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.

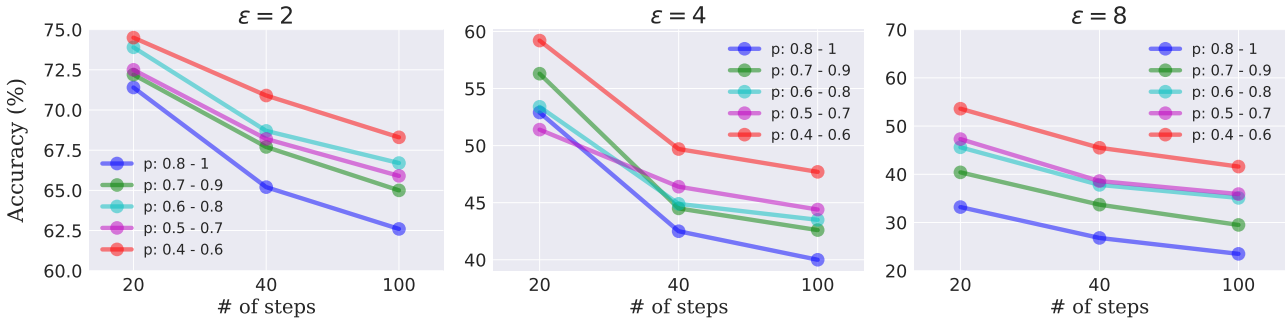


Figure 7. **CIFAR-10 white-box attack results of pure ME-Net with different perturbation** ε. We report ME-Net results with different training settings under various attack steps.

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.

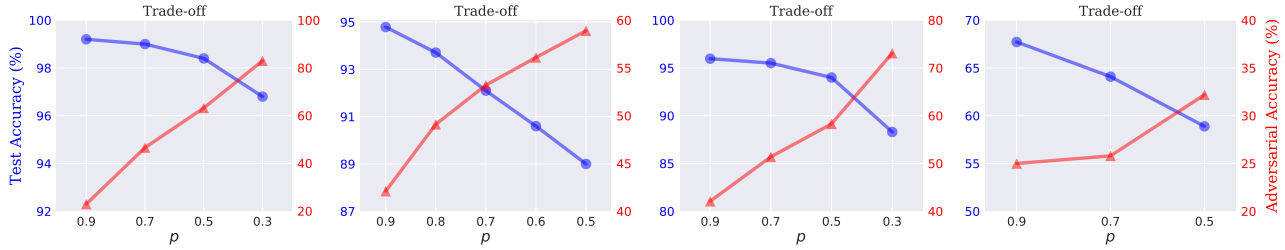
Metrics	Method	Type	Clean	Attack Steps				
				7	20	40	100	1000
Top-1	Madry	Adv. train	45.6%	23.3%	22.4%	22.4%	22.3%	22.1%
	ME-Net $p: 0.8 \rightarrow 1$	Prep. + Adv. train	53.9%	28.1%	25.7%	25.3%	25.0%	24.5%
	ME-Net $p: 0.6 \rightarrow 0.8$	Prep. + Adv. train	**57.0%**	33.7%	28.4%	27.3%	26.8%	26.3%
	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.

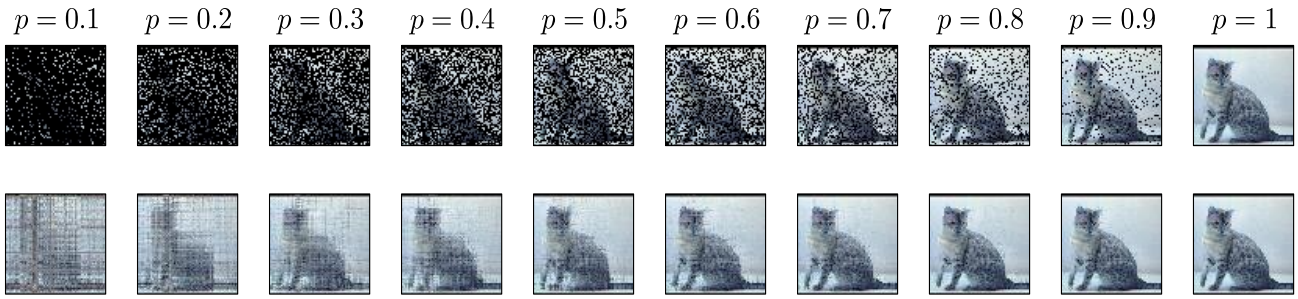


Figure 9. **Visualization of ME result with different observation probability p. First row:** Images after applying masks with different observation probabilities. **Second row:** The recovered images by applying ME. We can observe that the global structure of the image is maintained even when p is small.

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.

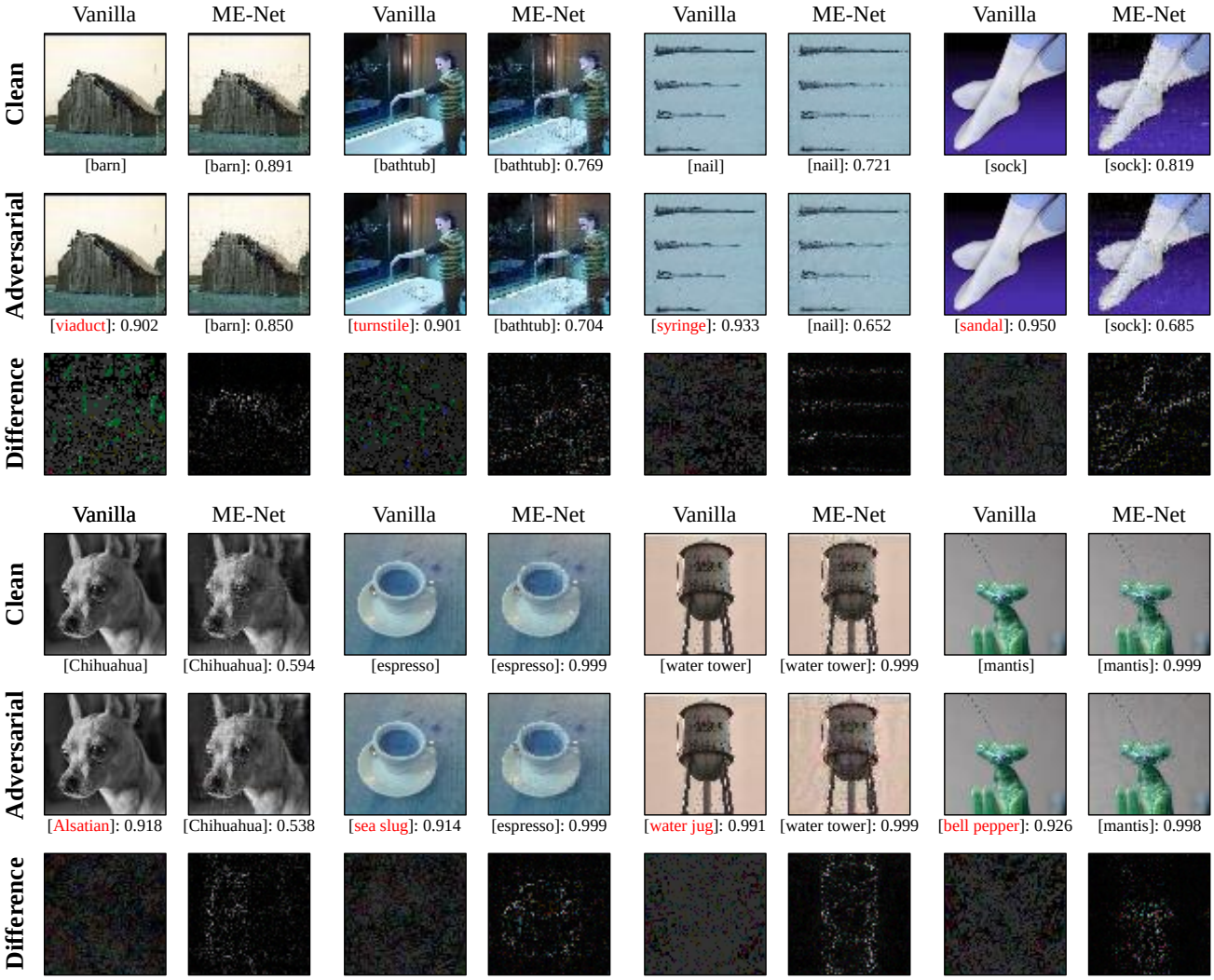


Figure 10. **Visualization of ME-Net applied to clean images, adversarial images, and their differences on Tiny-ImageNet. First column** from top to bottom: the clean image, the adversarial example generated by PGD attacks, the difference between them (i.e., the adversarial noises). **Second column** 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.