ENCRYPTION INSPIRED ADVERSARIAL DEFENSE FOR VISUAL CLASSIFICATION

MaungMaung AprilPyone and Hitoshi Kiya

Tokyo Metropolitan University, Tokyo, Japan

ABSTRACT

Conventional adversarial defenses reduce classification accuracy whether or not a model is under attacks. Moreover, most of image processing based defenses are defeated due to the problem of obfuscated gradients. In this paper, we propose a new adversarial defense which is a defensive transform for both training and test images inspired by perceptual image encryption methods. The proposed method utilizes a block-wise pixel shuffling method with a secret key. The experiments are carried out on both adaptive and non-adaptive maximum-norm bounded white-box attacks while considering obfuscated gradients. The results show that the proposed defense achieves high accuracy ($91.55\,\%$) on clean images and ($89.66\,\%$) on adversarial examples with noise distance of $8/255$ on CIFAR-10 dataset. Thus, the proposed defense outperforms state-of-the-art adversarial defenses including latent adversarial training, adversarial training and thermometer encoding.

Index Terms— Adversarial defense, adversarial machine learning, perceptual image encryption

1. INTRODUCTION

Security in computer vision systems is quintessential and high in demand. This is because computer vision technology has been deployed in many applications including safety and security critical applications such as self-driving cars, healthcare, facial recognition, etc. and many more visual recognition systems. Computer vision systems are primarily powered by deep neural networks (DNNs). It is proven that DNNs have brought impressive state-of-the-art results to computer vision. However, researchers have already discovered that neural networks in general are vulnerable towards certain alteration in the input known as adversarial examples [1, 2]. These adversarial examples can cause neural networks misclassify or force to classify a targeted class with high confidence. Incorrect decisions made by DNNs can cause serious and dangerous problems. As an example, self-driving cars may misclassify “Stop” sign as “Speed Limit” [3]. Due to this threat, adversarial machine learning research has got a significant amount of attention recently although it has been started over a decade ago [4].

Researchers have proposed various attacks and defenses. Ideally, provable robust models are desired. Inspiring works such as [5–7] proposed provable secure training. Although these methods are attractive and desirable, they are not available for larger datasets. One recent work [8] scaled up to CIFAR-10 [9] dataset in provable defense research. However, the accuracy is not comparable even on low adversarial noise distance. There is also an alternative approach to find a defensive transform $t(\cdot)$ so that the prediction of a classifier $f(\cdot)$ on clean image x is equal to that of an adversarial example x' (i.e., $f(x) = f(t(x')))$. Such works include [10–14], etc. They all have been defeated when accounting for obfuscated gradients (a way of gradient masking) [15]. To reinforce these weak defense methods, Raff et al. [16] proposed a stronger defense by combining a large number of transforms stochastically. However, applying many transforms drop in accuracy even though the model is not under attack and is computationally expensive. Our previous work removes adversarial noise generated on one-bit images by double quantization [17], but, clean images are limited to be in one-bit.

Therefore, in this work, we propose a new adversarial defense which has been inspired by perceptual image encryption methods [18–21]. It was reported that [20] can be used as a defensive transform [22]. However, it is not meant for adversarial defense and reduces accuracy. To defend adversarial examples and maintain high accuracy, we design a defensive transform that uses a block-wise pixel shuffling method. Similar to our work, Taran et al. proposed a key-based adversarial defense [23]. The main intellectual differences include: (1) the proposed defense is inspired by perceptual image encryption (specifically, block-wise image encryption), in contrast to traditional cryptographic methods and (2) we consider white-box attacks unlike the work by [23] that considered gray-box attacks. In an experiment, the proposed defense is confirmed to outperform state-of-the-art adversarial defenses including latent adversarial training, adversarial training and thermometer encoding under maximum-norm bounded threat model with the noise distance of $8/255$ on CIFAR-10 dataset.

2. PRELIMINARIES

2.1. Adversarial Examples

An adversarial example is a modified input x' (visually similar to x) to a classifier $f(\cdot)$ aiming $f(x) \neq f(x')$. An attacker

978-1-7281-6395-6/20/$31.00 ©2020 IEEE

finds perturbation δ under certain distance metric (usually ℓ_p norm) to construct an adversarial example. An attack algorithm usually minimizes the perturbation or maximizes the loss function, i.e.,

$$\underset{\delta}{\text{minimize}} \, \|\delta\|_p \,, \quad \text{s.t.} \;\; f(x+\delta) \neq y, \text{or} \tag{1}$$

$$\underset{\delta \in \Delta}{\text{maximize}} \, \mathcal{L}(f(x+\delta), y), \tag{2}$$

where $\Delta = \{\delta : \|\delta\|_p \leq \epsilon\}$. There are many attack algorithms such as Fast Gradient Sign Method (FGSM) [24], Projected Gradient Descent (PGD) [25], Carlini and Wagner (CW) [26], etc.

2.2. Threat Model

Following [27] and [2], first, we describe a threat model that we use to evaluate the proposed defense. We deploy PGD [25] because it is one of the strongest attacks under ℓ_∞ norm bounded metric.

Based on the goal of an adversary, the attack can be whether targeted ($f(x') = z$ where z is a class targeted by the adversary) or untargeted ($f(x') \neq y$ where y is a true class). We focus on untargeted attacks under $\|x' - x\|_\infty \leq \epsilon$, where ϵ is a given noise distance.

We evaluate the proposed defense in white-box settings. Therefore, we assume this adversary has full knowledge of the model, its parameters, trained weights, training data and the proposed defense mechanism except a secret key.

The adversary performs evasion attacks (i.e., test time attacks) in which small changes under ℓ_∞ metric change the true class of the input. The adversary's capability is to modify the test image where the noise distance is ϵ in the range of $[2/255, 32/255]$. Having full knowledge of the defense transform, our adversary also extends PGD. Fully accounting obfuscated gradients, the adversary implements an adaptive attack like Backward Pass Differentiable Approximation (BPDA) [15] to estimate the correct gradients with a guessed key.

3. PROPOSED METHOD

3.1. Overview

The goal of the proposed method is to hold high accuracy whether or not the model is under adversarial attacks. The overview of the proposed defense is depicted in Fig. 1. Training images are transformed by a secret key and a model is trained by the transformed images. Test images regardless of being clean images or adversarial examples are also transformed with the same key before classification process by the model.

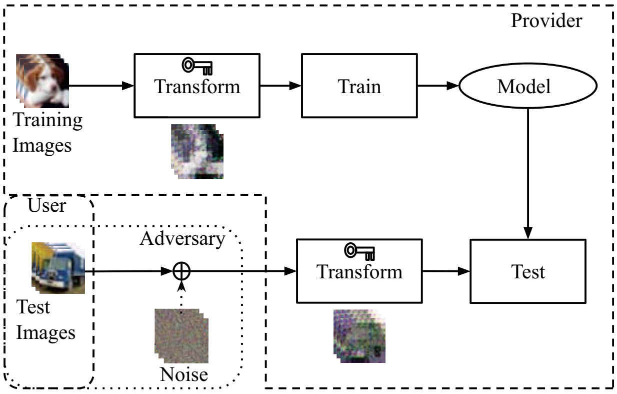

Fig. 1. Overview of the proposed defense.

3.2. Defensive Transform

We introduce a transform that exploits block-wise pixel shuffling with a secret key as an adversarial defense for the first time. Both training and test images are transformed with a common key. The transformation process is as follows.

A 3-channel (RGB), 8-bit image with a dimension of $X \times Y \times 3$ is divided into blocks (with the size of $M \times M \times 3$) where X and Y should be divisible by M. Otherwise, padding is required.

Let $p(i)$ and n be the pixel value and the number of pixels in each block (i.e., $M \times M \times 3$), where $i \in \{0, \ldots, n-1\}$. The new pixel value $p'(i)$ is given by

$$p'(i) = p(\alpha(i)), \tag{3}$$

where $\alpha = [\alpha(0), \alpha(1), \ldots, \alpha(n-2), \alpha(n-1)]$ is a random permutation vector of the integers from 0 to $n-1$ generated by a key K. Fig. 2 illustrates the process of block-wise pixel shuffling. The process is repeated for all the blocks in the image.

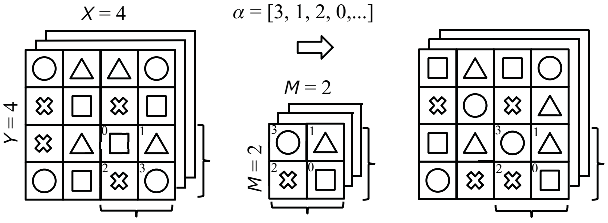

Fig. 2. Process of block-wise pixel shuffling.

3.3. Adaptive Attack

As pointed out by [27] and [2], adaptive attacks are necessary in evaluating adversarial defenses. Several recent defenses are defeated by adaptive attacks due to obfuscated gradients [15]. To ensure the strength of the proposed defense, we implement a BPDA-like attack so that the gradients are correct with respect to the attacker's guessed key as shown in Fig. 3. Basically, the adversary applies block-wise shuffling to a test image with a key, PGD is run on the shuffled image and the resulting adversarial example is de-shuffled with the adversary's assumed key. We used random keys to attack the proposed method in our experiments.

3.4. Key Management

The proposed method uses a shared secret key K to all the blocks in each of both training and test images. Its key space is defined as follows:

$$\mathcal{K}(n) = n!, \tag{4}$$

where n is the number of pixels in a block. Deep learning is often done in the cloud server (provider) and the key K should be saved securely at the server in deploying the proposed method.

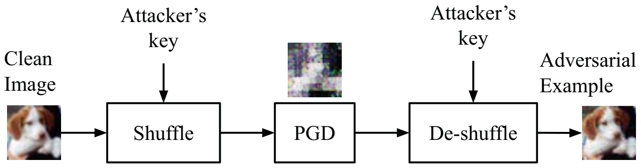

Fig. 3. Diagram of adaptive attack.

4. EXPERIMENTS

4.1. Setup

We used CIFAR-10 [9] dataset with a batch size of 128 and live augmentation (random cropping with padding of 4 and random horizontal flip) on training set. CIFAR-10 consists of 60,000 color images (dimension of $32 \times 32 \times 3$) with 10 classes (6000 images for each class) where 50,000 images are for training and 10,000 for testing. Both training and test images were preprocessed by the proposed method with a common shared secret key K.

The deep residual network [28] with 18 layers (ResNet18) was trained for 160 epochs by the stochastic gradient descent optimizer. The parameters are: momentum of 0.9, weight decay of 0.0005 and initial learning rate of 0.1. A step learning rate scheduler was used with the settings (lr_steps $= 40$, gamma $= 0.1$).

The parameters of PGD adversary are ϵ in the range of $[2/255, 32/255]$, and $\alpha = 2/255$. The attack was run for 20 and 40 iterations with/without random initialization. When random initializaition is set, perturbation is initialized with random values bounded by given ϵ.

Table 1. Accuracy of the proposed method under the use of various block sizes on PGD$_{20}$ ($\epsilon = 32/255$)

$M \times M$	Clean	PGD$_{20}$
2×2	**0.9408**	0.7157
4×4	0.9155	**0.8472**
8×8	0.8540	0.7892
16×16	0.7351	0.6756

We used publicly available ResNet18 implementation [29] on PyTorch. The proposed method was implemented by modifying the code base of [20]. We deployed traditional PGD implementation from [30] and implemented BPDA-like attack to make the adversary adaptive and effective.

4.2. Results

4.2.1. PGD Attack on Various Block Sizes

We evaluated the proposed method under the use of various block sizes, $M \in \{2, 4, 8, 16\}$ by PGD. We trained ResNet18 with images transformed by the proposed method with different block size $M \in \{2, 4, 8, 16\}$ resulting four models. The trained models were first attacked by PGD with $\epsilon = 32/255$ for 20 iterations (i.e., PGD$_{20}$) without random initialization.

Table 1 summarizes the results obtained from the experiment of the proposed method. The model trained with transformed images where $M = 2$ gave the best performance ($94.08\,\%$) when the model is not under attacks. However, $M = 4$ performed better under attacks (i.e., $84.72\,\%$). The results suggest that $M = 4$ provides the best overall performance.

4.2.2. PGD Attack in Various Settings

We further ran PGD attacks with various settings to the model trained by the proposed defense where $M = 4$. The attacks were executed for 20 and 40 iterations, and subscript r denotes random initialization (e.g., PGD$_{20r}$ stands for PGD attack for 20 iterations with random initialization and BPDA denotes the adaptive attack).

Table 2 captures the results of untargeted attacks where $\epsilon = 8/255$ and $32/255$. When $\epsilon = 8/255$, the model maintain $91.55\,\%$ accuracy on clean images and $89.66\,\%$ on BPDA$_{40r}$ attack. This confirms that the adaptive attack cannot reduce the accuracy when the attacker's key is not correct. However, when ϵ was increased to $32/255$, BPDA$_{40r}$ reduced the accuracy to $61.60\,\%$.

Our experiments show that BPDA$_{40r}$ is a better adversary. Therefore, we evaluated the proposed defense with various $\epsilon \in \{2/255, 4/255, 8/255, 16/255, 32/255\}$ by BPDA$_{40r}$. Moreover, to confirm the effectiveness of the proposed method, we also implemented state-of-the-art adversarial defense method, i.e., adversarial training (AT) [25]

Table 2. Accuracy of the proposed method ($M = 4$) by PGD attack in various settings

Epsilon ϵ	Clean	PGD_{20}	PGD_{20r}	PGD_{40}	PGD_{40r}	BPDA_{20}	BPDA_{20r}	BPDA_{40}	BPDA_{40r}
$8/255$	$\mathbf{0.9155}$	0.8948	0.8977	0.8917	0.8931	0.8988	0.8988	0.8988	$\mathbf{0.8966}$
$32/255$	$\mathbf{0.9155}$	0.8472	0.6645	0.7770	0.6323	0.8204	0.6505	0.7346	0.6160

on the same network specifications with $\epsilon = 8/255$ to compare the results. The accuracy versus various noise distances is plotted in Fig. 4. When $\epsilon < 8/255$, the model trained by the proposed defense provides more than $90\,\%$ accuracy. The accuracy gradually drops when ϵ is greater than $8/255$. Specifically, when $\epsilon = 16/255$, the model achieves $\approx 83\,\%$ accuracy. On the worst case scenario (i.e. $\epsilon = 32/255$), the accuracy of the model is $\approx 62\,\%$. Nevertheless, the proposed method outperforms AT in any given perturbation budget as shown in Fig. 4.

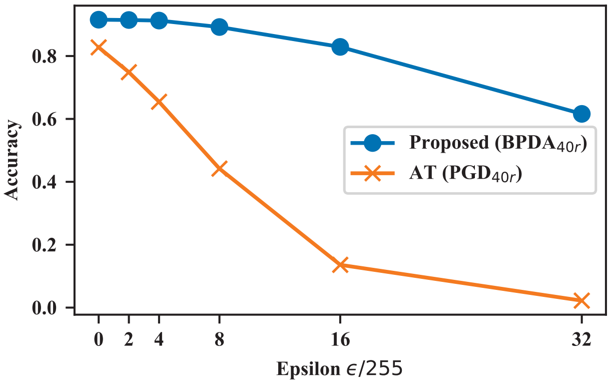

Fig. 4. Accuracy vs. perturbation budget.

4.3. Comparison with State-of-the-art Defenses

To confirm the effectiveness of the proposed defense, we made a comparison with state-of-the-art published defenses for CIFAR-10 dataset on RobustML catalog[1]. We compared the proposed defense with the recent three defenses: latent adversarial training (LAT) [31], adversarial training (AT) [25] and thermometer encoding (TE) [10]. All three defenses used wide residual network [32] and were evaluated on ℓ_∞ threat model with $\epsilon = 8/255$ except LAT (used $\epsilon = 0.03$). Table 3 shows the summary of the comparison. The proposed model was trained on ResNet18 and achieves superior accuracy (i.e., $91.55\,\%$ on clean images and $89.66\,\%$ on attacked ones). Even on the worst case scenario (i.e., $\epsilon = 32/255$), the accuracy of the proposed method was still higher than the state-of-the-art defenses whether or not the model was under attacks.

[1] https://www.robust-ml.org/

Table 3. Comparison with state-of-the-art defenses on CIFAR-10 dataset

Defense	Threat Model	Clean	Attacked
LAT [31]	$\ell_\infty(\epsilon = 0.03)$	87.80	53.82
AT [25]	$\ell_\infty(\epsilon = 8/255)$	87.00	46.00
TE [10]	$\ell_\infty(\epsilon = 8/255)$	90.00	30.00
Proposed	$\ell_\infty(\epsilon = 8/255)$	$\mathbf{91.55}$	$\mathbf{89.66}$
Proposed	$\ell_\infty(\epsilon = 16/255)$	$\mathbf{91.55}$	$\mathbf{82.90}$
Proposed	$\ell_\infty(\epsilon = 32/255)$	$\mathbf{91.55}$	$\mathbf{61.60}$

5. CONCLUSION

In this paper, we proposed a new adversarial defense that utilizes a key-based block-wise pixel shuffling method as a defensive transform for the first time. Specifically, both training and test images are transformed by the proposed method with a common key before training and testing. We also implemented an adaptive attack to verify the strength of the proposed defense. Our experiments suggest that the proposed defense is resistant to both adaptive and non-adaptive attacks. The results show that the proposed defense achieves higher accuracy, $91.55\,\%$ on clean images and $89.66\,\%$ on adversarial examples. Compared to state-of-the-art defenses, the accuracy of the proposed method is $35.84\,\%$ better than latent adversarial training, $43.66\,\%$ than adversarial training and $59.66\,\%$ than thermometer encoding under a maximum-norm bounded white-box threat model with the noise distance of $8/255$ on CIFAR-10 dataset.

6. REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, 2014.

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli, “Evasion attacks against machine learning at test time,” in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[3] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song, “Robust physical-world attacks on deep learning visual classification,” in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.

[4] Battista Biggio and Fabio Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[5] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang, "Certified defenses against adversarial examples," in *International Conference on Learning Representations*, 2018.

[6] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy A Mann, and Pushmeet Kohli, "A dual approach to scalable verification of deep networks.," in *UAI*, 2018, vol. 1, p. 2.

[7] Eric Wong and J. Zico Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 5283–5292.

[8] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter, "Scaling provable adversarial defenses," in *Advances in Neural Information Processing Systems*, 2018, pp. 8400–8409.

[9] Alex Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[10] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in *International Conference on Learning Representations*, 2018.

[11] Pouya Samangouei, Maya Kabkab, and Rama Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *International Conference on Learning Representations*, 2018.

[12] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten, "Countering adversarial images using input transformations," in *International Conference on Learning Representations*, 2018.

[13] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille, "Mitigating adversarial effects through randomization," in *International Conference on Learning Representations*, 2018.

[14] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," in *International Conference on Learning Representations*, 2018.

[15] Anish Athalye, Nicholas Carlini, and David A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 274–283.

[16] Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean, "Barrage of random transforms for adversarially robust defense," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6528–6537.

[17] MaungMaung AprilPyone, Yuma Kinoshita, and Hitoshi Kiya, "Adversarial robustness by one bit double quantization for visual classification," *IEEE Access*, vol. 7, pp. 177932–177943, 2019.

[18] Tatsuya Chuman, Warit Sirichotedumrong, and Hitoshi Kiya, "Encryption-then-compression systems using grayscale-based image encryption for jpeg images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1515–1525, 2019.

[19] Warit Sirichotedumrong and Hitoshi Kiya, "Grayscale-based block scrambling image encryption using ycbcr color space for encryption-then-compression systems," *APSIPA Transactions on Signal and Information Processing*, vol. 8, 2019.

[20] Masayuki Tanaka, "Learnable image encryption," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, 2018, pp. 1–2.

[21] Warit Sirichotedumrong, Yuma Kinoshita, and Hitoshi Kiya, "Pixel-based image encryption without key management for privacy-preserving deep neural networks," *IEEE Access*, vol. 7, pp. 177844–177855, 2019.

[22] MaungMaung AprilPyone, Warit Sirichotedumrong, and Hitoshi Kiya, "Adversarial test on learnable image encryption," in *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, 2019, pp. 693–695.

[23] Olga Taran, Shideh Rezaeifar, and Slava Voloshynovskiy, "Bridging machine learning and cryptography in defence against adversarial attacks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[24] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[26] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, May 2017, pp. 39–57.

[27] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] Kuangliu, "Train cifar10 with pytorch," https://github.com/kuangliu/pytorch-cifar, 2017.

[30] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin, "AdverTorch v0.1: An adversarial robustness toolbox based on pytorch," *arXiv preprint arXiv:1902.07623*, 2019.

[31] Nupur Kumari, Mayank Singh, Abhishek Sinha, Harshitha Machiraju, Balaji Krishnamurthy, and Vineeth N Balasubramanian, "Harnessing the vulnerability of latent layers in adversarially trained models," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 7 2019, pp. 2779–2785.

[32] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, September 2016, pp. 87.1–87.12.