimal perturbation with minimal classification distance to other categories.

Universal attacks. Moosavi-Dezfooli et al. [4] propose classical UAP to generate fixed image-agnostic perturbation for attacking most images in the dataset. Mopuri et al. [6] improve the transferability of UAP via fooling the features learned at multiple layers. Zhang et al. [19] propose class discriminative universal adversarial perturbation (CD-UAP) to control UAP over the targeted classes.

2.2. Adversarial training for UAP

Adversarial training is an effective adversarial defense method that injects adversaries into the training process to improve the robustness of the model. Existing adversarial training methods can be divided into *per-instance* adversarial training and *universal* adversarial training. The early work of per-instance adversarial training uses FGSM to improve adversarial robustness [1]. Later, Madry et al. [11] propose multi-step PGD for training and achieve state-of-the-art robustness levels against adversarial attacks.

There have been few works on defending against UAP with adversarial training. Mummadi et al. [8] propose shared adversarial training to utilize shared gradients of image heap. However, they mainly focus on the maintenance of accuracy under clean images. Shafahi et al. defend UAP with universal adversarial training [9] on FGSM-based UAP. Philipp Benz et al. [20] use proxy data-based UAP [21] to achieve class-wise universal adversarial training. However, the above methods simply apply existing UAPs in adversarial training without an in-depth analysis of the relationship between the curvature of decision boundary and injected samples during adversarial training. Note that our approach is different from the above works. First, we explore the geometric argument of adversaries used in adversarial training. Second, we use class distinctive momentum to leverage the inter-sample relationships of the same category to generate CoAM.

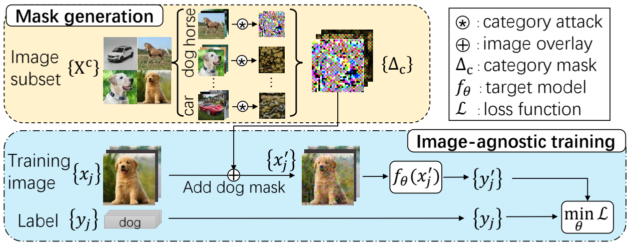


Fig. 2. **The framework of category adversarial training**, including mask generation and image-agnostic training.

3. ADVERSARIES ANALYSIS

To verify our hypothesis of excessive perturbations, we measure the reverse cost, RC, and boundary distance, BD, of adversaries in Table 1. We use target I-FGSM to reverse the adversaries into the original labels and record the iterative steps as reverse cost, RC. We set step size $\alpha = 1/255$. BD is used to estimate the semantic distance between the input to the actual adjacent decision boundary [22]. We generate a set of random directions in the input space, starting from a given image point. In each direction, we record the distance to the actual decision boundary when the prediction changes from the prediction on adversaries. We use 2 units for a step and perform about 1,000 random orthogonal directions. Smaller BD and RC mean more approaching decision boundaries.

	FGSM	PGD	Deepfool	F-UAP	Ours
RC	10.56	11.79	14.50	12.02	**9.79**
BD	1.68	1.48	**1.09**	-	1.38

Table 1. **Geometric arguments of adversaries on CIFAR-10.** *RC* is reverse cost, *BD* is boundary distance, *F-UAP* is the FGSM-based UAP.

As shown in Table 1, the results reveal that CoAM has the smallest RC and the second smallest BD. BD of FGSM based UAP [9], F-UAP, cannot be measured due to the most of the directions cannot reach the boundary in limited steps. Deepfool can approximate the boundary better with a smaller BD than CoAM but has more expensive overhead.

4. CATEGORY ADVERSARIAL TRAINING

Overview. As shown in Fig. 2, our category adversarial training splits the min-max adversarial training into two parts, mask generation, and image-agnostic training. The mask generation component generates CoAM Δ_c for each subset X^c

Algorithm 1: Updating CoAM

```
Input: Image $x$ with category $c$, model $f$, direction
       momentum $M$, epsilon $\epsilon$, old CoAMs $\Delta_c^{old}$
Output: Updated CoAMs $\Delta_c$
1  Set $\delta^c \leftarrow 0$, category flag $C \leftarrow c$, momentum $m \leftarrow 1$;
2  while $f(x+\delta_c) = c$ do
3      Get $\delta_j$, $c_{adj}$ at iteration $j$ by Eq. 2;
4      if $C = c_{adj}$ then
5          $m = Mm$;
6      else
7          $m = 1$;
8          $C = c_{adj}$;
9      $\delta^c = \delta^c + m \cdot \delta_j$;
10 $\Delta_c = \Delta_c + \delta^c$;
11 project $\Delta_c$ into $\epsilon$-ball according to $\mathcal{P}$;
12 return $\Delta_c$;
```

which specific to category c. We obtain the set of CoAMs $\{\Delta_c\}$ by Alg. 1. During the image-agnostic training, we add Δ_c to training image x_j to generate adversaries $x_j^{'}$. The training process will optimize the loss function $\mathcal{L}$ with the adversarial output $y_j^{'}$ of target model f_θ and true label y_j.

Generating CoAM. Our purpose is to generate CoAM that closes to the actual decision boundary for flattening the positive curvature. Traditional universal attack aims to attack all images with one image-agnostic universal perturbation. It has the highest degree of universality with the most excessive perturbations of other categories. We design a novel CoAM to balance the excessive perturbations and the degree of universality. CoAM integrates the local positive curvature of each image in a weighted form within the same category. Let $X^c = \{x_1^c, ..., x_n^c\}$ denotes a set of images of category c. Our method aims to find image-agnostic perturbations for each category to mislead all images in $\{X^c\}$:

$$f(x_i^c + \Delta_c) \neq f(x_i^c), x_i^c \in X^c \quad s.t. \quad ||\Delta_c||_\infty \leq \epsilon. \quad (1)$$

To obtain a common positive curvature part, we propose class distinctive momentum towards the decision boundary of a certain category instead of iterating trails of all the categories uniformly. Our design is based on the observation that the misclassification of samples is always traceable. In the real world, one category will always be mistaken for fixed several similar categories, such as truck and automobile or numbers 4 and 9. From the perspective of the neural network, these two categories belong to adjacent categories, the image data of these categories are close and their decision boundaries maybe local confusing (positive curvature). They are easy to be misclassified when the sample points close to this adjacent decision boundary. Therefore, adversaries labeled with the nearest adjacent category are close to the decision boundary with local positive curvature.

Based on this observation, we apply a cumulative momentum to the nearest adjacent category during the cumulative process of disturbance. By this means, CoAM can use decision boundary information to approach the decision boundary with local positive curvature more effectively than Deepfool or F-UAP. For j-th iteration on image x_i of category c:

$$\delta^c = Clip_{\delta,\epsilon}\{\delta^c + m \cdot \mathcal{D}(x_i)\}, x_i \in X^c, \quad (2)$$

where m is the class distinctive momentum, $\mathcal{D}$ is minimum adversarial perturbation with adjacent category c_{adj}:

$$\mathcal{D} = min_{f(x_i+\delta^c)\neq f(x_i)} \frac{|f(x_i+\delta^c) - f(x_i)|}{||\nabla f(x_i+\delta^c) - \nabla f(x_i)||_2^2}, \quad (3)$$

where ∇ is the gradient of classifier f. We select c_{adj} according to the ranking of adversarial perturbation, which is approximated by the ratio between model output and its category-oriented partial derivative. We update CoAM by $\Delta_c = \Delta_c + \delta^c$ and project Δ_c with budget ϵ:

$$\mathcal{P}(\Delta_c) = arg\,\underset{\Delta_c^{'}}{min}\,||\Delta_c - \Delta_c^{'}||_\infty \quad s.t.||\Delta_c||_\infty \leq \epsilon. \quad (4)$$

As shown in Alg. 1, we consider that the adjacent category is the target disturbance category. When the adjacent category is consistent with the last iteration, we accumulate momentum with multiplication (line 4-5). Otherwise, we reset the momentum to 1 on the iterative step (line 6-8). Then, we synthesize all δ for $x_j \in X^c$ to generate CoAM (line 9-11). This adjacent category-oriented attack will encourage adversaries to reach the most positive curvature area of the adjacent category. Our class distinctive momentum can accelerate the process of approaching the adjacent category decision boundary and significantly reduces generation cost as shown in Table 4.

Image-agnostic Adversarial Training. After obtaining the CoAM generated by a small subset of training data, we can apply the CoAM to the image-agnostic adversarial training on the whole training data. We split the two nested adversarial training loops of maximizing classification loss to generate adversaries and minimizing classification loss to update the model in parallel. The transformed optimation process is:

$$\underset{\theta}{min} \sum \underset{\delta<\epsilon}{max}\, \mathcal{L}(f_\theta(x+\delta^*), y), \quad (5)$$

where y is image label, $\mathcal{L}$ is loss function, δ is the adversarial perturbation, ϵ is the disturbance budget, into:

$$\begin{cases} \underset{\Delta_c<\epsilon}{max} \sum\limits_{x\in\{X^c\}} \mathcal{L}(f_\theta(x+\delta), y) \quad stage_1 \\ min_\theta \sum \mathcal{L}(f_\theta(x+\Delta_c), y) \quad stage_2, \end{cases} \quad (6)$$

where $\{X^c\}$ is image subset grouped by category c. $stage_1$ and $stage_2$ are category mask generation stage and image-agnostic training stage, respectively. We loop through each image in $\{X^c\}$ and synthesize CoAM according to the true labels at $stage_1$. If the class mask Δ_c fails to disturb the image $x \in \{X^c\}$, we will update Δ_c according to Alg. 1. The updated Δ_c will be injected in the following image-agnostic adversarial training. After the iteration of $\{X^c\}$, $stage_2$ is implemented during training to update the parameters θ.

5. EXPERIMENTS

Datasets. We conduct exclusive experiments on the benchmark datasets including CIFAR-10 [23] and ImageNet [24]. CIFAR-10 includes 10 categories with 32 × 32 resolution. ImageNet has 1,000 categories with 224 × 224 resolution.

		Clean	C-UAP	F-UAP	CD-UAP
Models trained with	Clean	96.4	11.7	13.7	10.2
	Madry	88.7	83.2	85.6	81.9
	SAT	93.2	86.5	88.7	84.4
	UAT	93.5	93.3	91.8	89.6
	Ours	**94.4**	**94.0**	**93.3**	**92.6**

Table 2. **Accuracy (%) comparison on CIFAR-10**. Baselines are *Clean*, *Madry*, *SAT*, and *UAT*.

Baselines. Baselines are (1) *Clean*, model trained with the standard approach using clean images. (2) *Madry*, model trained with PGD-based approach [11]. (3) *SAT*, model trained with shared adversarial traing [8]. (4) *UAT*, model trained with FGSM-based universal adversaries [9]. We demonstrate the defense performance under three widely used UAP, including classical UAP (C-UAP) [4], FGSM-based UAP (F-UAP) [9], and CD-UAP [19].

Training details. We use WideResnet 32-10 architecture on the CIFAR-10 [11] and AlexNet pre-trained model on ImageNet [9]. For CIFAR-10, the epoch is 200 and the data augmentation includes random crops with 4 pixels of padding and random horizontal flips during training. As for the ImageNet, the epoch is 30. The PGD perturbation budget for *Madry* is $8/255$ with step size $2/255$ of 7 iterations. We set the universal perturbation budget as $8/255$ for CIFAR-10 and $10/255$ for ImageNet. The initial learning rate is $10e-1$ and the weight decay is $2e-4$ for CIFAR-10, $10e-4$ for ImageNet. We generate CoAM every epoch in a subset of training data with 1,000 for CIFAR-10, 10,000 for ImageNet. The class distinctive momentum weight is 1.5 in this paper.

Results of CIFAR-10. The experimental results are shown in Table 2. It is obvious that the *Clean* fails under UAP and the accuracy is less than 10%. *Madry* achieves robustness accuracy with 85.6% under F-UAP. Under the clean images, it achieves 88.7% accuracy. Compared with *Madry*, *UAT* and *SAT* have an improvement but still worse than our method. Our method achieves the best accuracy under C-UAP (94.0%), F-UAP (93.3%), and CD-UAP (92.6%), receptively. Our method achieves higher accuracy than other baselines of prediction for clean images, about 94.4%, which proves CoAMs have excessive perturbations.

Results of ImageNet. We conduct experiments on ImageNet and the attack setting is the same as [9]. We calculate the top-1 and top-5 accuracy with the $10/255$ budget.

Trained with	Clean		C-UAP		F-UAP	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Clean	56.4	79.0	9.6	22.9	3.9	9.4
UAT	48.5	72.7	43.2	68.2	42.0	65.8
Ours	**49.6**	**73.5**	**47.8**	**72.4**	**46.2**	**71.6**

Table 3. **Accuracy (%) comparison under clean and different UAP on ImageNet.** Baselines include *Clean* and *UAT*.

The evaluation results are summarized in Table 3. Similar to the results on CIFAR-10, both the Top-1 and Top-5 accuracy of *Clean* are very low under different UAP. *UAT* improves the model robustness over the *Clean*. Our method achieves the best accuracy with 47.8%, 72.4% in C-UAP and 46.2%, 71.6% in F-UAP of Top-1 and Top-5 respectively.

Robustness and Generation Cost. To gather the robustness of a model's decision regions, we leverage BD in section 3 to estimate the distance from clean images to their decision boundary. We perform about 1,000 random orthogonal directions in 100 images on CIFAR-10. The results are shown in Table 4, the number of boundary distance represents the estimated robust distance to the adjacent decision boundaries. Therefore, the larger number of the BD means more robustness of the model when facing the adversarial attacks along each direction.

	Clean	Madry	SAT	UAT	Ours
BD	1.04	1.19	1.07	0.25	**1.42**
Cost (s)	-	-	-	2819	**316**

Table 4. **Comparison of *robust distance* and *training cost*.** Our method achieves the most robust model with the largest BD and minimum training cost.

To demonstrate the benefits of our splitting training, we compare the generation time of F-UAP with universal adversarial training with our CoAM on CIFAR-10 in 1 training epoch. The results are shown in Table 4, our CoAM achieves minimum training cost with 316s, which is 8.9× less than *UAT* [9] under the same device condition and batch size. Note that the adversarial samples in *Madry* and *SAT* are PGD-based per-instance adversarial attack, which has much larger generation time than F-UAP [9] and CoAM. Therefore, we leave out the training costs of them.

6. CONCLUSION

In this paper, we propose a curvature-aware category adversarial training method to defend against the UAP. We reveal that the excessive perturbations of different existing adversaries cause poor approximation of the actual decision boundary and larger local positive curvature. Based on this observation, we design CoAM that are synthesized the images within the category to eliminate the excessive perturbations while keeping the image-agonistic feature. We design the class distinctive momentum to effectively approach the desired adjacent decision boundary. Moreover, our method reduces the computation cost of adversarial training by splitting the min-max optimization loops into two parallel processes. The robust analysis and experimental results demonstrate that our method can achieve better training performance than existing methods under UAP.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No. 2018AAA0100500), the A3 Foresight Program of NSFC under Grant 62061146002, the Funds for Creative Research Groups of China under Grant 61921003, National Natural Science Foundation of China under Grant 61932013, and the 111 Project under Grant B18008.

7. REFERENCES

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[2] Yunhong Yin, Xiaolong Zheng, Peilun Du, Liang Liu, and Huadong Ma, "Scaling resilient adversarial patch," in *MASS*. 2021, pp. 189–197, IEEE.

[3] Xinyu Wang, Xiaolong Zheng, Peilun Du, Liang Liu, and Huadong Ma, "Occlusion resilient adversarial attack for person re-identification," in *MASS*, 2021, pp. 527–535.

[4] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[5] Valentin Khrulkov and Ivan V. Oseledets, "Art of singular vectors and universal adversarial perturbations," in *CVPR*, 2018, pp. 8562–8570.

[6] Konda Reddy Mopuri, Utsav Garg, and Venkatesh Babu Radhakrishnan, "Fast feature fool: A data independent approach to universal adversarial perturbations," in *BMVC*, 2017.

[7] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge J. Belongie, "Generative adversarial perturbations," in *CVPR*, 2018, pp. 4422–4431.

[8] Chaithanya Kumar Mummadi, Thomas Brox, and Jan Hendrik Metzen, "Defending against universal perturbations with shared adversarial training," in *ICCV*, 2019, pp. 4927–4936.

[9] Ali Shafahi, Mahyar Najibi, Zheng Xu, John P. Dickerson, Larry S. Davis, and Tom Goldstein, "Universal adversarial training," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. 2020, pp. 5636–5643, AAAI Press.

[10] Tejas S. Borkar, Felix Heide, and Lina J. Karam, "Defending against universal attacks through selective feature regeneration," in *CVPR*.

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[12] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein, "Adversarial training for free!," in *Advances in Neural Information Processing Systems*, 2019, pp. 3353–3364.

[13] Eric Wong, Leslie Rice, and J Zico Kolter, "Fast is better than free: Revisiting adversarial training," *arXiv preprint arXiv:2001.03994*, 2020.

[14] Haichao Zhang and Jianyu Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," in *Advances in Neural Information Processing Systems*, 2019, pp. 1829–1839.

[15] Saumya Jetley, Nicholas A. Lord, and Philip H. S. Torr, "With friends like these, who needs adversaries?," in *NeurIPS*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, Eds., 2018, pp. 10772–10782.

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto, "Robustness of classifiers to universal perturbations: A geometric perspective," in *ICLR*, 2018.

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[19] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon, "CD-UAP: class discriminative universal adversarial perturbation," in *AAAI*, 2020, pp. 6754–6761.

[20] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon, "Universal adversarial training with class-wise perturbations," pp. 1–6, 2021.

[21] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In So Kweon, "Understanding adversarial examples from the mutual influence of images and perturbations," in *CVPR*, 2020, pp. 14509–14518.

[22] Warren He, Bo Li, and Dawn Song, "Decision boundary analysis of adversarial examples," in *6th International Conference on Learning Representations*. 2018, OpenReview.net.

[23] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.