ME-Net: Towards Effective Adversarial Robustness with Matrix Estimation

Yuzhe Yang [1] **Guo Zhang** [1] **Dina Katabi** [1] **Zhi Xu** [1]

Abstract

Deep neural networks are vulnerable to adversarial attacks. The literature is rich with algorithms that can easily craft successful adversarial examples. In contrast, the performance of defense techniques still lags behind. This paper proposes ME-Net, a defense method that leverages matrix estimation (ME). In ME-Net, images are preprocessed using two steps: first pixels are randomly dropped from the image; then, the image is reconstructed using ME. We show that this process destroys the adversarial structure of the noise, while re-enforcing the global structure in the original image. Since humans typically rely on such global structures in classifying images, the process makes the network mode compatible with human perception. We conduct comprehensive experiments on prevailing benchmarks such as MNIST, CIFAR-10, SVHN, and Tiny-ImageNet. Comparing ME-Net with state-of-the-art defense mechanisms shows that ME-Net consistently outperforms prior techniques, improving robustness against both black-box and white-box attacks.

1. Introduction

State-of-the-art deep neural networks (NNs) are vulnerable to adversarial examples (Szegedy et al., 2013). By adding small human-indistinguishable perturbation to the inputs, an adversary can fool neural networks to produce incorrect outputs with high probabilities. This phenomena raises increasing concerns for safety-critical scenarios such as the self-driving cars where NNs are widely deployed.

An increasing body of research has been aiming to either generate effective perturbations, or construct NNs that are robust enough to defend against such attacks. Currently, many effective algorithms exist to craft these adversarial examples, but defense techniques seem to be lagging behind. For instance, the state-of-the-art defense can only achieve less than 50% adversarial accuracy for ℓ_∞ perturbations on datasets such as CIFAR-10 (Madry et al., 2017). Under recent strong attacks, most defense methods have shown to break down to nearly 0% accuracy (Athalye et al., 2018).

As adversarial perturbations are carefully generated structured noise, a natural conjecture for defending against them is to destroy their structure. A naive approach for doing so would randomly mask (i.e., zero out) pixels in the image. While such method can eliminate the adversarial structure within the noise through random information drop, it is almost certain to fail since it equally destroys the information of the original image, making NN inference even worse.

However, this naive starting point raises an interesting suggestion: instead of simply applying a random mask to the images, a preferable method should also reconstruct the images from their masked versions. In this case, the random masking destroys the crafted structures, but the reconstruction recovers the global structures that characterize the objects in the images. Images contain some global structures. An image classified as cat should have at least a cat as its main body. Humans use such global structure to classify images. In contrast the structure in adversarial perturbation is more local and defies the human eye. If both training and testing are performed under the same underlying global structures (i.e., there is no distributional shift in training and testing), the network should be generalizable and robust. If the reconstruction can successfully maintain the underlying global structure, the masking-and-reconstruction pipeline can redistribute the carefully constructed adversarial noises to non-adversarial structures.

In this paper, we leverage matrix estimation (ME) as our reconstruction scheme. ME is concerned with recovering a data matrix from noisy and incomplete observations of its entries, where exact or approximate recovery of a matrix is theoretically guaranteed if the true data matrix has some *global structures* (e.g., low rank). We view a masked adversarial image as a noisy and incomplete realization of the underlying clean image, and propose ME-Net, a preprocessing-based defense that reverts a noisy incomplete image into a denoised version that maintains the underlying global structures in the clean image. ME-Net realizes adversarial robustness by using such denoised global-structure preserving representations.

[1]MIT CSAIL, Cambridge, MA, USA. Correspondence to: Yuzhe Yang <yuzhe@mit.edu>, Zhi Xu <zhixu@mit.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

We note that the ME-Net pipeline can be combined with different training procedures. In particular, we show that ME-Net can be combined with standard stochastic gradient descent (SGD) or adversarial training, and in both cases improves adversarial robustness. This is in contrast with many preprocessing techniques which cannot leverage the benefits of adversarial training (Buckman et al., 2018; Song et al., 2018; Guo et al., 2017), and end up failing under the recent strong white-box attack (Athalye et al., 2018).

We provide extensive experimental validation of ME-Net under the strongest black-box and white-box attacks on established benchmarks such as MNIST, CIFAR-10, SVHN, and Tiny-ImageNet, where ME-Net outperforms state-of-the-art defense techniques. Our implementation is available at: `https://github.com/YyzHarry/ME-Net`.

We summarize our contributions as follows:

- We are the first to leverage matrix estimation as a general pipeline for image classification and defending against adversarial attacks.
- We show empirically that ME-Net improves the robustness of neural networks under various ℓ_∞ attacks:
  1. ME-Net alone significantly improves the state-of-the-art results on black-box attacks;
  2. Adversarially trained ME-Net consistently outperforms the state-of-the-art defense techniques on white-box attacks, including the strong attacks that counter gradient obfuscation (Athalye et al., 2018).

  Such superior performance is maintained across various datasets: CIFAR-10, MNIST, SVHN and Tiny-ImageNet.
- We show additional benefits of ME-Net such as improving generalization (performance on clean images).

2. ME-Net

We first describe the motivation and high level idea underlying our design. We then provide the formal algorithm.

2.1. Design Motivation

Images contain noise: even "clean" images taken from a camera contain white noise from the environment. Such small, unstructured noise seems to be tolerable for modern deep NNs, which achieve human-level performance. However, the story is different for carefully constructed noise. Structured, adversarial noise (i.e., adversarial examples) can easily corrupt the NN results, leading to incorrect prediction from human's perspective. This means that to achieve robustness to adversarial noise, we need to eliminate/reduce the crafted adversarial structure. Of course, while doing so, we need to maintain the intrinsic structures in the image that allow a human to make correct classifications.

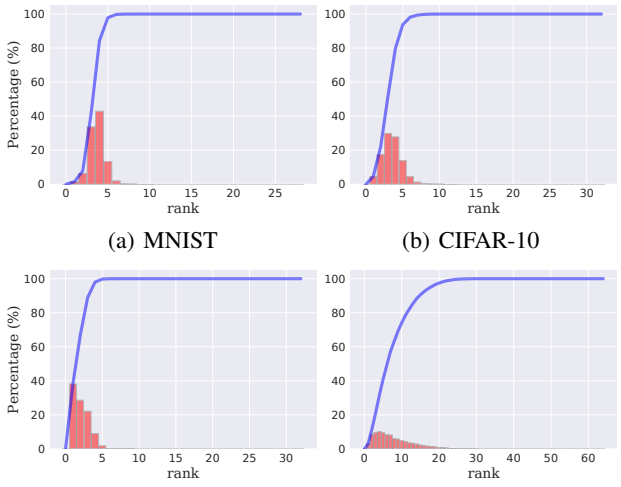
We can model the problem as follows: An image is a superposition of: 1) intrinsic true structures of the data in the scene, 2) adversarial carefully-structured noise, and 3) non-adversarial noise. Our approach is first to destroy much of the crafted structure of the adversarial noise by randomly masking (zeroing out) pixels in the image. Of course, this process also increases the overall noise in the image (i.e., the non-adversarial noise) and also negatively affects the underlying intrinsic structures of the scene. Luckily however there is a well-established theory for recovering the underlying intrinsic structure of data from noisy and incomplete (i.e., masked) observations. Specifically, if we think of an image as a matrix, then we can leverage a well-founded literature on matrix estimation (ME) which allows us to recover the true data in a matrix from noisy and incomplete observations (Candès & Recht, 2009; Keshavan et al., 2010; Chatterjee et al., 2015). Further, ME provides provable guarantees of exact or approximate recovery of the true matrix if the true data has some global structures (e.g., low rank) (Davenport & Romberg, 2016; Chen & Chi, 2018). Since images naturally have global structures (e.g., an image of a cat, has a cat as a main structure), ME is guaranteed to restore the intrinsic structures of the clean image.

Another motivation for our method comes from adversarial training, where an NN is trained with adversarial examples. Adversarial training is widely adopted to increase the robustness of neural networks. However, recent theoretical work formally argues that adversarial training requires substantially more data to achieve robustness (Schmidt et al., 2018). The natural question is then how to automatically obtain more data, with the purpose of creating samples that can help robustness. Our masking-then-reconstruction pipeline provides exactly one such automatic solutions. By using different random masks, we can create variations on each image, where all such variations maintain the image's underlying true global structures. We will see later in our results that this indeed provides significant gain in robustness.

2.2. Matrix Estimation Pipeline

Having described the intuition underlying ME-Net, we next provide a formal description of matrix estimation (ME), which constitutes the reconstruction step in our pipeline.

Matrix Estimation. Matrix estimation is concerned with recovering a data matrix from noisy and incomplete observations of its entries. Consider a true, unknown data matrix $M \in \mathbb{R}^{n\times m}$. Often, we have access to a subset Ω of entries from a noisy matrix $X \in \mathbb{R}^{n\times m}$ such that $\mathbb{E}[X] = M$. For example, in recommendation system, there are true, unknown ratings for each product from each user. One often observes a subset of noisy ratings if the user actually rates the product online. Technically, it is often assumed that each entry of X, X_{ij}, is a random variable independent of the



(a) MNIST (b) CIFAR-10

(c) SVHN (d) Tiny-ImageNet

Figure 1. **The approximate rank of different datasets.** We plot the histogram (in red) and the empirical CDF (in blue) of the approximate rank for images in each dataset.

others, which is observed with probability $p \in (0, 1]$ (i.e., missing with probability $1 - p$). The theoretical question is then formulated as finding an estimator $\hat{M}$, given noisy, incomplete observation matrix X, such that $\hat{M}$ is "close" to M. The closeness is typically measured by some matrix norm, $||\hat{M} - M||$, such as the Frobenius norm.

Over the years, extensive algorithms have been proposed. They range from simple spectral method such as universal singular value thresholding (USVT) (Chatterjee et al., 2015), which performs SVD on the observation matrix X and discards small singular values (and corresponding singular vectors), to convex optimization based methods, which minimize the nuclear norm (Candès & Recht, 2009), i.e.:

$$\min_{\hat{M} \in \mathbb{R}^{n \times m}} ||\hat{M}||_* \quad \text{s.t. } \hat{M}_{ij} \approx X_{ij}, \forall (i,j) \in \Omega, \qquad (1)$$

where $||\hat{M}||_*$ is the nuclear norm of the matrix (i.e., sum of the singular values). To speed up the computation, the Soft-Impute algorithm (Mazumder et al., 2010) reformulates the optimization using a regularization parameter $\lambda \geq 0$:

$$\min_{\hat{M} \in \mathbb{R}^{n \times m}} \frac{1}{2} \sum_{(i,j) \in \Omega} \left(\hat{M}_{ij} - X_{ij} \right)^2 + \lambda ||\hat{M}||_*. \qquad (2)$$

In this paper, we view ME as a reconstruction oracle from masked images, rather than focusing on specific algorithms.

The key message in the field of ME is: if the true data matrix M has some *global structures*, exact or approximate recovery of M can be theoretically guaranteed (Candès & Recht, 2009; Chatterjee et al., 2015; Chen & Chi, 2018). This strong theoretical guarantee serves as the foundation for employing ME to reconstruct structures in images. In

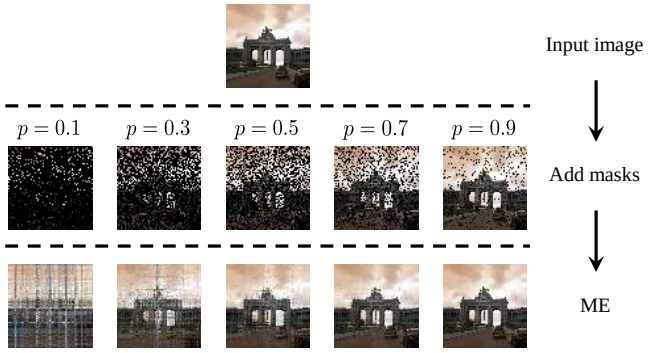


Figure 2. **An example of how ME affects the input images.** We apply different masks and show the reconstructed images by ME.

the literature, the most studied global structure is low rank. Latent variable models, where each row i and each column j are associated with some features $u_i \in \mathbb{R}^r$ and $v_j \in \mathbb{R}^r$ and $M_{ij} = f(u_i, v_j)$ for some function f, have also been investigated (Chatterjee et al., 2015; Borgs et al., 2017). To some extent, both could be good models for images.

Empirical Results. Before closing, we empirically show that images have strong global structures (i.e., low rank). We consider four datasets: MNIST, CIFAR-10, SVHN, and Tiny-ImageNet. We perform SVD on each image and compute its approximate rank, which is defined as the minimum number of singular values necessary to capture at least 90% of the energy in the image. Fig. 1 plots the histogram and the empirical CDF of the approximate ranks for each dataset. As expected, images in all datasets are relatively low rank. Specifically, the vast majority of images in MNIST, CIFAR-10, and SVHN have a rank less than 5. The rank of images in Tiny-ImageNet is larger but still significantly less than the image dimension ($\sim$10 vs. 64). This result shows that images tend to be low-rank, which implies the validity of using ME as our reconstruction oracle to find global structures.

Next, we show in Fig. 2 the results of ME-based reconstruction for different masks. Evidently, the global structure (the gate in the image) has been maintained even when p, the probability of observing the true pixel, is as low as 0.3. This shows that despite random masking we should be able to reconstruct the intrinsic global image structure from the masked adversarial images. Our intuition is that humans use such underlying global structures for image classification, and if we can maintain such global structures while weakening other potentially adversarial structures, we can force both training and testing to focus on human recognizable structures and increase robustness to adversarial attacks.

2.3. Model

We are now ready to formally describe our technique, which we refer as ME-Net. The method is illustrated in Fig. 3 and summarized as follows:

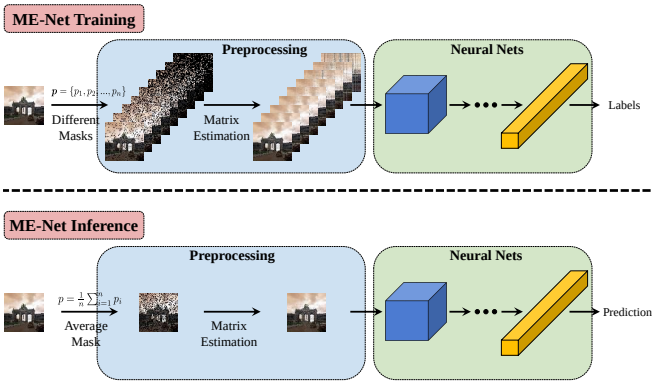


Figure 3. An illustration of ME-Net training and inference process.

- **ME-Net Training:** Define a mask as an image transform in which each pixel is preserved with probability p and set to zero with probability $1-p$. For each training image X, we apply n masks with probabilities $\{p_1, p_2, \ldots, p_n\}$, and obtain n masked images $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$. An ME algorithm is then applied to obtain reconstructed images $\{\hat{X}^{(1)}, \hat{X}^{(2)}, \ldots, \hat{X}^{(n)}\}$. We train the network on the reconstructed images $\{\hat{X}^{(1)}, \hat{X}^{(2)}, \ldots, \hat{X}^{(n)}\}$ as usual via SGD. Alternatively, adversarial training can also be readily applied in our framework.

- **ME-Net Inference:** For each test image X, we randomly sample a mask with probability $p = \frac{1}{n}\sum_{i=1}^{n} p_i$, i.e., the average of the masking probabilities during training. The masked image is then processed by the same ME algorithm used in training to obtain $\hat{X}$. Finally, $\hat{X}$ is fed to the network for prediction.

Note that we could either operate on the three RGB channels separately as independent matrices or jointly by concatenating them into one matrix. In this paper, we take the latter approach as their structures are closely related. We provide additional details of ME-Net in Appendix A and B.

3. Evaluation

We evaluate ME-Net empirically under ℓ_∞-bounded attacks and compare it with state-of-the-art defense techniques.

Experimental Setup: We implement ME-Net as described in Section 2.3. During training, for each image we randomly sample 10 masks with different p values and apply matrix estimation for each masked image to construct the training set. During testing, we sample a single mask with p set to the average of the values used during training, apply the ME-Net pipeline, and test on the reconstructed image. Unless otherwise specified, we use the Nuclear Norm minimization method (Candès & Recht, 2009) for matrix estimation.

We experiment with two versions of ME-Net: the first version uses standard stochastic gradient descent (SGD) to train the network, and the second version uses adversarial training, where the model is trained with adversarial examples.

For each attack type, we compare ME-Net with state-of-the-art defense techniques for the attack under consideration. For each technique, we report accuracy as the percentage of adversarial examples that are correctly classified.[1] As common in prior work (Madry et al., 2017; Buckman et al., 2018; Song et al., 2018), we focus on robustness against ℓ_∞-bounded attacks, and generate adversarial examples using standard methods such as the CW attack (Carlini & Wagner, 2017), Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015), and Projected Gradient Descent (PGD) which is a more powerful adversary that performs a multi-step variant of FGSM (Madry et al., 2017).

Organization: We first perform an extensive study on CIFAR-10 to validate the effectiveness of ME-Net against black-box and white-box attacks. We then extend the results to other datasets such as MNIST, SVHN, and Tiny-ImageNet. We also provide additional supporting results in Appendix C, D, E, F, G and J. Additional hyper-parameter studies, such as random restarts and different number of masks, can be found in Appendix I, H and K.

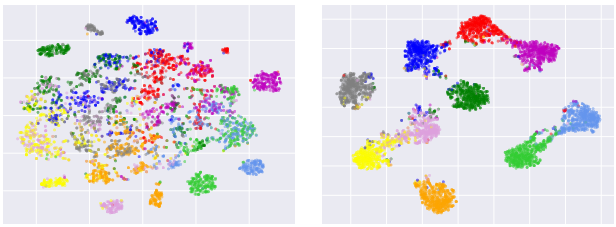
3.1. Black-box Attacks

In black-box attacks, the attacker has no access to the network model; it only observes the inputs and outputs. We evaluate ME-Net against three kinds of black-box attacks:

- **Transfer-based attack:** A copy of the victim network is trained with the same training settings. We apply CW, FGSM and PGD attacks on the copy network to generate black-box adversarial examples. We use the same attack parameters as in (Madry et al., 2017): total perturbation ε of $8/255$ (0.031), step size of $2/255$ (0.01). For PGD attacks, we use 7, 20 and 40 steps. Note that we only consider the *strongest* transfer-based attacks, i.e., we use *white-box* attacks on the independently trained copy to generate black-box examples.
- **Decision-based attack:** We apply the newly proposed Boundary attack (Brendel et al., 2017) which achieves better performance than transfer-based attacks. We apply 1000 attack steps to ensure convergence.
- **Score-based attack:** We also apply the state-of-the-art SPSA attack (Uesato et al., 2018) which is strong enough to bring the accuracy of several defenses to near zero. We use a batch-size of 2048 to make the SPSA strong, and leave other hyper-parameters unchanged.

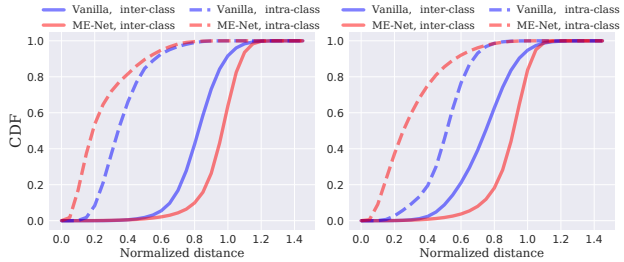
As in past work that evaluates robustness on CIFAR-10

[1] To be consistent with literature, we generate adversarial examples from the whole dataset and use all of them to report accuracy.



(a) Vanilla under adv. attack. (b) ME-Net under adv. attack.

Figure 4. **Class separation under black-box adversarial attack.** The vectors right before the softmax layer are projected to a 2D plane using t-SNE (Maaten & Hinton, 2008).



(a) Clean data. (b) Black-box adv. attack.

Figure 5. **The empirical CDF of the distance within and among classes.** We quantitatively show the intra-class and inter-class distances between vanilla model and ME-Net on clean data and under black-box adversarial attacks.

(Madry et al., 2017; Buckman et al., 2018), we use the standard ResNet-18 model in (He et al., 2016). In training ME-Net, we experiment with different settings for p. We report the results for $p \in [0.8, 1]$ below, and refer the reader to the Appendix for the results with other p values.

Since most defenses experimented only with transfer-based attacks, we first compare ME-Net to past defenses under transfer-based attacks. For comparison, we select a state-of-the-art adversarial training defense (Madry et al., 2017) and a preprocessing method (Buckman et al., 2018). We compare these schemes against ME-Net with standard SGD training. The results are shown in Table 1. They reveal that even without adversarial training, ME-Net is much more robust than prior work to black-box attacks, and can improve accuracy by 13% to 25%, depending on the attack.

To gain additional insight, we look at the separation between different classes under black-box transfer-based attack, for the vanilla network and ME-Net. Fig. 4(a) and 4(b) show the 2D projection of the vectors right before the output layer (i.e., softmax layer), for the test data in the vanilla model and ME-Net. The figures show that when the vanilla model is under attack, it loses its ability to separate different classes. In contrast, ME-Net can sustain clear separation between classes even in the presence of black-box attack.

To further understand this point, we compute the Euclidean distance between classes and within each class. Fig. 5 plots

Method	Training	CW	FGSM	PGD (7 steps)
Vanilla	SGD	8.9%	24.8%	7.6%
Madry	Adv. train	78.7%	67.0%	64.2%
Thermometer	SGD	–	–	53.5%
Thermometer	Adv. train	–	–	77.7%
ME-Net	SGD	**93.6%**	**92.2%**	**91.8%**

Table 1. **CIFAR-10 black-box results under transfer-based attacks.** We compare ME-Net with state-of-the-art defense methods under both SGD and adversarial training.

Attacks	CW	FGSM	PGD			Boundary	SPSA
			7 steps	20 steps	40 steps		
Vanilla	8.9%	24.8%	7.6%	1.8%	0.0%	3.5%	1.4%
ME-Net	**93.6%**	**92.2%**	**91.8%**	**91.8%**	**91.3%**	**87.4%**	**93.0%**

Table 2. **CIFAR-10 extensive black-box results.** We show significant adversarial robustness of ME-Net under different strong black-box attacks.

the empirical CDFs of the intra-class and inter-class distance between the vectors before the output layer, for both the vanilla classifier and ME-Net. The figure shows results for both clean data and adversarial examples. Comparing ME-Net (in red) with the vanilla classifier (in blue), we see that ME-Net both reduces the distance within each class, and improves the separation between classes; further this result applies to both clean and adversarial examples. Overall, these visualizations offer strong evidence supporting the improved robustness of ME-Net.

Finally, we also evaluate ME-Net under other strong black-box attacks. Table 2 summarizes these results demonstrating that ME-Net consistently achieves high robustness under different black-box attacks.

3.2. White-box Attacks

In white-box attacks, the attacker has full information about the neural network model (architecture and weights) and defense methods. To evaluate robustness against such white-box attacks, we use the BPDA attack proposed in (Athalye et al., 2018), which has successfully circumvented a number of previously effective defenses, bringing them to near 0 accuracy. Specifically, most defense techniques rely on preprocessing methods which can cause *gradient masking* for gradient-based attacks, either because the preprocessing is not differentiable or the gradient is useless. BPDA addresses this issue by using a "differentiable approximation" for the backward pass. As such, until now no preprocessing method is effective under white-box attacks. In ME-Net, the backward pass is not differentiable, which makes BPDA the strongest white-box attack. We use PGD-based BPDA and experiment with different number of attack steps.

Method	Type	Steps	Accuracy
Thermometer	Prep.	40	0.0%*
PixelDefend	Prep.	100	9.0%*
TV Minimization	Prep.	100	0.4%
ME-Net	Prep.	1000	**40.8%**

Table 3. **White-box attack against pure preprocessing schemes.** We use PGD or BPDA attacks in white-box setting. Compared to other pure preprocessing methods, ME-Net can increase robustness by a significant margin. *Data from (Athalye et al., 2018).

Network	Method	Type	Steps	Accuracy
ResNet-18	Madry	Adv. train	1000	45.0%
	ME-Net	Prep. + Adv. train	1000	**52.8%**
WideResNet	Madry	Adv. train	1000	46.8%
	Thermometer	Prep. + Adv. train	1000	12.3%
	ME-Net	Prep. + Adv. train	1000	**55.1%**

Table 4. **White-box attack results for adversarial training**. We use 1000 steps PGD or BPDA attacks in white-box setting to ensure the results are convergent. ME-Net achieves state-of-the-art white-box robustness when combined with adversarial training.

For white box attacks, we distinguish two cases: defenses that use only preprocessing (without adversarial training), and defenses that incorporate adversarial training. All defenses that incorporate adversarial training, including ME-Net, are trained with PGD with 7 steps.

Table 3 shows a comparison of the performance of various preprocessing methods against the BPDA white-box attack. We compare ME-Net with three preprocessing defenses, i.e., the PixelDefend method (Song et al., 2018), the Thermometer method (Buckman et al., 2018), and the total variation (TV) minimization method (Guo et al., 2017). The results in the table for (Song et al., 2018; Buckman et al., 2018) are directly taken from (Athalye et al., 2018). Since the TV minimization method is not tested on CIFAR-10, we implement this method using the same setting used with ME-Net. The table shows that preprocessing alone is vulnerable to the BPDA white-box attack, as all schemes perform poorly under such attack. Interestingly however, the table also shows that ME-Net's preprocessing is significantly more robust to BPDA than other preprocessing methods. We attribute this difference to that ME-Net's preprocessing step focuses on protecting the global structures in images.

Next we report the results of white-box attacks on schemes that use adversarial training. One key characteristic of ME-Net is its orthogonality with adversarial training. Note that many preprocessing methods propose combining adversarial training, but the combination actually performs worse than adversarial training alone (Athalye et al., 2018). Since ME-Net's preprocessing already has a decent accuracy under the strong white-box attacks, we envision a further improvement when combining with adversarial training. We compare ME-Net against two baselines: we compare against (Madry et al., 2017), which is the state-of-the-art in defenses against white-box attacks. We also compare with the Thermometer technique in (Buckman et al., 2018), which like ME-Net, combines a preprocessing step with adversarial training. For all compared defenses, adversarial training is done using PGD with 7 steps. We also use BPDA to approximate the gradients during the backward pass. For our comparison we use ResNet-18 and its wide version since they were used in past work on robustness with adversarial training. As for the attacker, we allow it to use the *strongest possible* attack, i.e., it uses BPDA with 1000 PGD attack steps to ensure the results are convergent. Note that previous defenses (including the state-of-the-art) only consider up to 40 steps.

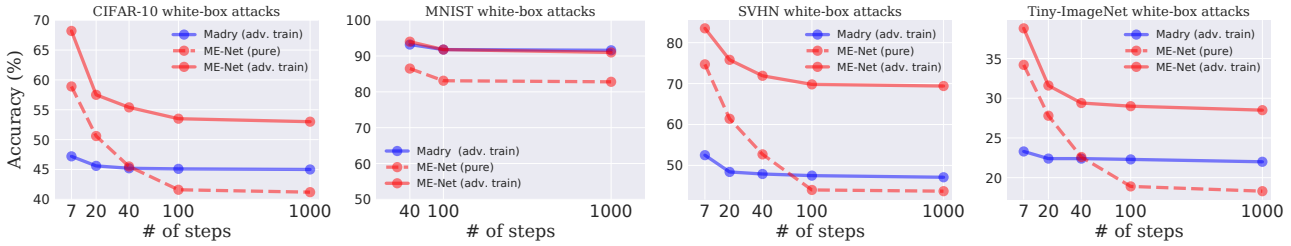
Table 4 summarizes the results. As shown in the table, ME-Net combined with adversarial training outperforms the state-of-the-art results under white-box attacks, achieving a 52.8% accuracy with ResNet and a 55.1% accuracy with WideResNet. In contrast, the Thermometer method that also uses preprocessing plus adversarial training cannot survive the strong white-box adversary.

3.3. Evaluation with Different Datasets

We evaluate ME-Net on MNIST, SVHN, CIFAR-10, and Tiny-ImageNet and compare its performance across these datasets. For space limitations, we present only the results for the white-box attacks. We provide results for black-box attacks and additional attacks in Appendix C, D, E, and F.

For each dataset, we use the network architecture and parameters commonly used in past work on adversarial robustness to help in comparing our results to past work. For MNIST, we use the LeNet model with two convolutional layers as in (Madry et al., 2017). We also use the same attack parameters as total perturbation scale of $76.5/255$ (0.3), and step size $2.55/255$ (0.01). Besides using 40 and 100 total attack steps, we also increase to 1000 steps to further strengthen the adversary. For ME-Net with adversarial training, we follow their settings to use 40 steps PGD during training. We use standard ResNet-18 for SVHN and CIFAR-10, and DenseNet-121 for Tiny-ImageNet, and set attack parameters as follows: total perturbation of $8/255$ (0.031), step size of $2/255$ (0.01), and with up to 1000 total attack steps. Since in (Madry et al., 2017) the authors did not examine on SVHN and Tiny-ImageNet, we follow their methods to retrain their model on these datasets. We use 7 steps PGD for adversarial training. We keep all the training hyper-parameters the same for ME-Net and (Madry et al., 2017).

Fig. 6 shows the performance of ME-Net on the four datasets and compares it with (Madry et al., 2017), a state-of-the-art



(a) CIFAR-10 (b) MNIST (c) SVHN (d) Tiny-ImageNet

Figure 6. **White-box attack results on different datasets**. We compare ME-Net with (Madry et al., 2017) under PGD or BPDA attack with different attack steps up to 1000. We show both the pure ME-Net without adversarial training, and ME-Net with adversarial training. For Tiny-ImageNet, we report the Top-1 adversarial robustness.

defense against white-box attacks. We plot both the result of a pure version of ME-Net, and ME-Net with adversarial training. The figure reveals the following results. First, it shows that ME-Net with adversarial training outperforms the state-of-the-art defense against white-box attacks. Interestingly however, the gains differ from one dataset to another. Specifically, ME-Net is comparable to (Madry et al., 2017) on MNIST, provides about 8% gain on CIFAR-10 and Tiny-ImageNet, and yields 23% gain on SVHN.

We attribute the differences in accuracy gains across datasets to differences in their properties. MNIST is too simple (single channel with small 28×28 pixels), and hence ME-Net and (Madry et al., 2017) both achieve over 90% accuracy. The other datasets are all more complex and have 3 RGB channels and bigger images. More importantly, Fig. 1 shows that the vast majority of images in SVHN have a very low rank, and hence very strong global structure, which is a property that ME-Net leverages to yield an accuracy gain of 23%. CIFAR-10 and Tiny-ImageNet both have relatively low rank images but not as low as SVHN. The CDF shows that 90% of the images in CIFAR have a rank lower than 5, whereas 90% of the images in Tiny-ImageNet have a rank below 10. When taking into account that the dimension of Tiny-ImageNet is twice as CIFAR (64×64 vs. 32×32), one would expect ME-Net's gain on these datasets to be comparable, which is compatible with the empirical results.

3.4. Evaluation against Adaptive Attacks

Since ME-Net provides a new preprocessing method, we examine customized attacks where the adversary takes advantage of knowing the details of ME-Net's pipeline. We propose two kinds of white-box attacks: 1) *Approximate input attack*: since ME-Net would preprocess the image, this adversary attacks not the original image, but uses the exact preprocess method to approximate/reconstruct an input, and attacks the newly constructed image using the BPDA procedure (Athalye et al., 2018). 2) *Projected BPDA attack*: since ME-Net focuses on the global structure of an image, this adversary aims to attack directly the main structural space of the image. Specifically, it uses BPDA to approximate the gradient, and then projects the gradient to the low-rank space of the image iteratively, i.e., it projects on the space constructed by the top few singular vectors of the original image, to construct the adversarial noise. Note that these two attacks are based on the BPDA white-box attack which has shown most effective against preprocessing. Table 5 shows the results of these attacks, which demonstrates that ME-Net is robust to these adaptive white-box attacks.

Method	Training	Steps	Approx. Input	Projected BPDA
ME-Net	Pure	1000	**41.5%**	**64.9%**
	Adversarial	1000	**62.5%**	**74.7%**

Table 5. **Results of ME-Net against adaptive white-box attacks on CIFAR-10.** We use 1000 steps PGD-based BPDA for the two newly proposed attacks, and report the accuracy of ME-Net.

3.5. Comparison of Different ME Methods

Matrix estimation (ME) is a well studied topic with several established ME techniques. The results in the other sections are with the Nuclear Norm minimization algorithm (Candès & Recht, 2009). Here we compare the performance of three ME methods: the Nuclear Norm minimization algorithm, the Soft-Impute algorithm (Mazumder et al., 2010), and the universal singular value thresholding (USVT) approach (Chatterjee et al., 2015).

We train ME-Net models using different ME methods on CIFAR-10 with ResNet-18. We apply transfer-based PGD black-box attacks with 40 attack steps, as well as white-box BPDA attack with 1000 attack steps. We compare the complexity, generalization and adversarial robustness of these methods. More details can be found in Appendix H.

Table 6 shows the results of our comparison. The table shows that all the three ME methods are able to improve the original standard generalization, and achieve almost the same test accuracy. The nuclear norm minimization algorithm takes much longer time and more computation power. The Soft-Impute algorithm simplifies the process but still requires certain computation resources, while the USVT approach is much simpler and faster. The performance of

Method	Complexity	Clean	Black-box	White-box
Vanilla	–	93.4%	0.0%	0.0%
ME-Net - USVT	Low	**94.8%**	**89.4%**	**51.9%**
ME-Net - Soft-Imp.	Medium	**94.9%**	**91.3%**	**52.3%**
ME-Net - Nuc. Norm	High	**94.8%**	**91.0%**	**52.8%**

Table 6. **Comparisons between different ME methods.** We report the generalization and adversarial robustness of three ME-Net models using different ME methods on CIFAR-10. We apply transfer-based 40 steps PGD attack as black-box adversary, and 1000 steps PGD-based BPDA as white-box adversary.

Method	Training	MNIST	CIFAR-10	SVHN	Tiny-ImageNet
Vanilla	Pure	98.8%	93.4%	95.0%	66.4%
ME-Net	Pure	**99.2%**	**94.9%**	**96.0%**	**67.7%**
Madry	Adversarial	98.5%	79.4%	87.4%	45.6%
ME-Net	Adversarial	**98.8%**	**85.5%**	**93.5%**	**57.0%**

Table 7. **Generalization performance on clean data.** For each dataset, we use the same network for all the schemes. ME-Net improves generalization for both adversarial and non-adversarial training. For Tiny-ImageNet, we report the Top-1 accuracy.

different ME methods is slightly different, as more complex algorithms may gain better performances.

3.6. Improving Generalization

As a preprocessing method, ME-Net also serves as a data augmentation technique during training. We show that besides adversarial robustness, ME-Net can also improve generalization (i.e., the test accuracy) on clean data. We distinguish between two training procedures: 1) non-adversarial training, where the model is trained only with clean data, and 2) adversarial training where the model is trained with adversarial examples. For each case we compare ME-Net with the best performing model for that training type. We show results for different datasets, where each dataset is trained with the typical model in past work as stated in Section 3.3. Table 7 shows the results, which demonstrate the benefit of ME-Net as a method for improving generalization under both adversarial and non-adversarial training.

4. Related Work

Due to the large body of work on adversarial robustness, we focus on methods that are most directly related to our work, and refer readers to the survey (Akhtar & Mian, 2018) for a more comprehensive and broad literature review.

Adversarial Training. Currently, the most effective way to defend against adversarial attacks is adversarial training, which trains the model on adversarial examples generated by different kinds of attacks (Madry et al., 2017; Szegedy et al., 2013; Goodfellow et al., 2015). Authors of (Madry et al., 2017) showed that training on adversarial examples generated by PGD with a random start can achieve state-of-the-art performance on MNIST and CIFAR-10 under ℓ_∞ constraint. One major difficulty of adversarial training is that it tends to overfit to the adversarial examples. Authors in (Schmidt et al., 2018) thus demonstrated and proved that much more data is needed to achieve good generalization under adversarial training. ME-Net can leverage adversarial training for increased robustness. Further its data augmentation capability helps improving generalization.

Preprocessing. Many defenses preprocess the images with a transformation prior to classification. Typical preprocessing includes image re-scaling (Xie et al., 2018), discretization (Chen et al., 2018), thermometer encoding (Buckman et al., 2018), feature squeezing (Xu et al., 2017), image quilting (Guo et al., 2017), and neural-based transformations (Song et al., 2018; Samangouei et al., 2018). These defenses can cause *gradient masking* when using gradient-based attacks. However, as shown in (Athalye et al., 2018), by applying the Backward Pass Differentiable Approximation (BPDA) attacks designed for obfuscated gradients, the accuracy of all of these methods can be brought to near zero. ME-Net is the first preprocessing method that remains effective under the strongest BPDA attack, which could be attributed to its ability to leverage adversarial training.

Matrix Estimation. Matrix estimation recovers a data matrix from noisy and incomplete samples of its entries. A classical application is recommendation systems, such as the Netflix problem (Bell & Koren, 2007), but it also has richer connections to other learning challenges such as graphon estimation (Airoldi et al., 2013; Borgs et al., 2017), community detection (Abbe & Sandon, 2015b;a) and time series analysis (Agarwal et al., 2018). Many efficient algorithms exist such as the universal singular value thresholding approach (Chatterjee et al., 2015), the convex nuclear norm minimization formulation (Candès & Recht, 2009) and even non-convex methods (Jain et al., 2013; Chen & Wainwright, 2015; Ge et al., 2016). The key promise is that as long as there are some structures underlying the data matrix, such as being low-rank, then exact or approximate recovery can be guaranteed. As such, ME is an ideal reconstruction scheme for recovering global structures.

5. Conclusion

We introduced ME-Net, which leverages matrix estimation to improve the robustness to adversarial attacks. Extensive experiments under strong black-box and white-box attacks demonstrated the significance of ME-Net, where it consistently improves the state-of-the-art robustness in different benchmark datasets. Furthermore, ME-Net can easily be embedded into existing networks, and can also bring additional benefits such as improving standard generalization.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments in revising the paper. We are grateful to the members of NETMIT and CSAIL for their insightful discussions and supports. Zhi Xu is supported by the Siemens FutureMakers Fellowship.

References

Abbe, E. and Sandon, C. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pp. 670–688. IEEE, 2015a.

Abbe, E. and Sandon, C. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in neural information processing systems*, pp. 676–684, 2015b.

Agarwal, A., Amjad, M. J., Shah, D., and Shen, D. Model agnostic time series analysis via matrix estimation. *ACM SIGMETRICS performance evaluation review*, 2(3), 2018.

Airoldi, E. M., Costa, T. B., and Chan, S. H. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pp. 692–700, 2013.

Akhtar, N. and Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *arXiv preprint arXiv:1801.00553*, 2018.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, July 2018. URL https://arxiv.org/abs/1802.00420.

Bell, R. M. and Koren, Y. Lessons from the netflix prize challenge. *SIGKDD Explor. Newsl.*, 9(2):75–79, December 2007. ISSN 1931-0145. doi: 10.1145/1345448.1345465. URL http://doi.acm.org/10.1145/1345448.1345465.

Borgs, C., Chayes, J., Lee, C. E., and Shah, D. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. In *Advances in Neural Information Processing Systems*, pp. 4715–4726, 2017.

Brendel, W., Rauber, J., and Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. Thermometer encoding: One hot way to resist adversarial examples. 2018. URL https://openreview.net/pdf?id=S18Su--CW.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Chatterjee, S. et al. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Chen, J., Wu, X., Liang, Y., and Jha, S. Improving adversarial robustness by data-specific discretization. *CoRR*, abs/1805.07816, 2018.

Chen, Y. and Chi, Y. Harnessing structures in big data via guaranteed low-rank matrix estimation. *arXiv preprint arXiv:1802.08397*, 2018.

Chen, Y. and Wainwright, M. J. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Davenport, M. A. and Romberg, J. An overview of low-rank matrix recovery from incomplete observations. *arXiv preprint arXiv:1601.06422*, 2016.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Guo, C., Rana, M., Cisse, M., and van der Maaten, L. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM, 2013.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*, 2018.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug): 2287–2322, 2010.

Mosbach, M., Andriushchenko, M., Trost, T., Hein, M., and Klakow, D. Logit pairing methods can fool gradient-based attacks. 2018.

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *NIPS*, 2018. URL http://arxiv.org/abs/1804.11285.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Uesato, J., O'Donoghue, B., Oord, A. v. d., and Kohli, P. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Supplementary Material

A. Pseudo Code for ME-Net

Algorithm 1 ME-Net training & inference

```
/* ME-Net Training */
```
Input: training set $S = \{(X_i, y_i)\}_{i=1}^{M}$, prescribed masking probability $\boldsymbol{p} = \{p_1, p_2, \ldots, p_n\}$, network N
for all $X_i \in S$ **do**
  Randomly sample n masks with probability $\{p_1, p_2, \ldots, p_n\}$
  Generate n masked images $\{X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(n)}\}$
  Apply ME to obtain reconstructed images $\{\hat{X}_i^{(1)}, \hat{X}_i^{(2)}, \ldots, \hat{X}_i^{(n)}\}$
  Add $\{\hat{X}_i^{(1)}, \hat{X}_i^{(2)}, \ldots, \hat{X}_i^{(n)}\}$ into new training set S'
end for
Randomly initialize network N
for number of training iterations **do**
  Sample a mini-batch $B = \{(\hat{X}_i, y_i)\}_{i=1}^{m}$ from S'
  Do one training step of network N using mini-batch B
end for

```
/* ME-Net Inference */
```
Input: test image X, masking probability $\boldsymbol{p} = \{p_1, p_2, \ldots, p_n\}$ used during training
Output: predicted label y
Randomly sample one mask with probability $p = \frac{1}{n}\sum_{i=1}^{n} p_i$
Generate masked image and apply ME to reconstruct $\hat{X}$
Input $\hat{X}$ to the trained network N to get the predicted label y

B. Training Details

Training settings. We summarize our training hyper-parameters in Table 8. We follow the standard data augmentation scheme as in (He et al., 2016) to do zero-padding with 4 pixels on each side, and then random crop back to the original image size. We then randomly flip the images horizontally and normalize them into $[0, 1]$. Note that ME-Net's preprocessing is performed before the training process as in Algorithm 1.

Dataset	Model	Data Aug.	Optimizer	Momentum	Epochs	LR	LR decay
CIFAR-10	ResNet-18 Wide-ResNet	$\surd$	SGD	0.9	200	0.1	step (100, 150)
MNIST	LeNet	$\times$	SGD	0.9	200	0.01	step (100, 150)
SVHN	ResNet-18	$\surd$	SGD	0.9	200	0.01	step (100, 150)
Tiny-ImageNet	DenseNet-121	$\surd$	SGD	0.9	90	0.1	step (30, 60)

Table 8. **Training details of ME-Net on different datasets.** Learning rate is decreased at selected epochs with a step factor of 0.1.

ME-Net details. As was mentioned in Section 2.3, one could either operate on the three RGB channels separately as independent matrices or jointly by concatenating them into one wide matrix. For the former approach, given an image, we can apply the same mask to each channel and then separately run ME to recover the matrix. For the latter approach, the RGB channels are first concatenated along the column dimension to produce a wide matrix, i.e., if each channel is of size 32×32, then the concatenated matrix, [RGB], is of size 32×96. A mask is applied to the wide matrix and the whole matrix is then recovered. This approach is a common, simple method for estimating tensor data. Since this work focuses on structures of the image and channels within an image are closely related, we adopt the latter approach in this paper.

In our experiments, we use the following method to generate masks with different observing probability: for each image, we

select n masks in total with observing probability p ranging from $a \to b$. We use $n = 10$ for most experiments. To provide an example, "$p : 0.6 \to 0.8$" indicates that we select 10 masks in total with observing probability from 0.6 to 0.8 with an equal interval of 0.02, i.e., $0.6, 0.62, 0.64, \ldots$. Note that we only use this simple selection scheme for mask generation. We believe further improvement can be achieved with better designed selection schemes, potentially tailored to each image.

C. Additional Results on CIFAR-10

C.1. Black-box Attacks

We provide additional results of ME-Net against different black-box attacks on CIFAR-10. We first show the complete results using different kinds of black-box attacks, i.e., transfer-based (FGSM, PGD, CW), decision-based (Boundary) and score-based (SPSA) attacks. For CW attack, we follow the settings in (Madry et al., 2017) to use different confidence values κ. We report ME-Net results with different training settings on Table 9. Here we use pure ME-Net as a preprocessing method without adversarial training. As shown, previous defenses only consider limited kinds of black-box attacks. We by contrast show extensive and also advanced experimental results.

Method		Clean	FGSM	PGD			CW		Boundary	SPSA
				7 steps	20 steps	40 steps	$\kappa = 20$	$\kappa = 50$		
Vanilla		93.4%	24.8%	7.6%	1.8%	0.0%	9.3%	8.9%	3.5%	1.4%
Madry		79.4%	67.0%	64.2%	–	–	78.7%	–	–	–
Thermometer		87.5%	–	77.7%	–	–	–	–	–	–
ME-Net	$p : 0.8 \to 1$	**94.9%**	**92.2%**	**91.8%**	**91.8%**	**91.3%**	**93.6%**	**93.6%**	**87.4%**	**93.0%**
	$p : 0.6 \to 0.8$	92.1%	85.1%	84.5%	83.4%	81.8%	89.2%	89.0%	81.8%	90.9%
	$p : 0.4 \to 0.6$	89.2%	75.7%	74.9%	73.0%	70.9%	82.0%	82.0%	77.5%	87.1%

Table 9. **CIFAR-10 extensive black-box attack results.** Different kinds of strong black-box attacks are used, including transfer-, decision-, and score-based attacks.

Further, we define and apply another stronger black-box attack, where we provide the architecture and weights of our trained model to the black-box adversary to make it stronger. This kind of attack is also referred as "semi-black-box" or "gray-box" attack in some instances, while we still view it as a black-box one. This time the adversary is not aware of the preprocessing layer but has full access to the trained network, and directly performs white-box attacks to the network. We show the results in Table 10.

Method		FGSM	PGD			CW	
			7 steps	20 steps	40 steps	$\kappa = 20$	$\kappa = 50$
ME-Net	$p : 0.8 \to 1$	**85.1%**	**84.9%**	**84.0%**	**82.9%**	75.8%	75.2%
	$p : 0.6 \to 0.8$	83.2%	82.8%	81.7%	79.6%	81.5%	76.8%
	$p : 0.4 \to 0.6$	80.5%	80.2%	79.2%	76.4%	**84.0%**	**77.1%**

Table 10. **CIFAR-10 additional black-box attack results where adversary has limited access to the trained network.** We provide the architecture and weights of our trained model to the black-box adversary to make it stronger.

C.2. White-box Attacks

C.2.1. Pure ME-Net

We first show the extensive white-box attack results with pure ME-Net in Table 11. We use strongest white-box BPDA attack (Athalye et al., 2018) with different attack steps. We select three preprocessing methods (Song et al., 2018; Buckman et al., 2018; Guo et al., 2017) as competitors. We re-implement the total variation minimization approach (Guo et al., 2017) and apply the same training settings as ME-Net on CIFAR-10. The experiments are performed under total perturbation

ε of $8/255$ (0.031). By comparison, ME-Net is demonstrated to be the first preprocessing method that is effective under strongest white-box attacks.

Method		Type	Attack Steps			
			7	**20**	**40**	**100**
Vanilla		–	0.0%	0.0%	0.0%	0.0%
Thermometer		Prep.	–	–	0.0%*	0.0%*
PixelDefend		Prep.	–	–	–	9.0%*
TV Minimization		Prep.	14.7%	5.1%	2.7%	0.4%
ME-Net	$p: 0.8 \rightarrow 1$	Prep.	46.2%	33.2%	26.8%	23.5%
	$p: 0.7 \rightarrow 0.9$	Prep.	50.3%	40.4%	33.7%	29.5%
	$p: 0.6 \rightarrow 0.8$	Prep.	53.0%	45.6%	37.8%	35.1%
	$p: 0.5 \rightarrow 0.7$	Prep.	55.7%	47.3%	38.6%	35.9%
	$p: 0.4 \rightarrow 0.6$	Prep.	**59.8%**	**52.6%**	**45.5%**	**41.6%**

Table 11. **CIFAR-10 extensive white-box attack results with pure ME-Net.** We use the strongest PGD or BPDA attacks in white-box setting with different attack steps. We compare ME-Net with other pure preprocessing methods (Buckman et al., 2018; Song et al., 2018; Guo et al., 2017). We show that ME-Net is the first preprocessing method to be effective under white-box attacks. *Data from (Athalye et al., 2018).

Further, we study the performance of ME-Net under different ε in Fig. 7. Besides using $\varepsilon = 8$ which is commonly used in CIFAR-10 attack settings (Madry et al., 2017), we additionally provide more results including $\varepsilon = 2$ and 4 to study the performance of pure ME-Net under strongest BPDA white-box attacks.

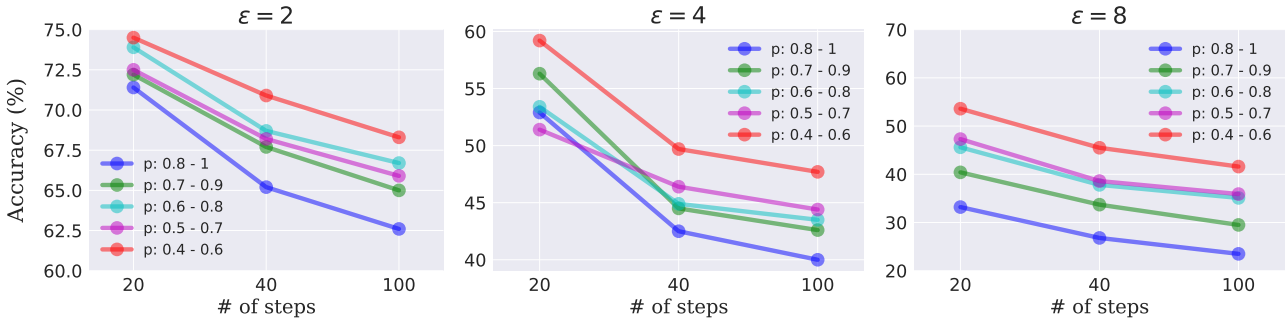


Figure 7. **CIFAR-10 white-box attack results of pure ME-Net with different perturbation** ε. We report ME-Net results with different training settings under various attack steps.

Besides the strongest BPDA attack, we also design and apply another white-box attack to further study the effect of the preprocessing layer. We assume the adversary is aware of the preprocessing layer, but not use the backward gradient approximation. Instead, it performs iterative attacks only for the network part after the preprocessing layer. This attack helps study how the preprocessing affects the network robustness against white-box adversary. The results in Table 12 shows that pure ME-Net provides sufficient robustness if the white-box adversary does not attack the preprocessing layer.

C.2.2. Combining with Adversarial Training

We provide more advanced and extensive results of ME-Net when combining with adversarial training in Table 13. As shown, preprocessing methods are not necessarily compatible with adversarial training, as they can perform worse than adversarial training alone (Buckman et al., 2018). Compared to current state-of-the-art (Madry et al., 2017), ME-Net achieves consistently better results under strongest white-box attacks.

Method		FGSM	PGD			CW	
			7 steps	20 steps	40 steps	$\kappa = 20$	$\kappa = 50$
ME-Net	$p : 0.8 \rightarrow 1$	**84.3%**	**83.7%**	**83.1%**	**82.5%**	77.0%	75.9%
	$p : 0.6 \rightarrow 0.8$	82.6%	82.1%	81.5%	80.3%	76.9%	76.4%
	$p : 0.4 \rightarrow 0.6$	79.1%	79.0%	78.3%	77.4%	**77.5%**	**77.2%**

Table 12. **CIFAR-10 additional white-box attack results where the white-box adversary does not attack the preprocessing layer.** We remain the same attack setups as in the white-box BPDA attack, while only attacking the network part after the preprocessing layer of ME-Net.

Network	Method	Type	Clean	Attack Steps				
				7	20	40	100	1000
ResNet-18	Madry	Adv. train	79.4%	47.2%	45.6%	45.2%	45.1%	45.0%
	ME-Net $p : 0.8 \rightarrow 1$	Prep. + Adv. train	**85.5%**	57.4%	51.5%	49.3%	48.1%	47.4%
	ME-Net $p : 0.6 \rightarrow 0.8$	Prep. + Adv. train	84.8%	62.1%	53.0%	51.2%	50.0%	49.6%
	ME-Net $p : 0.4 \rightarrow 0.6$	Prep. + Adv. train	84.0%	**68.2%**	**57.5%**	**55.4%**	**53.5%**	**52.8%**
Wide-ResNet	Madry	Adv. train	87.3%	50.0%	47.1%	47.0%	46.9%	46.8%
	Thermometer	Prep. + Adv. train	89.9%	59.4%	34.9%	26.0%	18.4%	12.3%
	ME-Net $p : 0.6 \rightarrow 0.8$	Prep. + Adv. train	**91.0%**	69.7%	58.0%	54.9%	53.4%	52.9%
	ME-Net $p : 0.4 \rightarrow 0.6$	Prep. + Adv. train	88.7%	**74.1%**	**61.6%**	**57.4%**	**55.9%**	**55.1%**

Table 13. **CIFAR-10 extensive white-box attack results.** We apply up to 1000 steps PGD or BPDA attacks in white-box setting to ensure the results are convergent. We use the released models in (Madry et al., 2017; Athalye et al., 2018) but change the attack steps up to 1000 for comparison. ME-Net shows significant advanced results by consistently outperforming the current state-of-the-art defense method (Madry et al., 2017).

D. Additional Results on MNIST

D.1. Black-box Attacks

In Table 14, we report extensive results of ME-Net under different strong black-box attacks on MNIST. We follow (Madry et al., 2017) to use the same LeNet model and the same attack parameters with a total perturbation scale of $76.5/255$ (0.3). We use a step size of $2.55/255$ (0.01) for PGD attacks. We use the same settings as in CIFAR-10 for Boundary and SPSA attacks (i.e., 1000 steps for Boundary attack, and a batch size of 2048 for SPSA attack) to make them stronger. Note that we only use the *strongest* transfer-based attacks, i.e., we use *white-box* attacks on the independently trained copy to generate black-box examples. As shown, ME-Net shows significantly more effective results against different strongest black-box attacks.

We further provide the architecture and weights of our trained model to the black-box adversary to make it stronger, and provide the results in Table 15. As shown, ME-Net can still maintain high adversarial robustness against stronger black-box adversary under this setting.

D.2. White-box Attacks

Table 16 shows the extensive white-box attack results on MNIST. As discussed, we follow (Madry et al., 2017) to use 40 steps PGD during training when combining ME-Net with adversarial training. We apply up to 1000 steps strong BPDA-based PGD attack to ensure the results are convergent. For the competitor, we use the released model in (Madry et al., 2017), but change the attack steps to 1000 for comparison.

Method		Clean	FGSM	PGD		CW		Boundary	SPSA
				40 steps	100 steps	$\kappa = 20$	$\kappa = 50$		
Vanilla		98.8%	28.2%	0.1%	0.0%	14.1%	12.6%	3.7%	6.2%
Madry		98.5%	**96.8%**	**96.0%**	**95.7%**	96.4%	97.0%	–	–
Thermometer		99.0%	–	41.1%	–	–	–	–	–
ME-Net	$p : 0.8 \to 1$	**99.2%**	77.4%	73.9%	73.6%	**98.8%**	**98.7%**	**89.3%**	**98.1%**
	$p : 0.6 \to 0.8$	99.0%	87.1%	85.1%	84.9%	98.6%	98.4%	88.6%	97.5%
	$p : 0.4 \to 0.6$	98.4%	91.1%	90.7%	88.9%	98.4%	98.3%	88.0%	97.0%
	$p : 0.2 \to 0.4$	96.8%	**93.2%**	**92.8%**	**92.2%**	96.6%	96.5%	88.1%	96.1%

Table 14. **MNIST extensive black-box attack results.** Different kinds of strong black-box attacks are used, including transfer-, decision-, and score-based attacks.

Method		FGSM	PGD		CW	
			40 steps	100 steps	$\kappa = 20$	$\kappa = 50$
ME-Net	$p : 0.8 \to 1$	93.0%	91.9%	85.5%	**98.8%**	**98.7%**
	$p : 0.6 \to 0.8$	95.0%	94.2%	93.7%	98.3%	98.2%
	$p : 0.4 \to 0.6$	**96.2%**	**95.9%**	**95.3%**	98.3%	98.0%
	$p : 0.2 \to 0.4$	94.5%	94.2%	93.4%	96.5%	96.5%

Table 15. **MNIST additional black-box attack results where adversary has limited access to the trained network.** We provide the architecture and weights of our trained model to the black-box adversary to make it stronger.

Method		Type	Clean	Attack Steps		
				40	100	1000
Madry		Adv. train	98.5%	93.2%	91.8%	**91.6%**
ME-Net	$p : 0.8 \to 1$	Prep.	**99.2%**	22.9%	21.8%	18.9%
	$p : 0.6 \to 0.8$	Prep.	99.0%	47.6%	42.4%	40.8%
	$p : 0.4 \to 0.6$	Prep.	98.4%	65.2%	62.1%	60.6%
	$p : 0.2 \to 0.4$	Prep.	96.8%	**86.5%**	**83.1%**	**82.6%**
ME-Net	$p : 0.8 \to 1$	Prep. + Adv. train	97.6%	87.8%	81.7%	78.0%
	$p : 0.6 \to 0.8$	Prep. + Adv. train	97.7%	90.5%	88.1%	86.5%
	$p : 0.4 \to 0.6$	Prep. + Adv. train	**98.8%**	92.1%	89.4%	88.2%
	$p : 0.2 \to 0.4$	Prep. + Adv. train	97.4%	**94.0%**	**91.8%**	**91.0%**

Table 16. **MNIST extensive white-box attack results.** We apply up to 1000 steps PGD or BPDA attacks in white-box setting to ensure the results are convergent. We use the released models in (Madry et al., 2017) but change the attack steps up to 1000 for comparison. We show both pure ME-Net results and the results when combining with adversarial training.

E. Additional Results on SVHN

E.1. Black-box Attacks

Table 17 shows extensive black-box attack results of ME-Net on SVHN. We use standard ResNet-18 as the network, and use a total perturbation of $\varepsilon = 8/255$ (0.031). We use the same strong black-box attacks as previously used (i.e., transfer-,

decision-, and score-based attacks), and follow the same attack settings and parameters. As there are few results on SVHN dataset, we compare only with the vanilla model which uses the same network and training process as ME-Net. As shown, ME-Net provides significant adversarial robustness against various black-box attacks.

Method		Clean	FGSM	PGD			CW		Boundary	SPSA
				7 steps	20 steps	40 steps	$\kappa = 20$	$\kappa = 50$		
Vanilla		95.0%	31.2%	8.5%	1.8%	0.0%	20.4%	7.6%	4.5%	3.7%
ME-Net	$p : 0.8 \rightarrow 1$	**96.0%**	**91.8%**	**91.1%**	**90.9%**	**89.8%**	**95.5%**	**95.2%**	79.2%	**95.5%**
	$p : 0.6 \rightarrow 0.8$	95.5%	88.9%	88.7%	86.4%	86.2%	95.1%	94.9%	80.6%	94.6%
	$p : 0.4 \rightarrow 0.6$	94.0%	87.0%	86.4%	85.8%	84.4%	93.6%	93.4%	**85.3%**	93.8%
	$p : 0.2 \rightarrow 0.4$	88.3%	80.7%	76.4%	75.3%	74.2%	87.4%	87.4%	83.3%	87.6%

Table 17. **SVHN extensive black-box attack results.** Different kinds of strong black-box attacks are used, including transfer-, decision-, and score-based attacks.

Again, we strengthen the black-box adversary by providing the network architecture and weights of our trained model. We then apply various attacks and report the results in Table 18. ME-Net can still maintain high adversarial robustness under this setting.

Method		FGSM	PGD			CW	
			7 steps	20 steps	40 steps	$\kappa = 20$	$\kappa = 50$
ME-Net	$p : 0.8 \rightarrow 1$	83.8%	83.3%	81.3%	78.6%	**95.2%**	**95.0%**
	$p : 0.6 \rightarrow 0.8$	85.8%	85.7%	84.0%	82.1%	94.9%	94.8%
	$p : 0.4 \rightarrow 0.6$	**88.8%**	**88.6%**	**87.4%**	**86.8%**	93.5%	93.3%
	$p : 0.2 \rightarrow 0.4$	86.6%	86.3%	85.7%	85.5%	88.2%	88.2%

Table 18. **SVHN additional black-box attack results where adversary has limited access to the trained network.** We provide the architecture and weights of our trained model to the black-box adversary to make it stronger.

E.2. White-box Attacks

For white-box attacks, we set attack parameters the same as in CIFAR-10, and use strongest white-box BPDA attack with different attack steps (up to 1000 for convergence). We show results of both pure ME-Net and adversarially trained one. We use 7 steps for adversarial training. Since in (Madry et al., 2017) the authors did not provide results on SVHN, we follow their methods to retrain a model. The training process and hyper-parameters are kept identical to ME-Net.

Table 19 shows the extensive results under white-box attacks. ME-Net achieves significant adversarial robustness against the strongest white-box adversary, as it can consistently outperform (Madry et al., 2017) by a certain margin.

F. Additional Results on Tiny-ImageNet

In this section, we extend our experiments to evaluate ME-Net on a larger and more complex dataset. We use Tiny-ImageNet, which is a subset of ImageNet and contains 200 classes. Each class has 500 images for training and 50 for testing. All images are 64×64 colored ones. Since ME-Net requires to train the model from scratch, due to the limited computing resources, we do not provide results on even larger dataset such as ImageNet. However, we envision ME-Net to perform better on such larger datasets as it can leverage the global structures of those larger images.

F.1. Black-box Attacks

For black-box attacks on Tiny-ImageNet, we only report the Top-1 adversarial accuracy. We use standard DenseNet-121 (Huang et al., 2017) as our network, and set the attack parameters as having a total perturbation $\varepsilon = 8/255$ (0.031). We

Method		Type	Clean	Attack Steps				
				7	20	40	100	1000
Madry		Adv. train	87.4%	52.5%	48.4%	47.9%	47.5%	47.1%
ME-Net	$p: 0.8 \to 1$	Prep.	**96.0%**	42.1%	27.2%	14.2%	8.0%	7.2%
	$p: 0.6 \to 0.8$	Prep.	95.5%	52.4%	39.6%	28.2%	17.1%	15.9%
	$p: 0.4 \to 0.6$	Prep.	94.0%	60.3%	48.7%	40.1%	27.4%	25.8%
	$p: 0.2 \to 0.4$	Prep.	88.3%	**74.7%**	**61.4%**	**52.7%**	**44.0%**	**43.4%**
ME-Net	$p: 0.8 \to 1$	Prep. + Adv. train	**93.5%**	62.2%	41.4%	37.5%	35.5%	34.3%
	$p: 0.6 \to 0.8$	Prep. + Adv. train	92.6%	72.1%	57.1%	49.6%	47.8%	46.5%
	$p: 0.4 \to 0.6$	Prep. + Adv. train	91.2%	79.9%	69.1%	64.2%	62.3%	61.7%
	$p: 0.2 \to 0.4$	Prep. + Adv. train	87.6%	**83.5%**	**75.8%**	**71.9%**	**69.8%**	**69.4%**

Table 19. **SVHN extensive white-box attack results.** We apply up to 1000 steps PGD or BPDA attacks in white-box setting to ensure the results are convergent. We show results of both pure ME-Net and adversarially trained ones. ME-Net shows significantly better results as it consistently outperforms (Madry et al., 2017) by a certain margin.

use the same black-box attacks as before and follow the same attack settings. The extensive results are shown in Table 20.

Method		Clean	FGSM	PGD			CW		Boundary	SPSA
				7 steps	20 steps	40 steps	$\kappa = 20$	$\kappa = 50$		
Vanilla		66.4%	15.2%	1.3%	0.0%	0.0%	8.0%	7.7%	2.6%	1.2%
ME-Net	$p: 0.8 \to 1$	**67.7%**	**67.1%**	**66.3%**	**66.0%**	**65.8%**	**67.6%**	**67.4%**	**62.4%**	**67.4%**
	$p: 0.6 \to 0.8$	64.1%	63.6%	63.1%	63.1%	62.4%	63.8%	63.6%	61.9%	63.8%
	$p: 0.4 \to 0.6$	58.9%	54.8%	51.7%	51.6%	50.4%	58.2%	58.2%	58.9%	58.1%

Table 20. **Tiny-ImageNet extensive black-box attack results.** Different kinds of strong black-box attacks are used, including transfer-, decision-, and score-based attacks.

Further, additional black-box attack results are provided in Table 21, where the black-box adversary has limited access to ME-Net. The results again demonstrate the effectiveness of the preprocessing layer.

Method		FGSM	PGD			CW	
			7 steps	20 steps	40 steps	$\kappa = 20$	$\kappa = 50$
ME-Net	$p: 0.8 \to 1$	**66.5%**	**64.0%**	**62.6%**	59.1%	55.8%	56.0%
	$p: 0.6 \to 0.8$	61.1%	60.9%	60.7%	**59.2%**	57.6%	57.6%
	$p: 0.4 \to 0.6$	58.8%	58.2%	57.5%	56.9%	**58.3%**	**58.2%**

Table 21. **Tiny-ImageNet additional black-box attack results where adversary has limited access to the trained network.** We provide the architecture and weights of our trained model to the black-box adversary to make it stronger.

F.2. White-box Attacks

In white-box settings, we set the attack hyper-parameters as follows: a total perturbation of $8/255$ (0.031), a step size of $2/255$ (0.01), and 7 steps PGD for adversarial training. We still use strongest BPDA attack with different attack steps up to 1000. We re-implement (Madry et al., 2017) to be the baseline, and keep all training process the same for ME-Net

and (Madry et al., 2017). Finally, we report both Top-1 and Top-5 adversarial accuracy in Table 22, which demonstrates the significant adversarial robustness of ME-Net.

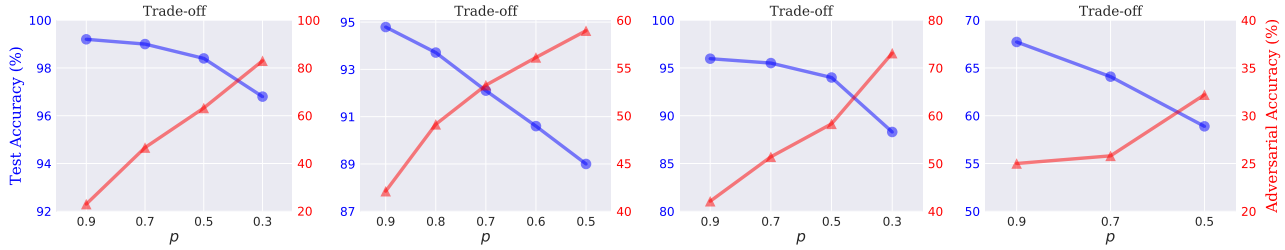
Metrics	Method	Type	Clean	Attack Steps				
				7	20	40	100	1000
Top-1	Madry	Adv. train	45.6%	23.3%	22.4%	22.4%	22.3%	22.1%
	ME-Net $p: 0.8 \rightarrow 1$	Prep. + Adv. train	53.9%	28.1%	25.7%	25.3%	25.0%	24.5%
	ME-Net $p: 0.6 \rightarrow 0.8$	Prep. + Adv. train	**57.0%**	33.7%	28.4%	27.3%	26.8%	26.3%
	ME-Net $p: 0.4 \rightarrow 0.6$	Prep. + Adv. train	55.6%	**38.8%**	**30.6%**	**29.4%**	**29.0%**	**28.5%**
Top-5	Madry	Adv. train	71.4%	47.5%	46.0%	45.9%	45.8%	45.0%
	ME-Net $p: 0.8 \rightarrow 1$	Prep. + Adv. train	77.4%	54.8%	52.2%	51.9%	51.2%	50.6%
	ME-Net $p: 0.6 \rightarrow 0.8$	Prep. + Adv. train	**80.3%**	62.1%	57.1%	56.7%	56.4%	55.1%
	ME-Net $p: 0.4 \rightarrow 0.6$	Prep. + Adv. train	78.8%	**66.7%**	**59.5%**	**58.5%**	**58.0%**	**56.9%**

Table 22. **Tiny-ImageNet extensive white-box attack results.** We apply up to 1000 steps PGD or BPDA attacks in white-box setting to ensure the results are convergent. We select (Madry et al., 2017) as the baseline and keep the training process the same for both (Madry et al., 2017) and ME-Net. We show both Top-1 and Top-5 adversarial accuracy under different attack steps. ME-Net shows advanced results by outperforming (Madry et al., 2017) consistently in both Top-1 and Top-5 adversarial accuracy.

G. Trade-off between Adversarial Robustness and Standard Generalization

In this section, we briefly discuss the trade-off between standard generalization and adversarial robustness, which can be affected by training ME-Net with different hyper-parameters. When the masks are generated with higher observing probability p, the recovered images will contain more details and are more similar to the original ones. In this case, the generalization ability will be similar to the vanilla network (or even be enhanced). However, the network will be sensible to the adversarial noises, as the adversarial structure in the noise is only destroyed a bit, and thus induces low robustness. On the other hand, when given lower observing probability p, much of the adversarial structure in the noise will be eliminated, which can greatly increase the adversarial robustness. Nevertheless, the generalization on clean data can decrease as it becomes harder to reconstruct the images and the input images may not be similar to the original ones. In summary, there exists an inherent trade-off between standard generalization and adversarial robustness. The trade-off should be further studied to acquire a better understanding and performance of ME-Net.

We provide results of the inherent trade-off between adversarial robustness and standard generalization on different datasets. As shown in Fig. 8, we change the observing probability p of the masks to train different ME-Net models, and apply 7 steps white-box BPDA attack to each of them. As p decreases, the generalization ability becomes lower, while the adversarial robustness grows rapidly. We show the consistent trade-off phenomena on different datasets.



(a) MNIST (b) CIFAR-10 (c) SVHN (d) Tiny-ImageNet

Figure 8. **The trade-off between adversarial robustness and standard generalization on different datasets**. We use pure ME-Net during training, and apply 7 steps white-box BPDA attack for the adversarial accuracy. For Tiny-ImageNet we only report the Top-1 accuracy. The results verify the consistent trade-off across different datasets.

H. Additional Results of Different ME Methods

H.1. Black-box Attacks

We first provide additional experimental results using different ME methods against black-box attacks. We train different ME-Net models on CIFAR-10 using three ME methods, including the USVT approach, the Soft-Impute algorithm and the Nuclear Norm minimization algorithm. The training processes are identical for all models. For the black-box adversary, we use different transfer-based attacks and report the results in Table 23.

Method	Complexity	Type	Clean	FGSM	PGD			CW	
					7 steps	20 steps	40 steps	$\kappa = 20$	$\kappa = 50$
Vanilla	–	–	93.4%	24.8%	7.6%	1.8%	0.0%	9.3%	8.9%
ME-Net - USVT	Low	Prep.	94.8%	90.5%	90.3%	89.4%	88.9%	**93.6%**	**93.6%**
ME-Net - Soft-Imp.	Medium	Prep.	**94.9%**	**92.2%**	**91.8%**	**91.8%**	**91.3%**	**93.6%**	93.5%
ME-Net - Nuc. Norm	High	Prep.	94.8%	92.0%	91.7%	91.4%	91.0%	93.3%	93.4%

Table 23. **Comparison between different ME methods against black-box attacks.** We report the generalization and adversarial robustness of three ME-Net models using different ME methods on CIFAR-10. We apply transfer-based black-box attacks as the adversary.

H.2. White-box Attacks

We further report the white-box attack results of different ME-Net models in Table 24. We use 7 steps PGD to adversarially train all ME-Net models with different ME methods on CIFAR-10. We apply up to 1000 steps strongest white-box BPDA attacks as the adversary. Compared to the previous state-of-the-art (Madry et al., 2017) on CIFAR-10, all the three ME-Net models can outperform them by a certain margin, while also achieving higher generalizations. The performance of different ME-Net models may vary slightly, where we can observe that more complex methods can lead to slightly better performance.

Method	Complexity	Type	Clean	Attack Steps				
				7	20	40	100	1000
Madry	–	Adv. train	79.4%	47.2%	45.6%	45.2%	45.1%	45.0%
ME-Net - USVT	Low	Prep. + Adv. train	**85.5%**	67.3%	55.8%	53.7%	52.6%	51.9%
ME-Net - Soft-Imp.	Medium	Prep. + Adv. train	**85.5%**	67.5%	56.5%	54.8%	53.0%	52.3%
ME-Net - Nuc. Norm	High	Prep. + Adv. train	85.0%	**68.2%**	**57.5%**	**55.4%**	**53.5%**	**52.8%**

Table 24. **Comparison between different ME methods against white-box attacks.** We adversarially trained three ME-Net models using different ME methods on CIFAR-10, and compare the results with (Madry et al., 2017). We apply up to 1000 steps PGD or BPDA white-box attacks as adversary.

I. Additional Studies of Attack Parameters

We present additional studies of attack parameters, including different random restarts and step sizes for further evaluations of ME-Net. Authors in (Mosbach et al., 2018) show that using multiple random restarts and different step sizes can drastically affect the performance of PGD adversaries. We consider the same white-box BPDA-based PGD adversary as in Table 4, and report the results on CIFAR-10. Note that with n random restarts, given an image, we consider a classifier successful only if it was not fooled by any of these n attacks. In addition, this also significantly increases the computational overhead. We hence fix the number of attack steps as 100 (results are almost flattened; see for example Fig. 6), and select three step sizes and restart values. We again compare ME-Net with (Madry et al., 2017).

As shown in Table 25, with different step sizes, the performance of ME-Net varies slightly. Specifically, the smaller the step

Method	Step sizes	Random restarts		
		10	20	50
Madry	$2/255$	43.4%	42.7%	41.7%
	$4/255$	43.8%	43.3%	41.9%
	$8/255$	44.0%	43.3%	41.9%
ME-Net	$2/255$	**48.7%**	**47.2%**	**44.8%**
	$4/255$	**49.7%**	**48.4%**	**45.2%**
	$8/255$	**50.8%**	**49.8%**	**46.0%**

Table 25. **Results of white-box attacks with different random restarts and step sizes on CIFAR-10.** We compare ME-Net with (Madry et al., 2017) using three different step sizes and random restart values. We apply 100 steps PGD or BPDA white-box attacks as adversary.

size (e.g., $2/255$) is, the stronger the adversary becomes for both ME-Net and (Madry et al., 2017). This is as expected, since a smaller step size enables a finer search for the adversarial perturbation.

ME-Net leverages randomness through masking, and it would be helpful to understand how random restarts, with a hard success criterion, affect the overall pipeline. As observed in Table 25, more restarts can reduce the robust accuracy by a few percent. However, we note that ME-Net can still outperform (Madry et al., 2017) by a certain margin across different attack parameters. We remark that arguably, one could potentially always handle such drawbacks by introducing restarts during training as well, so as to maximally match the training and testing conditions. This introduces in unnecessary overhead that might be less meaningful. We hence focus on other parameters such as the number of attack steps in the main paper.

J. Additional Benefits by Majority Voting

It is common to apply an ensemble or vote scheme during the prediction stage to further improve accuracy. ME-Net naturally provides a majority voting scheme. As we apply masks with different observation probability p during training, an intuitive method is to also use multiple masks with the same p (rather than only one p) for each image during inference, and output a majority vote over predicted labels. One can even use more masks with different p within the training range. By such, the training procedure and model can remain unchanged while the inference overhead only gets increased by a small factor.

Attack Steps	Method	MNIST	CIFAR-10	SVHN	Tiny-ImageNet	
					Top-1	Top-5
40	Standard	94.0%	55.4%	71.9%	29.4%	58.5%
	Vote	**95.9%**	**59.3%**	**76.0%**	**33.8%**	**68.9%**
100	Standard	91.8%	53.5%	69.8%	29.0%	58.0%
	Vote	**94.2%**	**56.2%**	**73.1%**	**31.2%**	**65.4%**
1000	Standard	91.0%	52.8%	69.4%	28.5%	56.9%
	Vote	**92.6%**	**54.2%**	**71.4%**	**29.8%**	**59.5%**

Table 26. **Comparison between majority vote and standard inference.** For each image, we apply 10 masks with same p used during training, and the model outputs a majority vote over predicted labels. The standard inference only uses one mask with the mean probability of those during training. We use 40, 100 and 1000 steps white-box BPDA attack and report the results on each dataset.

In Table 26, we report the majority voting result of ME-Net on different datasets, where voting can consistently improve the adversarial robustness of the standard one by a certain margin. This is especially helpful in real-world settings where the defender can get more robust output without highly increasing the computational overhead. Note that by using majority vote, we can further improve the state-of-the-art white-box robustness.

K. Hyper-Parameters Study

K.1. Observation Probability p

As studied previously, by applying different masks with different observation probability p, the performance of ME-Net can change differently. We have already reported extensive quantitative results of different ME-Net models trained with different p. Here we present the qualitative results by visualizing the effect of different p on the original images. As illustrated in Fig. 9, the first row shows the masked image with different p, and the second row shows the recovered image by ME. It can be observed that the global structure of the image is maintained even when p is small.

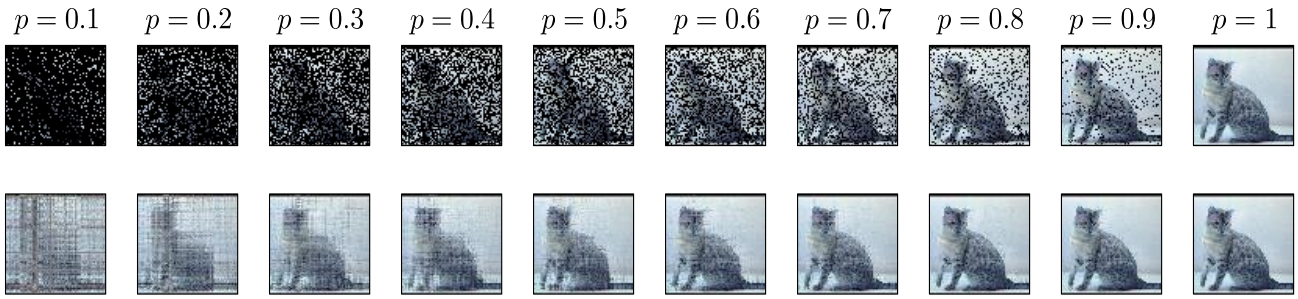


Figure 9. **Visualization of ME result with different observation probability p. First row:** Images after applying masks with different observation probabilities. **Second row:** The recovered images by applying ME. We can observe that the global structure of the image is maintained even when p is small.

K.2. Number of Selected Masks

Another hyper-parameter of ME-Net is the number of selected masked images for each input image. In the main paper, all experiments are carried out using 10 masks. We here provide the hyper-parameter study on how the number of masks affects the performance of ME-Net. We train ME-Net models on CIFAR-10 using different number of masks and keep other settings the same. In Table 27, we show the results of both standard generalization and adversarial robustness. We use transfer-based 40 steps PGD as black-box adversary, and 1000 steps BPDA as white-box adversary. As expected, using more masks can lead to better performances. Due to the limited computation resources, we only try a maximum of 10 masks for each image. However, we expect ME-Net to perform even better with more sampled masks and better-tuned hyper-parameters.

# of Masks	Method		Clean	Black-box	White-box
–	Vanilla		93.4%	0.0%	0.0%
1	ME-Net	$p : 0.9$	92.7%	82.3%	44.1%
		$p : 0.5$	79.8%	59.7%	47.4%
5	ME-Net	$p : 0.8 \rightarrow 1$	94.1%	87.8%	46.5%
		$p : 0.4 \rightarrow 0.6$	86.3%	68.5%	49.3%
10	ME-Net	$p : 0.8 \rightarrow 1$	**94.9%**	**91.3%**	47.4%
		$p : 0.4 \rightarrow 0.6$	89.2%	70.9%	**52.8%**

Table 27. **Comparisons between different number of masked images used for each input image.** We report the generalization and adversarial robustness of ME-Net models trained with different number of masks on CIFAR-10. We apply transfer-based 40 steps PGD attack as black-box adversary, and 1000 steps PGD-based BPDA as white-box adversary.

L. Additional Visualization Results

We finally provide more visualization results of ME-Net applied to clean images, adversarial images, and their differences. We choose Tiny-ImageNet since it has a higher resolution. As shown in Fig. 10, for vanilla model, the highly structured adversarial noises are distributed over the entire image, containing human imperceptible adversarial structure that is very

likely to fool the network. In contrast, the redistributed noises in the reconstructed images from ME-Net mainly focus on the global structure of the images, which is well aligned with human perception. As such, we would expect ME-Net to be more robust against adversarial attacks.

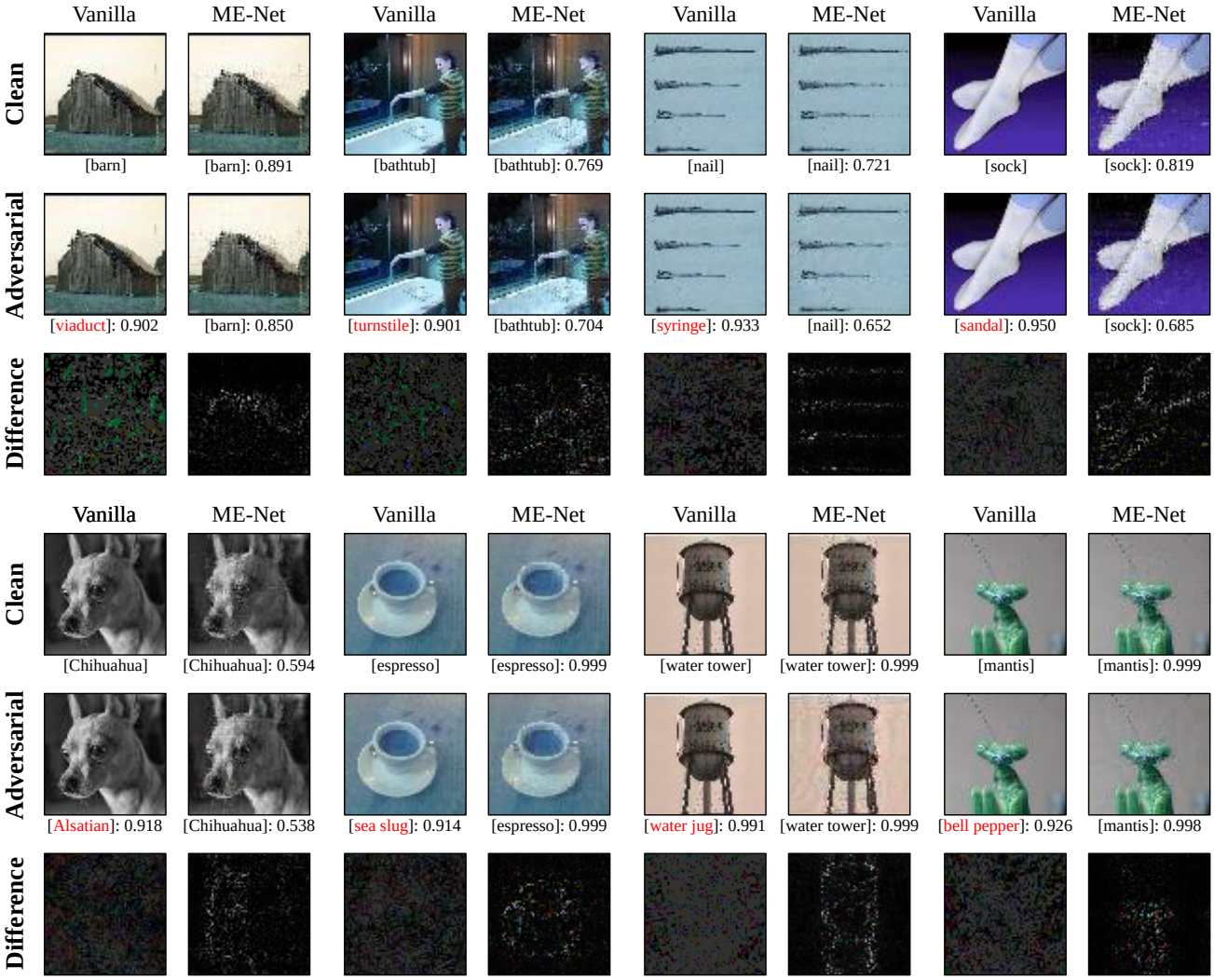


Figure 10. **Visualization of ME-Net applied to clean images, adversarial images, and their differences on Tiny-ImageNet. First column** from top to bottom: the clean image, the adversarial example generated by PGD attacks, the difference between them (i.e., the adversarial noises). **Second column** from top to bottom: the reconstructed clean image by ME-Net, the reconstructed adversarial example by ME-Net after performing PGD attacks, the difference between them (i.e., the redistributed noises). Underlying each image is the predicted class and its probability. We multiply the difference images by a constant scaling factor to increase the visibility. The differences between the reconstructed clean image by ME-Net and the reconstructed adversarial example by ME-Net after performing PGD attacks, i.e., the new adversarial noises, are redistributed to the global structure.