STOCHASTIC DEPTH BOOSTS TRANSFERABILITY OF NON-TARGETED AND TARGETED ADVERSARIAL ATTACKS

Chaoning Zhang[*], **Philipp Benz**[*], **Adil Karjauv**[*] **& In So Kweon**
Korea Advanced Institute of Science and Technology (KAIST)
{chaoningzhang1990, mikolez}@gmail.com, {pbenz, iskweon77}@kaist.ac.kr

ABSTRACT

Deep Neural Networks (DNNs) are widely known to be vulnerable to adversarial examples which have a surprising property of being transferable (or generalizable) to unknown networks. This property has been exploited in numerous works for achieving transfer-based black-box attacks. In contrast to most existing works that manipulate the image input for boosting transferability, our work manipulates the model architecture. Specifically, we boost the transferability with stochastic depth by randomly removing a subset of layers in networks with skip connections. Technical-wise, our proposed approach is mainly inspired by previous work improving the network generalization with stochastic depth. Motivation-wise, our approach of removing residual module instead of skip connection is inspired by the known finding that transferability of adversarial examples are positively related to local linearity of DNNs. The experimental results demonstrate that our approach outperforms existing methods by a large margin, resulting in SOTA performance, for both non-targeted and targeted attacks. Moreover, our approach is also complementary to the existing input manipulation approaches, combined with which the performance can be further boosted.

1 INTRODUCTION

Since the discovery of adversarial examples (Szegedy et al., 2013), numerous works have proposed various attack methods, both image-specific ones (Szegedy et al., 2013; Moosavi-Dezfooli et al., 2016; Carlini & Wagner, 2017; Goodfellow et al., 2015; Kurakin et al., 2017; Madry et al., 2018; Dong et al., 2018) and universal adversarial perturbations (Moosavi-Dezfooli et al., 2017; Poursaeed et al., 2018; Zhang et al., 2020b;a; Benz et al., 2020a; Zhang et al., 2021a; Benz et al., 2021; Zhang et al., 2021b). Adversarial attacks are often roughly divided into white-box attacks (Szegedy et al., 2013; Goodfellow et al., 2015; Madry et al., 2018) and black-box attacks (Dong et al., 2019; Wu et al., 2020). Black-box attacks constitute a larger threat since they do not require direct access to the target model. One widely recognized property of adversarial examples is their transferability (or generalization) capability of adversarial examples, *i.e.* adversarial examples generated on a substitute model are often also effective in fooling another, unknown target model. This property has been exploited in numerous works for developing transfer-based black-box attacks. Nonetheless, how to improve the transferability of adversarial examples remains an open question, especially for the more challenging targeted attack scenario.

Early works directly apply the white-box attack techniques, such as FGSM and I-FGSM, however, the transferability performance is not satisfactory. For example, I-FGSM often increases the attack success rate on the white-box model but at the cost of lower transferability to unknown target models. Multiple seminal works, such as MI-FGSM (Dong et al., 2018), DI-FGSM (Xie et al., 2019), and TI-FGSM (Dong et al., 2019), have attempted to enhance the transferability through some manipulation on the network input, showing significant success compared with the vanilla I-FGSM. In contrast to them, our work proposes to boost transferability through manipulating the model architecture. Specifically, we randomly drop some layers in networks with skip connections, such as ResNet.

[*] Equal Contribution

Our work is mainly inspired by two existing works: SGM (Wu et al., 2020) and stochastic depth (Huang et al., 2016). Residual networks (He et al., 2016a;b) have two information propagation pathways: the skip connection pathway and the residual module pathway. It has been shown in (Huang et al., 2016) that the gradient from the skip connection path results in significantly higher transferability than that of the residual module pathway. Somewhat surprisingly, only utilizing the gradients from the skip connection path in the backward propagation results in even higher transferability (refer to Figure 1 in SGM (Wu et al., 2020)). An insightful takeaway from their results is that skip connections are crucial for enhancing transferability. This also aligns well with the widely reported finding that utilizing networks with skip connections, such as ResNet and DenseNet, as the substitute models yield much higher transferability than those without skip connections, such as VGG or Inception family. In other words, networks with skip connections are more suitable for serving as a substitute model. The reason might be attributed to an early hypothesis established in (Goodfellow et al., 2015) that the transferability of adversarial examples roots in the “linear nature” of adversarial examples. This hypothesis was first supported by the success of their simple FGSM attack and has been revisited in one recent work (Guo et al., 2020) showing that back-propagating linearly increases the transferability, which constitutes another new empirical evidence that the transferability of adversarial examples is highly positively related to the “linearity”. Inspired by this, we propose to drop residual modules while keeping the (linear) shortcut path, which is, in essence, equivalent to dropping some layers during the optimization of the adversarial examples. (Huang et al., 2016) has proposed to train ResNets by dropping some layers resulting in a stochastic depth of the model during the training stage. The training with stochastic depth results in reduced training time due to the reduced depth during training while providing a non-trivial performance boost on the CIFAR10 and CIFAR100 datasets. However, their approach does not provide superior performance on the more large-scale ImageNet dataset. In contrast to their approach focusing on improving the network generalization performance, our work aims to improve the generalization (or transferability) of adversarial examples. Moreover, following previous works in transfer-based black-box attacks, our work conducts experiments on the ImageNet dataset and shows that this simple technique inspired by SGM (Wu et al., 2020) and stochastic depth (Huang et al., 2016) can significantly improve the transferability of adversarial examples.

2 Related Work

Black-box Attacks Depending on whether full knowledge of a target model is available to the attacker, most adversarial attacks are roughly categorized into white-box ones and black-box ones. White-box attacks are stronger but black-box attacks are often considered a more realistic attack. Black-box attacks can further be divided into query-based attacks and transfer-based attacks. Query-based ones (Chen et al., 2017; Papernot et al., 2017; Ilyas et al., 2019a; 2018; Guo et al., 2019; Tu et al., 2019; Shi et al., 2019; Rahmati et al., 2020), which require numerous queries, might be infeasible in practice if the API of the target model is not available or easily causes suspicion due to repeated query with the same adversarial example. The latter generated adversarial examples on a substitute model for attacking an unknown target model. Exploiting the prior of transfer-based attacks to reduce the number of queries has also recently been discussed in (Cheng et al., 2019; Yan et al., 2019; Huang & Zhang, 2020).

Adversarial Transferability. One property of adversarial examples is their transferability (Kurakin et al., 2017; Zhou et al., 2018), meaning that an adversarial example crafted on one model is also effective on another, previously unknown one. The adversarial vulnerability has been attributed to the existence of non-robust features (Ilyas et al., 2019b) and also transferability is believed to be related to the features that DNNs learn (Benz et al., 2020b; Ortiz-Jimenez et al., 2020). White-box attacks, such as FGSM and I-FGSM have been directly adopted in early investigation (Goodfellow et al., 2015; Kurakin et al., 2017; Tramèr et al., 2017) to demonstrate the transferability of adversarial examples. One explanation for this phenomenon has been attributed to the *linear nature* of deep classifiers, which is supported by both the success of FGSM as a single-step attack as well as the latter finding (Kurakin et al., 2017) that FGSM transfers better I-FGSM FGSM is found to transfer better than I-FGSM even though I-FGSM is stronger on the white-box model. Follow-up works enhance the transferability by ensembles of white-box models (Liu et al., 2017; Tramèr et al., 2018). Numerous works have then attempted to alleviate the over-fitting issue of I-FGSM and among them, there is a major line of work that manipulates the image input, such as MI-FGSM (Dong et al., 2018) introducing momentum on the input gradient, DI-FGSM (Xie et al., 2019) applying random resizing and padding on the input image, and TI-FGSM (Dong et al., 2019) smoothing the gradient with a convolution kernel. In contrast to them, our work proposes to manipulate the model architecture by randomly removing some layers of a model with skip connections. Technically, our approach is similar to the stochastic depth approach adopted in (Huang et al., 2016) for improving the network generalization while our work aims to improve the generalization of adversarial examples. Fine-tuning adversarial examples with the so-called intermediate-level attack has been discussed in (Huang et al., 2019; Li et al., 2020b). Our work is also highly related to recent works (Wu et al., 2020). Recently, adjusting the gradients in the backward propagation has been shown to improve the transferability by weighting the gradients through the shortcut and residual module (Wu et al., 2020; Guo et al., 2020) which exploits backpropagating with more linear gradients through either giving lower weight to the residual module or replacing the ReLU with an identity function. Most of the transfer-based works have mainly or exclusively studied non-targeted attacks, while the success of the targeted attack is still limited. (Li et al., 2020a) shows that adopting Poincaré ball distance and the triplet loss instead of the widely used CE loss can improve the performance. Another line of works (Inkawhich et al., 2019; 2020a;b) achieve transferable targeted attack by generating perturbation in feature space.

3 Methodology

Stochastic Depth. (Huang et al., 2016) has pointed out that training a deep ResNet can be time-consuming and they propose a straightforward approach to remove a subset of layers in the network during training. Since the to-be-removed layers are randomly chosen in each iteration, it also helps improve the model generalization. Inspired by their promising results, we conjecture that such an approach might also help improve the generalization, *i.e.* transferability, of adversarial examples. More clearly, (Huang et al., 2016) has the goal to increase the model generalization to unseen data, while ours is to enhance the adversarial example generalization to unseen models. We adopt the widely used cross-entropy loss to iteratively update the perturbation, however, in each iteration, the layers are randomly removed with a certain probability. Empirically, we find our approach achieves satisfactory performance when the probability is set to 0.2. In contrast to (Huang et al., 2016) that only investigates ResNet, inspired by the investigation in (Wu et al., 2020), we further evaluate our approach on DenseNet that also has dense skip connections. Note that to improve the generalization capability, we can also remove the skip connections instead of the residual module layers. Empirically, however, we find that it leads to much worse performance. This aligns well with the finding in (Wu et al., 2020) that the gradient from the skip connection pathway is more useful for enhancing the transferability. Different from (Wu et al., 2020) that only removes the residual module pathway in the backpropagation, our approach removes it in both forward and backward propagation. Similar to (Wu et al., 2020), our approach mainly works on substitute models with skip connections. This is practically meaningful as a substitute model, networks without skip connections, *e.g.* from the VGG or Inception family, perform much worse than those with skip connections (Inkawhich et al., 2019; 2020a).

Overall, we term our approach stochastic depth (SD), which is a technique complementary to the existing input manipulation techniques, such as MI-FGSM, DI-FGSM, TI-FGSM without causing any additional overhead. Due to the removed layers, our approach also helps to reduce the computation resources. We use MI-DI-TI-FGSM to indicate a method that simultaneously combines all the above three input manipulation techniques, and constitutes a very strong transfer-based attack method.

4 Experiments

4.1 Experimental Setup

Following previous works (Dong et al., 2018; 2019; Li et al., 2020a), we evaluate our proposed techniques on an ImageNet-compatible dataset composed of 1000 images. This dataset was originally introduced in the NeurIPS 2017 adversarial challenge. Throughout this work, following prior works we adopt the l_∞-norm constraint and set $l_\infty = 16/255$ for all our experiments. We set step size to 2 and set the number of iterations to 20 for non-targeted attack and 320 for the targeted attack, respectively. We experimented with 20/40/80/160/320 iterations for the targeted attack case and found a small number of iterations, such as 20, is not enough for generating transferable adversarial examples. To compare and use previous approaches, we follow the previous hyper-parameter settings as close as possible. Following (Dong et al., 2018) we set the momentum to 1, following (Xie et al., 2019) we set the probability of the stochastic input diversity p to 0.7, and adopt a kernel length of value 5 as in (Dong et al., 2019). For the non-targeted attack, following previous works, we report attack success rate (ASR). For the targeted attack, we mainly report the targeted ASR as the main evaluation metric; for completeness, (non-targeted) ASR is also reported.

4.2 Results

Transfer-Based Non-Targeted Attacks. The non-targeted results are shown in Table 1. Applying our proposed stochastic depth approach to the MI-FGSM attack, we observe that it can significantly outperform not only MI-FGSM but also DI-FGSM and TI-FGMS. The MI-FGSM with the stochastic gradient method (SGM) is also outperformed by the proposed stochastic depth by around 5% and 2% for substitute model ResNet50 and DenseNet121, respectively. We will now further test the performance of the stochastic depth approach for the targeted transferability attack scenario.

Table 1: Non-targeted ASR (%) for transfer-based non-targeted attacks with different substitute models and under different attack scenarios.

Substitute	Attack	DN121	VGG16	RN152	MNv2	IncV3	Avg.
RN50	MI	86.0	83.3	90.8	84.5	50.4	79.0
	DI	97.4	96.9	97.9	94.6	57.9	88.9
	TI	79.6	77.7	88.3	78.2	39.1	72.6
	MI-DI-TI	99.2	98.4	99.2	98.4	84.2	95.9
	MI (SGM)	92.9	90.7	96.2	91.1	60.4	86.3
	MI-DI-TI (SGM)	99.7	99.3	99.7	99.1	90.7	97.7
	MI (SD)	96.6	95.9	97.4	97.3	72.0	91.8
	MI-DI-TI (SD)	99.4	98.4	99.2	98.5	85.6	96.2
Substitute	Attack	RN50	VGG16	DN201	MNv2	IncV3	Avg.
DN121	MI	88.8	85.9	95.1	84.0	58.1	82.4
	DI	96.3	96.3	98.1	91.7	62.1	88.9
	TI	86.2	82.7	93.2	78.9	46.7	77.5
	MI (SGM)	95.8	93.4	98.2	89.9	70.8	89.6
	MI (SD)	96.9	95.1	98.2	95.5	71.1	91.4

Transfer-based Targeted Attacks. Overall, the attack scenario of transferable targeted adversarial examples constitutes a relatively more difficult one than the previously reported transferable non-targeted ones. The results of the adversarial examples crafted with stochastic depth compared to the baseline techniques are shown in Table 2. We observe, that for ResNet50 and DenseNet121 as the substitute model, MI-FGMS with the proposed SD outperforms all other attack methods in terms of non-targeted ASR and targeted ASR, including the MI-DI-TI attack. MI-DI-TI (SD) outperforms MI-DI-TI by a large margin from 25.3% to 58.8% (15.2% to 48.7%) for ResNet50 (DenseNet121) as the substitute model. Applying the SD to MI-DI-TI further significantly boosts the performance to 87.1% (72.6%) for ResNet50 (DenseNet121) as the substitute model.

Table 2: Targeted attack success rate (%) with a single substitute model.

Substitute	Attack	DN121	VGG16bn	RN152	MNv2	IncV3	Avg.
RN50	MI	47.8/1.3	56.5/0.5	48.9/1.8	64.5/0.5	26.3/0.1	48.8/0.8
	DI	57.7/14.1	71.6/13.6	54.2/13.8	64.1/2.9	25.6/0.2	54.6/8.9
	TI	39.7/0.4	46.3/0.5	42.3/1.6	52.7/0.4	21.4/0.0	40.5/0.6
	MI-DI-TI	84.2/40.2	88.6/28.0	82.6/43.1	84.7/10.4	52.9/4.6	78.6/25.3
	MI (SD)	90.8/75.1	89.2/59.4	95.4/86.9	91.0/56.2	59.9/16.4	85.3/58.8
	MI-DI-TI (SD)	99.0/95.6	98.9/92.1	99.1/96.7	98.7/90.5	86.6/60.6	96.5/87.1
Substitute	Attack	RN50	VGG16bn	DN201	MNv2	IncV3	Avg.
DN121	MI	58.0/1.9	59.7/0.4	51.7/4.4	70.3/0.4	31.4/0.1	54.2/1.4
	DI	60.0/9.0	67.5/8.2	57.4/20.9	63.5/2.5	28.7/0.0	55.4/8.1
	TI	49.4/1.6	52.0/0.5	41.6/3.6	56.5/0.5	26.7/0.1	45.2/1.3
	MI-DI-TI	78.4/16.4	80.3/11.3	79.2/40.5	83.6/5.3	51.6/2.3	74.6/15.2
	MI (SD)	89.0/64.6	87.4/46.3	95.0/88.7	84.2/31.1	57.9/13.0	82.7/48.7
	MI-DI-TI (SD)	96.8/85.0	95.6/77.2	98.6/95.9	94.3/62.7	79.2/42.1	92.9/72.6

5 Conclusion

In this work, we explore the application of stochastic depth through randomly removing a subset of layers in networks with skip connections. Our experimental results show the effectiveness of our approach, for example, we improve the targeted transferability performance on InceptionV3 from 4.6% to 60.6%.

REFERENCES

Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *ACCV*, 2020a.

Philipp Benz, Chaoning Zhang, and In So Kweon. Batch normalization increases adversarial vulnerability: Disentangling usefulness and robustness of model features. *arXiv preprint arXiv:2010.03316*, 2020b.

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Universal adversarial training with class-wise perturbations. *ICME*, 2021.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM workshop on artificial intelligence and security*, 2017.

Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *NeurIPS*, 2019.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Chuan Guo, Jacob R Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q Weinberger. Simple black-box adversarial attacks. *ICML*, 2019.

Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. *arXiv preprint arXiv:2012.03528*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016b.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016.

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *ICCV*, 2019.

Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. *ICLR*, 2020.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018.

Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *ICLR*, 2019a.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019b.

Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, 2019.

Nathan Inkawhich, Kevin J Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. *ICLR*, 2020a.

Nathan Inkawhich, Kevin J Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *NeurIPS*, 2020b.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017.

Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *CVPR*, 2020a.

Qizhang Li, Yiwen Guo, and Hao Chen. Yet another intermediate-level attack. In *ECCV*, 2020b.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *ICLR*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017.

Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Hold me tight! influence of discriminative features on deep network boundaries. In *NeurIPS*, 2020.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *ACM on Asia conference on computer and communications security*, 2017.

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *CVPR*, 2018.

Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: a geometric framework for black-box adversarial attacks. In *CVPR*, 2020.

Yucheng Shi, Siyu Wang, and Yahong Han. Curls & whey: Boosting black-box adversarial attacks. In *CVPR*, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *ICLR*, 2018.

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *AAAI*, 2019.

Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*, 2020.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019.

Ziang Yan, Yiwen Guo, and Changshui Zhang. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. *NeurIPS*, 2019.

Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Cd-uap: Class discriminative universal adversarial perturbation. In *AAAI*, 2020a.

Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *CVPR*, 2020b.

Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Universal adversarial perturbations through the lens of deep steganography: Towards a fourier perspective. *AAAI*, 2021a.

Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. *IJCAI*, 2021b.

Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *ECCV*, 2018.