Evidence-based Multi-Feature Fusion for Adversarial Robustness

Zheng Wang, Xing Xu, Lei Zhu, Yi Bin, Guoqing Wang,
Yang Yang *Senior Member, IEEE,* Heng Tao Shen *Fellow, IEEE*

Abstract—The accumulation of adversarial perturbations in the feature space makes it impossible for Deep Neural Networks (DNNs) to know what features are robust and reliable, and thus DNNs can be fooled by relying on a single contaminated feature. Numerous defense strategies attempt to improve their robustness by denoising, deactivating, or recalibrating non-robust features. Despite their effectiveness, we still argue that these methods are under-explored in terms of determining how trustworthy the features are. To address this issue, we propose a novel **E**vidence-based **M**ulti-**F**eature **F**usion (termed **EMFF**) for adversarial robustness. Specifically, our EMFF approach introduces evidential deep learning to help DNNs quantify the belief mass and uncertainty of the contaminated features. Subsequently, a novel multi-feature evidential fusion mechanism based on Dempster's rule is proposed to fuse the trusted features of multiple blocks within an architecture, which further helps DNNs avoid the induction of a single manipulated feature and thus improve their robustness. Comprehensive experiments confirm that compared with existing defense techniques, our novel EMFF method has obvious advantages and effectiveness in both scenarios of white-box and black-box attacks, and also prove that by integrating into several adversarial training strategies, we can improve the robustness of across distinct architectures, including traditional CNNs and recent vision Transformers with a few extra parameters and almost the same cost.

Index Terms—Adversarial Attack, Adversarial Robustness, Multi-feature Fusion, Evidential Deep Learning

1 Introduction

The success of deep neural networks (DNNs) [1], [2], [3] has greatly boosted the intelligence of various agents in computer vision [4], [5], and even makes them exceed human-level performance. However, adversarial examples [6], [7], [8], [9] constructed by maliciously adding imperceptible perturbations to benign images are notorious for subverting the decisions of these DNNs. Even if the adversaries are denied access to the details of target models, the adversarial examples are still effective [10], [11]. Obviously, this phenomenon poses a great threat to real-world cases, especially those models deployed in safety-critical fields, including autonomous driving, medical image analysis, and face recognition. Therefore, the design of various effective defense strategies to improve the adversarial robustness of DNNs has motivated a widespread concern and become imperative.

Till now, various defense technologies have been proposed and can be roughly divided into two categories including adversarial training [8], [12], [13] and feature manipulation [14], [15]. The former is the most successful and effective strategy to enhance the robustness of models by training them on a series of adversarial examples, and considerable efforts have recently been continuously invested in improving its efficiency [16] and efficacy [17]. However, even with adversarial training, the perturbations in the feature space will be gradually increased by the transformations of neural layers on the imperceptible noise at pixel level, which eventually causes the network to make incorrect predictions [14]. Thus, another parallel line of defense made the utmost attempt to learn robust representations by explicitly manipulating intermediate features, including batch normalization [18], [19], non-robust representation separation [20], [21], feature denoising [14], [22] and deactivation [23], [24], and frequency bias exploration [25].

Despite their remarkable advancements, the aforementioned defensive techniques are under-explored in terms of how to learn more trustworthy features. Nevertheless, this issue is a challenging task due to the requirements of: 1) **Feature confidence quantification**. The existing robust representation learning can not explicitly tell us "How confident is the learned feature" and "Why is the model more or less robust?", while they solely present the predictions. Toward this end, an intuitive idea is to introduce uncertainty into the robust representation learning of the model, which makes the learned features more reliable and trustworthy, then lowers or defends the attack of adversarial examples. 2) **Evidential feature aggregation**. Generally, the dominant backbones [5], [26] only regard the output of their final block as discriminative representation. Such a plain paradigm may be very vulnerable to providing unreliable decisions, especially when the final features are contaminated by adversarial perturbations. Visualized by Figure 1, the possible reason for this case is that every block can capture different proportions of the ground-truth feature, but only a single view from the final output, especially based on gradually exacerbated perturbations, does not capture the comprehensive range

- *Zheng Wang, Lei Zhu, Yi Bin, and Heng Tao Shen are with Tongji University, China; Xing Xu, Guoqing Wang and Yang Yang are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. (e-mail: zh_wang@hotmail.com; xing.xu@uestc.edu.cn; leizhu0608@gmail.com; yi.bin@hotmail.com; gqwang0420@hotmail.com; yang.yang@uestc.edu.cn; shenhengtao@hotmail.com).*

This work was supported by the National Natural Science Foundation of China (62306065, 62220106008, and 62476201), the Fundamental Research Funds for the Central Universities. (Corresponding author: Xing Xu, Heng Tao Shen.) Manuscript received Jun xx, 2024; revised xx xx, 202x.

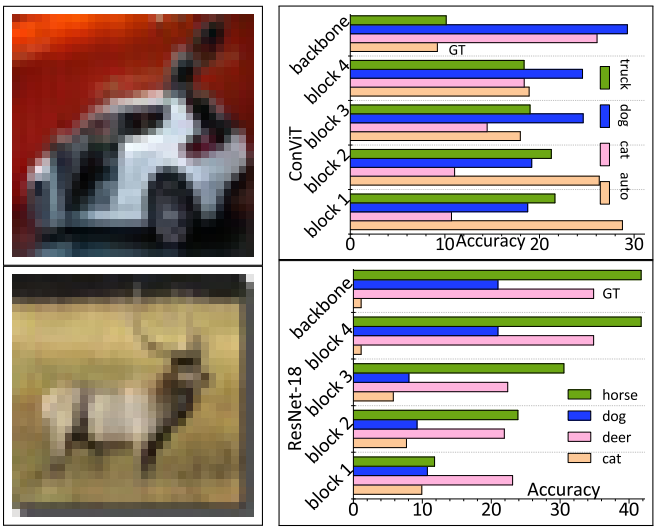

Fig. 1. Motivation illustration: Each block in a backbone can capture somewhat ground-truth features, if the classifier only with the perturbed output from the final block, it is thus easily induced by locally contaminated features to make incorrect predictions. **Left**: Adversarial examples based on CIFAR-10; **Right**: Top-4 classification performance of different blocks from diverse backbones.

of robust features, which thus makes the model prone to incorrect predictions.

Enlightened by evidential deep learning [27], [28], [29], we propose a novel **E**vidence-based **M**ulti-**F**eature **F**usion (termed **EMFF**) method to solve the aforementioned challenges. To be specific, our EMFF method exploits the evidential learning paradigm to capture and leverage the uncertainty brought by adversarial perturbation to enhance the robust representation learning, thus defending the adversarial attack. Furthermore, to address the issue that a single block is not comprehensive, another evidential feature fusion mechanism is introduced to elegantly absorb the strengths of features from different blocks for adversarial robustness. Firstly, the Dirichlet distribution is utilized to model the distribution of class probabilities, parameterized with evidence from different blocks. The Dempster-Shafer theory [30] is then leveraged to fuse multi-evidential feature for robustness enhancement. Additionally, our EMFF method can be regarded as an easy plugin due to its simplicity, which can be efficiently implanted into various backbones of different architectures for defense promotion.

- We introduce evidential deep learning to enhance the adversarial robustness of DNNs by quantifying the uncertainty of the learned representation.
- To further promote the reliability of representation learning, we present a novel fusion mechanism of different features from multiple blocks within the backbone.
- Additionally, our EMFF method can be leveraged as an easy plugin to effectively promote the robustness of different backbones of classical CNNs and advanced vision Transformers.
- Experimentally, comprehensive results and solid ablation studies demonstrate the efficiency and efficacy of our EMFF method in defending various white- and black-box attacks on several datasets.

The paper is organized as follows: We make a comprehensive investigation of the related work in Section 2. Section 3 dwells on our novel EMFF method for adversarial robustness in different backbones. In Section 4, extensive experiments on several widely-used datasets are conducted to evaluate our effectiveness, superiority and efficiency. The final section concludes our work.

2 Related Work

We comprehensively surveyed the literature on four topics including: adversarial examples, adversarial training and feature space manipulation for adversarial defense, and evidential deep learning.

Adversarial Examples created by the attackers with perturbations that are imperceptible to humans and added to benign images, are fed into DNNs that will cause them to produce wrong outputs [8], [31]. The classical work of Fast Gradient Sign Method (FGSM) [8] pioneered the gradient information of a given image to generate effective adversarial examples, which can be formalized as:

$$\hat{x} = x + \Delta = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f_\theta(x, y)), \tag{1}$$

where $\hat{x}$ defines the crafted adversarial example, Δ is a perturbation constrained by a ℓ_p-norm ball space with radius ϵ at the benign image x, f_θ is a DNN-based classifier with the parameter of θ, and $\mathcal{L}$ defines the classification loss. Within expectation, the perturbed image $\hat{x}$ can cause the victim classifier f_θ to output wrong results, *i.e.* $f_\theta(\hat{x}) \neq y$, where y denotes the ground-truth label of the benign image x. Subsequently, many elegant strategies with better performance of adversarial attack are constructed based on the Iterative version of FGSM (I-FGSM) [32] by combining with some advanced technologies, such as momentum [33], nesterov's accelerated gradient [34], customized iteration and sampling [35], an acceleration framework composed of sLocator and sRudder [36], and so on.

Project Gradient Descent (PGD) [12] is another well-known and widely used method to generate adversarial examples via multi-step projected gradient descent on a ϵ-ball, formulated as:

$$x_{t+1} = \Pi\left(x_t + \eta \cdot \text{sign}(\nabla_x \mathcal{L}(f_\theta(x, y)))\right), \tag{2}$$

where η is the step size, Π is the project function, and x_t is the adversarial example generated at the t-th step. Recently, AutoAttack [37] developed two extensions of PGD-attack to break through many of the recent defense strategies. Hence, FGSM, PGD, C&W [38], and AutoAttack are treated as popular attacks to evaluate the robustness of DNN models.

However, the aforementioned attack strategies fall in white-box attacks, which may not be practical in the real world as they require the full access to the knowledge of the DNN model. As a result, recent efforts have been spent on designing black-box attacks, which only need the inputs and outputs of the DNNs. Related recent approaches [39], [40], [41] study the transferability of adversarial examples, where they commonly use a known model to generate adversarial examples that can attack an unknown target model. We call this kind of method a transfer-based attack.

Differently, our main focus is to design an effective robust strategy for various backbones, albeit evaluating the

robustness of different models after combining them with our EMFF method against white- and black-box attacks, such as PGD-attack, AutoAttack, and transfer-based attacks.

Adversarial Training has been adopted in many real-world cases including privacy preserving [42], X-Rays classification [43] as the most effective strategy to defend against adversarial attacks, which can steer the models towards robustness by training them on a set of worst-case adversarial examples, generated by the minimax optimization:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\Delta} \mathcal{L}(f_\theta(x+\Delta), y) \right], \tag{3}$$

where $x \in \mathcal{D}$ is a benign image with ground-truth label y. Obviously, the inner maximization corresponds to numerous optimization strategies for the generation of adversarial examples, such as Fast Gradient Sign Method (FGSM) [8], Projected Gradient Descent (PGD) attack [12], FGSM variants with momentum [33], diverse input [40], translation-invariant attack [39], scale-invariant attack [44], variance tuning [45], and so on. While the outer minimization aims to robustify the model with the generated adversarial examples of the worst cases. This is the process of adversarial training for robustness enhancement.

Many elegant strategies of adversarial training [46], [47], [48], [49], [50], [51] have been proposed. For instance, TRADES [46] decomposed the prediction error of adversarial examples as the classification error and the boundary error, and then improved robustness via the theory of classification-calibrated loss. MART [47] additionally investigated the influence of misclassified examples on the final robustness of adversarial training. S^2O [50] treated model weights as random variables to enhance robustness through second-order statistics optimization. Recently, single-step adversarial training [51], [52] has received wide attention to improve the efficiency [16] and effectiveness [17]. Stylized Adversarial Defense (SAD) [53] proposes a max-margin adversarial training approach that minimizes the distance between source image and its adversary and maximizes the distance between the adversary and the target image. Lately, Variational Adversarial Defense [54] introduces a novel training scheme based on the distribution of adversarial examples to upgrade the defend scheme from local point-wise to distribution-wise, which yields an enlarged support region for safeguarding robust training and thus possesses a higher promising in defense. In the meantime, Topology Aligning Adversarial Training (TAAT) [31] algorithm takes full advantage of the topology information to maintain consistency in the topological structure within the feature space of both natural and adversarial examples. Adaptive Attention Scaling (AAS) [55] attack automatically finds the optimal scaling factors of pre-softmax outputs using gradient-based optimization. Self-Guided Label Refinement [56] self-refines a more accurate and informative label distribution from over-confident hard labels for adversarial training. Random entangled image Transformer (ReiT) [57] seamlessly integrates adversarial training and randomization to only bolster the adversarial robustness of vision transformers.

Unlike them, our EMFF method introduces evidence to steer representation learning towards robustness. Due to its simplicity, our novel plugin can be easily integrated into these adversarial training strategies across distinct backbones for robustness enhancement.

Feature Space Manipulation is another line to defend against adversarial attacks, which aims to learn robust features via numerous elegant strategies. For example, several works [58], [59], [60] attempted to prune out a random set of activations with small magnitudes or vulnerable to adversarial attacks. In the meantime, some other attempts explicitly manipulated the features to reduce abnormalities by masking out the feature that is sensitive to perturbations [61] or deactivating unreliable features [62]. From the view of noise removal, there are also considerable efforts on denoising techniques. Specifically, feature denoising [14], compress and defend [15], feature purification [63], high-level representation guided denoiser [22] were subsequently proposed to deactivate abnormal features activated by noises. Inspired by [20] which provided the existence of non-robust features in the dataset, non-robust feature separations were recently explored. Channel Activation Suppression (CAS) [23] and Channel-wise Importance-based Feature Selection (CIFS) [24] deactivated the activation of non-robust features from the perspective of channels. Additionally, Marine Picot et. al [64] proposed an information-geometric formulation of adversarial defense, which is based on the geodesic distance between the softmax outputs corresponding to natural and perturbed input features. SAD [53] utilized the style and content information of the target from another class, alongside its class-boundary information to create adversarial perturbations. Semantic Constraint Adversarial Robust Learning (SCARL) [65] maximized mutual information to bridge the gap between the visual representations and the corresponding semantic word vectors in the embedding space. Principal Latent Space (PrincipaLS) [66] manipulated the latent space of novelty detectors to improve the robustness against adversarial examples. The latest work of Feature Separation and Recalibration (FSR) [21] firstly disentangled the feature map into the robust part and non-robust one, and then the latter was re-calibrated to restore the potentially useful cues for model predictions. Another latest effort of Frequency Preference Control Module (FPCM) [25] adaptively reconfigured the low- and high-frequency components of intermediate features to make full use of frequency in robust learning.

In contrast, we propose an explainable multi-evidence fusion strategy that makes robust learning more reliable through deep evidential learning, and thus eliminating the interference of non-robust features in a more certain way.

Evidential Deep Learning. Most DNNs are deterministic predictions and cannot quantify their predictive uncertainty, which thus causes the false predictions to be overconfident [67]. Therefore, Evidential Deep Learning (EDL) [28], [29] aims to make DNNs know "what they don't know" and fall back onto a prior belief. Combing the Dempster-Shafer theory [30], [68] of evidence with Subjective Logic theory [69], EDL explicitly assesses the uncertainty and collects evidence for each category by placing the Dirichlet distribution on the class probability. Recently, EDL has successfully been applied in various computer vision tasks, including open-set recognition [70], [71], stereo matching [72], regression [73] and long-tail learning [74]. Some pioneering works also transfer the evidence collection into video understanding, such as temporal localization [75], [76], audio-visual event

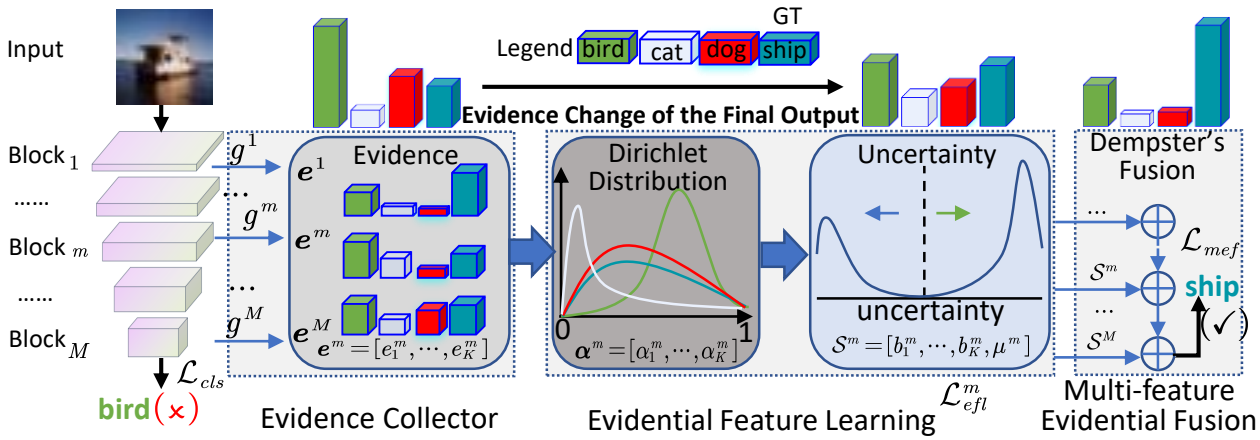

Fig. 2. Flowchart of our EMFF method includes three parts: Evidence Collector, Evidential Feature Learning and Multi-Evidential Fusion.

perception [27]. However, the above approaches neglect the enhancement of robustness against adversarial perturbations brought by evidence learning, which is our focus first presented here.

3 Our Approach

Conventional neural-network-based classifiers f_θ of K categories often directly take the final output feature as the distinctive representation of the ith image to make predictions. Their training is based on the classical cross-entropy loss, which can be denoted as:

$$\mathcal{L}_{cls} = \sum_{k=1}^{K} -y_{ik} \log(p_{ik}), \tag{4}$$

where p_{ik} is the predicted probability of the final feature for class k and y_{ik} is the corresponding ground-truth label.

However, we argue that such a plain paradigm may be very vulnerable to providing unreliable predictions, especially when the final feature is heavily contaminated by adversarial perturbations. To this end, the proposed method mainly includes three parts: the first part attempts to collect the evidence for each class; The second part associates the evidence with Dirichlet distribution to quantify the uncertainty of the representation; The third part combines multiple evidential features through Dempster-Shafer theory to aggregate evidence and thus enhance robustness through training for more evidential outputs, displayed as Figure 2.

3.1 Evidence Collector

As aforementioned, the traditional cross-entropy loss is incapable of explicitly collecting evidential outputs in an uncertainty-aware manner. Thus, we follow recent EDL-based works [27], [71] to generate evidence from perturbed features output by different blocks. Specifically, the evidence on robustness of the k^{th} category for the feature $\mathbf{x}^m$ output by the m^{th} block is a scalar, denoted as:

$$e_k^m = g^m(f_\theta^m(\mathbf{x}^m)), \tag{5}$$

where $f(\cdot)$ is the backbone of a common classifier parameterized with θ, such as ResNet [5], wide-ResNet [77], Transformers [4], [26], etc. $f^m(\cdot)$ accordingly denotes its m^{th} block. $g^m(\cdot)$ defines the evidence function, which is a full connection layer configured by $d_{\mathbf{x}^m} * K$ ($d_{\mathbf{x}^m}$ indicates the corresponding dimension of m^{th} feature, K is the category number) and then followed by an activation function, such as SoftPlus, Exp or ReLU, to keep the obtained evidence non-negative. For brevity, we have omitted the subscript i.

3.2 Evidential Feature Learning

Inspired by evidential feature learning that follows Subjective Logic [69] to quantify the classification uncertainty, we introduce it to steer blocks towards evidential outputs for robust representation learning. It can establish a connection between the parameters of the Dirichlet distribution and the belief distribution. In this context, the Dirichlet distribution can be conceptualized as the conjugate prior to the classification distribution. Subjective logic, on the other hand, involves the allocation of a belief mass to each class label and an overall uncertainty mass based on the collected evidence for the feature. The sum of these masses is constant and always equal to 1, denoted as:

$$u^m + \sum_{k=1}^{K} b_k^m = 1, \tag{6}$$

where both u^m and b_k^m are non-negative and indicate the overall uncertainty and the probability (belief mass) about the m^{th} block output for the k^{th} class, respectively. Their derivations can be straightforwardly achieved by employing the combination of Eq. 5 and Eq. 7.

The evidence $\mathbf{e}^m = [e_1^m, \cdots, e_K^m]$ collected by Eq. 5 can then be related to the belief mass assignment, which can be connected to the parameters of the Dirichlet distribution $\boldsymbol{\alpha}^m = [\alpha_1^m, \cdots, \alpha_K^m]$ by the formulas of $\alpha_k^m = e_k^m + 1$. Accordingly, b_k^m and u^m can be derived as:

$$b_k^m = \frac{e_k^m}{L^m} = \frac{\alpha_k^m - 1}{L^m} \quad \text{and} \quad u^m = \frac{K}{L^m}, \tag{7}$$

where $L^m = \sum_{k=1}^{K} (e_k^m + 1) = \sum_{k=1}^{K} \alpha_i^m$ is configured as the Dirichlet distribution strength. The readers can refer to [68] for more details. Obviously, if more evidence about the k^{th} category can be collected on the feature $\mathbf{x}^m$ of the m^{th} block, the probability (belief mass) to the k^{th} class would be greater. Consequently, the belief assignment

$\mathbf{b}^m = [b_1^m, \cdots, b_K^m]$ can be treated as subjective opinions associated to the Dirichlet distribution with parameters $\boldsymbol{\alpha}^m = [\alpha_1^m, \cdots, \alpha_K^m]$ that behaves as humans and is immune from those imperceptible perturbations. Given an opinion, the expected class probability p_k^m of the corresponding Dirichlet distribution $\mathbf{p}^m$ can be estimated as $p_k^m = \frac{\alpha_k^m}{L^m}$ [78].

Based on the obtained parameter $\boldsymbol{\alpha}^m$, the multinomial opinions $D(\mathbf{p}^m|\boldsymbol{\alpha}^m)$ for i^{th} image could be formed over evidence to represent the density of each probability assignment, where $\mathbf{p}^m$ is the class assignment probabilities on a K-dimensional unit simplex. This process can be defined as:

$$D(\mathbf{p}_i^m|\boldsymbol{\alpha}_i^m) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha}_i^m)} \prod_{k=1}^K (p_{ik}^m)^{\alpha_{ik}^m - 1} d\mathbf{p}_i^m, \\ 0, \quad otherwise. \end{cases} \tag{8}$$

We then combine the classical cross-entropy loss (Eq. 4) with evidential learning via a simple modification, derived as:

$$\begin{aligned} &\mathcal{L}_e(\boldsymbol{\alpha}_i^m, \mathbf{y}_i) = \int [\textstyle\sum_{k=1}^K -y_{ik} \log(p_{ik}^m)] D(\mathbf{p}_i^m|\boldsymbol{\alpha}_i^m) d\mathbf{p}_i^m \\ &= \int [\textstyle\sum_{k=1}^K -y_{ik} \log(p_{ik}^m)] \frac{1}{B(\boldsymbol{\alpha}_i^m)} \prod_{k=1}^K (p_{ik}^m)^{\alpha_{ik}^m - 1} d\mathbf{p}_i^m \\ &= \textstyle\sum_{k=1}^K y_{ik} \left(\psi\left(L_i^m\right) - \psi\left(\alpha_{ik}^m\right)\right), \end{aligned} \tag{9}$$

where $\psi(\cdot)$ is the digamma function and $B(\boldsymbol{\alpha}_i^m)$ is the K-dimensional multinomial beta function for the m^{th}-block feature of the i^{th} image. Eq. 9 is the integral of the cross-entropy loss function on the simplex determined by $\boldsymbol{\alpha}$.

For clarity, we provide an example to illustrate the above formulation only for the final block. Assuming that the belief mass assignment for a triple classification is represented by $\mathbf{b} = [0, \cdots, 0]$. Then, the prior distribution for a given image becomes a uniform, i.e., $D(\mathbf{p}|(1, \cdots, 1))$ — a Dirichlet distribution whose parameters are all one, as there is no observed evidence and the belief masses are all zero. This means that the uniform opinion does not contain any information, and implies total uncertainty, i.e., $u = 1$. Through the process of training, the belief masses may evolve to become $\mathbf{b} = [0.7, 0, 0]$, indicating that the total belief in the opinion is 0.7 and the remaining 0.3 is attributed to uncertainty. Consequently, the Dirichlet strength $L = 3/0.3 = 10$ as $K = 3$. Therefore, the new evidence for the first class is derived as $10 \times 0.7 = 7$. In this case, the opinion corresponds to the Dirichlet distribution $D(\mathbf{p}|(7, 1, 1))$.

The ultimate goal of the adversarial attack is to make the classifier classify adversarial examples into incorrect labels. While $\mathcal{L}_e$ can improve robustness by guaranteeing that the model generates more evidence for the correct labels than for other classes, it cannot guarantee that the model collects less evidence for the wrong classes. Consequently, the evidence for incorrect labels is best narrowed down to 0. To this end, a Kullback-Leibler (KL) divergence term is incorporated to regularize our predictive distribution by penalizing the divergences from the uncertainty, defined as:

$$\begin{aligned} &\mathcal{L}_{KL}(\boldsymbol{\alpha}_i^m, \mathbf{y}_i) = KL\left[D\left(\mathbf{p}_i^m|\tilde{\boldsymbol{\alpha}}_i^m\right) \| D\left(\mathbf{p}_i^m|\mathbf{1}\right)\right] \\ &= \log\left(\frac{\Gamma\left(\sum_{k=1}^K \tilde{\alpha}_{ik}^m\right)}{\Gamma(K) \prod_{k=1}^K \Gamma\left(\tilde{\alpha}_{ik}^m\right)}\right) \\ &+ \textstyle\sum_{k=1}^K \left(\tilde{\alpha}_{ik}^m - 1\right) \left[\psi\left(\tilde{\alpha}_{ik}^m\right) - \psi\left(\sum_{k=1}^K \tilde{\alpha}_{ik}^m\right)\right], \end{aligned} \tag{10}$$

where $\mathbf{1}$ indicates a constant vector of K ones, $\tilde{\boldsymbol{\alpha}}_i^m = \mathbf{y}_i + (1 - \mathbf{y}_i) \odot \boldsymbol{\alpha}_i^m$ is the adjusted Dirichlet parameter by removing the unreliable evidence from the predicted Dirichlet distribution, and $\Gamma(\cdot)$ is the gamma function. Hence, the loss of evidential feature learning for the i^{th} image on the m^{th} block feature can be easily figured out:

$$\mathcal{L}_{efl}^m = \mathcal{L}_e(\boldsymbol{\alpha}_i^m, \mathbf{y}_i) + \lambda_t \mathcal{L}_{KL}(\boldsymbol{\alpha}_i^m, \mathbf{y}_i), \tag{11}$$

where $\lambda_t = min(1.0, t/max\_epoch) \in [0, 1)$ is the annealing factor, and t indicates the current epoch. Experimentally, the effect of gradually increasing KL regularization by λ_t can prevent less exploration of parameter space and premature convergence to a flat uniform distribution.

3.3 Multi-feature Evidential Fusion

As aforementioned, only with the final feature, especially under the case of adversarial perturbations, the neural models cannot comprehensively capture all the reliable cues and thus be prone to make incorrect predictions. Motivated by the Dempster-Shafer theory [68], we accordingly propose a novel multi-feature evidential fusion to promote robustness. Practically, these evidences coming from different blocks within the same architecture can be fused only with such minor changes of adding evidence collectors, which facilitates our integration with existing methods. Empirically, when the aggregation of evidence from various sources comes to a degree of belief, the classifier can be promoted not to overfocus on local features contaminated by perturbations that make DNNs lose the global picture on robustness.

Specifically, we take two subjective logics $\mathcal{S}^1 = \left\{\{b_k^1\}_{k=1}^K, u^1\right\}$ and $\mathcal{S}^2 = \left\{\{b_k^2\}_{k=1}^K, u^2\right\}$ as an example to elaborate on Dempster's combination for the joint opinion $\mathcal{S} = \left\{\{b_k\}_{k=1}^K, u\right\}$, displayed in the following manner: $\mathcal{S} = \mathcal{S}^1 \oplus \mathcal{S}^2$, where $\oplus$ defines the fusion operator as:

$$b_k = \frac{b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1}{1 - C}, \quad u = \frac{u^1 u^2}{1 - C}, \tag{12}$$

where $C = \sum_{i \neq j} b_k^i b_k^j$ is designed to measure conflicts on the k^{th} class between $\mathcal{S}^i$ and $\mathcal{S}^j$. Intuitively, the fusion mechanism make the joint opinion $\mathcal{S}$ work like logic 'OR': Only when all the features are uncertain (large u), its joint confidence is relatively low (small b_k); If any feature was with high confidence (large b_k), and then the output decision would depend on this block. Accordingly, we can obtain a robust feature of higher confidence fused by various blocks in a redundant manner, which thus defends against attack.

Shown as Figure 2, based on the proposed fusion strategy, we can easily combine with evidential features from last $M \geq 2$ blocks in the given backbone for robustness enhancement:

$$\mathcal{S} = \mathcal{S}^1 \oplus \cdots \oplus \mathcal{S}^m \oplus \cdots \oplus \mathcal{S}^M. \tag{13}$$

According to Eq. 7, the corresponding Dirichlet's parameter of joint opinion on k^{th} category and its joint evidence could be figured out. Hence, we can follow Eq. 9, Eq. 10 and Eq. 11 to derive the loss of multi-feature evidential fusion:

$$\mathcal{L}_{mef} = \mathcal{L}_e(\boldsymbol{\alpha}_i, \mathbf{y}_i) + \lambda_t \mathcal{L}_{KL}(\boldsymbol{\alpha}_i, \mathbf{y}_i). \tag{14}$$

To obtain a better robust representation, we need to collect evidence not only from individual blocks but also from the fused block to form a more reliable opinion. Additionally, the classifier should also learn discriminant representation for different classes, and the original cross-entropy loss can

TABLE 1
Robustness (accuracy (%)) comparisons of different adversarial training strategies with normal, FSR, and our EMFF method against various white-box attacks on ResNet-18. The best results are in **bold**, and the performance of FPCM [25] is our reproduced results.

ResNet-18	CIFAR-10						SVHN					
Method	Natural	FGSM	PGD-20	PGD-100	C&W	Ens	Natural	FGSM	PGD-20	PGD-100	C&W	Ens
TAAT [31]	**83.46**	59.81	53.82	53.60	51.34	50.05	91.41	75.48	58.57	57.98	54.92	54.08
TAAT + **EMFF**	82.33	**60.34**	**54.55**	**54.24**	**52.06**	**50.88**	**92.02**	**76.04**	**59.13**	**58.69**	**55.73**	**54.78**
AT [12]	**85.02**	56.21	48.22	46.37	47.38	45.51	91.21	55.55	40.85	37.54	40.61	37.41
AT + FSR [21]	81.46	58.07	52.47	51.02	49.44	48.34	91.28	60.46	43.94	39.01	43.22	38.81
AT + **EMFF**	81.14	**58.50**	**53.59**	**52.35**	**51.98**	**50.04**	**91.74**	**65.22**	**51.87**	**47.56**	**50.10**	**46.51**
TRADES [46]	86.31	57.21	50.74	49.44	48.66	47.89	90.99	61.31	47.12	43.55	45.48	42.99
TRADES + FSR [21]	84.49	58.29	52.27	51.28	49.92	49.28	91.39	68.85	51.49	47.50	46.70	46.17
TRADES + FPCM [25]	82.61	58.94	52.78	51.98	50.37	49.51	89.98	65.77	53.81	50.75	50.77	48.94
TRADES + **EMFF**	**86.71**	**60.51**	**53.80**	**52.62**	**51.24**	**50.75**	**92.61**	**70.27**	**57.02**	**52.16**	**54.18**	**49.75**
MART [47]	82.73	56.65	50.88	49.15	47.21	45.98	90.50	58.21	43.61	40.43	42.20	40.07
MART + FSR [21]	**83.28**	59.55	54.80	53.69	48.98	48.36	89.87	61.06	46.51	42.94	43.89	42.40
MART + FPCM [25]	81.04	58.93	54.47	53.40	49.80	48.86	91.00	61.58	46.48	42.72	43.70	41.73
MART + **EMFF**	80.71	**59.78**	**55.52**	**54.32**	**51.53**	**50.30**	**94.07**	**69.83**	**53.80**	**47.87**	**49.89**	**46.70**

TABLE 2
Robustness (accuracy (%)) comparisons of different adversarial training strategies with normal, FSR, and our EMFF method against various white-box attacks on WideResNet-34-10. The best results are in **bold**, and we reproduced the performance of FPCM [25].

WideResNet-34-10	CIFAR-10						SVHN					
Method	Natural	FGSM	PGD-20	PGD-100	C&W	Ens	Natural	FGSM	PGD-20	PGD-100	C&W	Ens
AT [12]	**87.49**	59.47	50.72	48.75	50.42	48.52	91.20	61.27	47.50	43.26	44.94	42.61
AT + FSR [21]	87.02	61.40	53.78	52.04	52.35	50.36	92.73	68.34	49.79	43.58	47.34	43.22
AT + **EMFF**	85.98	**63.40**	**56.37**	**54.27**	**54.91**	**52.99**	**94.42**	**71.43**	**58.21**	**53.07**	**55.78**	**52.60**
TRADES [46]	86.06	60.78	51.77	49.66	51.34	49.27	92.01	65.13	49.54	44.17	47.39	43.72
TRADES + FSR [21]	86.88	62.97	54.37	51.98	53.19	51.34	**92.75**	72.83	61.87	54.72	55.26	51.42
TRADES + FPCM [25]	85.29	61.00	53.67	51.95	52.76	51.03	91.13	67.52	52.42	47.90	49.93	47.02
TRADES + **EMFF**	**89.83**	**65.71**	**58.06**	**56.05**	**56.09**	**53.72**	92.24	**76.14**	**63.71**	**59.30**	**62.32**	**55.77**
MART [47]	85.81	61.22	52.49	49.88	49.67	48.81	91.98	69.18	50.78	45.74	47.77	45.00
MART + FSR [21]	**86.21**	62.61	54.23	52.00	51.25	50.10	92.46	73.56	56.54	47.48	53.77	46.85
MART + FPCM [25]	85.29	61.84	55.25	53.39	51.95	50.95	92.03	65.65	49.39	44.43	45.47	43.40
MART + **EMFF**	86.06	**64.32**	**58.26**	**56.51**	**54.64**	**53.53**	**94.69**	**83.14**	**61.99**	**57.33**	**56.84**	**54.26**

not be discarded. As a consequence, the overall loss for our EMFF plugin is derived as:

$$\mathcal{L}_{overall} = \mathcal{L}_{cls} + \sum_{m=1}^{M} \mathcal{L}_{efl}^{m} + \mathcal{L}_{mef}. \tag{15}$$

4 Experiments

4.1 Experimental Setups

Datasets and Evaluations. We evaluate our effectiveness and superiority for the adversarial robustness of different backbones on several benchmark datasets, including CIFAR-10 [79], CIFAR-100, SVHN [80], Imagenette [81], and Tiny-ImageNet [82]. Specifically, the mainstream CNNs baselines including ResNet-18 [5] and WideResNet-34-10 [77], and the recent popular vision transformer architectures containing DeiT [4] and ConViT [26] are utilized in our experiments for robustness evaluation.

Following the recent works [21], [25], [83], we also combine our novel EMFF method to PGD adversarial training (AT) [12], or other common variants including TRADES [46] and MART [47], to demonstrate our generalization and practicability on robustness enhancement. For evaluation, we also take FGSM [8], PGD [12] (PGD-20, PGD-100) and C&W [38] (C&W-30 for CNNs and C&W-20 for Transformers), and AutoAttack [37] as adversarial attacks. All PGD optimization is with step size of $\Delta/10$. To better compare the robustness of different defense strategies, we also follow FSR [21] to report the average per-example ensemble (Ens) robustness of the models including FGSM, PGD-20, PGD-100, and C&W-30 for CNNs robustness evaluation.

TABLE 3
Robustness (accuracy (%)) of adversarial training strategies (A: AT, T: TRADES, M: MART) with different plug-ins against diverse white-box attacks in ResNet-18 on Tiny-ImageNet [82] dataset.

Method	Natural	FGSM	PGD-20	PGD-100	C&W	Ens
A [12]	51.13	22.54	18.69	17.87	17.83	16.34
A + FSR [21]	51.77	24.19	20.95	20.06	19.32	18.02
A + **EMFF**	**52.10**	**25.02**	**21.43**	**20.92**	**20.14**	**18.83**
T [46]	**50.41**	23.79	21.16	20.72	17.24	17.02
T + FSR [21]	49.53	24.87	23.22	23.09	19.22	19.04
T + **EMFF**	49.82	**25.33**	**23.71**	**23.52**	**20.02**	**19.74**
M [47]	**46.21**	23.84	21.75	21.35	18.34	17.71
M + FSR [21]	46.02	26.02	24.05	23.82	20.63	20.24
M + **EMFF**	46.14	**26.32**	**24.51**	**24.31**	**21.04**	**20.82**

Implementation Details. For all experiments, we follow PGD-10 [12] to constrain the perturbation bound within $\Delta = 8/255$ under ℓ_∞-norm to craft adversarial examples. For

TABLE 4
Robustness (accuracy (%)) comparisons of different adversarial training strategies with normal, FSR, and our EMFF method against various white-box attacks on CIFAR-100. The best results are in **bold**, and we reproduced the performance of WideResNet-34-10 [77].

CIFAR-100	ResNet-18						WideResNet-34-10					
Method	Natural	FGSM	PGD-20	PGD-100	C&W	Ens	Natural	FGSM	PGD-20	PGD-100	C&W	Ens
TAAT [31]	**56.99**	33.11	30.56	30.32	26.76	26.01	**60.78**	34.23	30.99	30.68	27.57	27.31
TAAT + **EMFF**	**56.78**	**34.57**	**32.15**	**32.05**	**28.11**	**27.26**	60.55	**35.08**	**32.45**	**32.28**	**28.62**	**28.38**
AT [12]	**59.25**	28.80	24.39	23.43	23.92	22.46	62.18	35.32	32.50	32.34	29.14	28.87
AT + FSR [21]	58.23	29.58	25.33	24.30	24.54	22.95	62.03	35.48	32.67	32.58	29.68	29.32
AT + **EMFF**	52.77	**32.27**	**29.84**	**29.29**	**30.21**	**27.84**	**63.02**	**36.05**	**33.90**	**33.17**	**30.26**	**30.14**
TRADES [46]	**61.87**	30.77	26.37	25.76	24.08	23.45	62.05	35.02	31.43	31.27	29.32	28.88
TRADES + FSR [21]	57.27	31.66	27.70	27.27	24.82	24.40	62.23	35.34	31.92	31.88	29.91	29.54
TRADES + **EMFF**	59.53	**34.06**	**31.30**	**30.82**	**30.26**	**28.70**	**62.78**	**35.78**	**32.31**	**32.12**	**32.04**	**30.22**
MART [47]	**57.13**	31.32	27.40	26.80	25.24	24.42	58.89	32.18	30.19	30.05	28.58	28.33
MART + FSR [21]	56.51	32.08	27.90	27.28	25.91	24.98	58.93	32.56	31.08	30.62	29.22	28.95
MART + **EMFF**	54.85	**33.10**	**31.00**	**30.47**	**30.63**	**28.53**	**59.38**	**33.22**	**31.82**	**31.27**	**30.46**	**29.32**

TABLE 5
Robustness (accuracy (%)) comparisons of different adversarial training strategies with normal, ATMT [83] and our EMFF method against various white-box attacks on Transformer backbones. The best results are in **bold**. AA indicates AutoAttack [37].

Trans-formers	Method	CIFAR-10					Imagenette				
		Natural	C&W-20	PGD-20	PGD-100	AA	Natural	C&W-20	PGD-20	PGD-100	AA
DeiT-Tiny [4]	TRADES [46]	78.70	46.78	49.63	49.58	46.25	88.00	63.00	62.60	62.40	61.20
	TRADES + ATMT [83]	**80.24**	47.60	51.02	50.97	47.02	**89.00**	63.20	64.40	64.00	61.80
	TRADES + **EMFF**	77.35	**63.90**	**61.61**	**61.43**	**75.74**	88.20	**75.00**	**70.00**	**69.80**	**85.40**
	MART [47]	71.70	45.95	49.52	49.37	44.34	80.40	55.40	56.20	56.00	52.60
	MART + ATMT [83]	**74.89**	47.60	51.18	51.16	45.97	**86.40**	61.80	63.40	63.20	62.20
	MART + **EMFF**	72.35	**64.80**	**63.31**	**63.31**	**70.25**	85.20	**73.00**	**71.60**	**71.60**	**82.20**
ConViT-Tiny [26]	TRADES [46]	77.70	45.09	48.71	48.63	44.65	83.80	58.80	60.40	60.20	57.80
	TRADES + ATMT [83]	**80.02**	47.33	50.10	50.08	46.75	**89.20**	66.20	65.60	65.00	64.60
	TRADES + **EMFF**	77.95	**62.17**	**59.80**	**59.84**	**77.79**	86.80	**74.20**	**70.40**	**70.00**	**84.60**
	MART [47]	63.68	38.97	42.80	42.77	37.62	61.80	36.40	41.80	41.60	35.40
	MART + ATMT [83]	**74.89**	47.60	51.18	51.16	45.97	**88.00**	65.00	64.40	64.40	63.40
	MART + **EMFF**	73.14	**60.31**	**59.81**	**59.90**	**69.34**	81.20	**72.20**	**71.80**	**71.80**	**80.40**

optimal performance on different backbones, we empirically incorporate the features from last four blocks within the architecture ($M = 4$), to perform our evidential fusion.

For CNNs, we train them under PGD-10 attack for 100 epochs with the iterative step size $\Delta/4$ for CIFAR-10 and $\Delta/8$ for SVHN respectively. The SGD optimizer (momentum=0.9, weight decay=5×10^{-4}) with an initial learning rate of 0.1, which was then gradually reduced by 0.1 at the 75^{th} epoch and the 90^{th} epoch for both CIFAR-10 and SVHN, was adopted during the training process.

For Vision Transformers (ViTs), we first obtain pre-trained models on ImageNet, and continue to train them under PGD-10 attack with 40 epochs, where the iterative step size is set to $\Delta/4$. The optimizer and initial learning rate are consistent with that of CNNs, and the differences are that the weight decay is 1×10^{-4} and the learning rate gradually decreases at both the 36^{th} and 38^{th} epochs.

4.2 Defense against White-box Attacks

Classical Backbones. For fair comparisons, we follow the latest works [21], [25] to combine our novel EMFF method with three different adversarial training strategies (AT, TRADES, MART) and the latest topoloy information aligning method TAAT [31] for the robustness enhancement of the classical backbones. Specifically, we report the performance for ResNet-18 [5] and WideResNet-34-10 [77] in CIFAR-10 [79], SVHN [80], Tiny-ImageNet [82], and CIFAR-100 in Table 1, Table 2, Table 3, and Table 4, respectively. Analyzing these results, we can observe several important findings:

1) Smilar to the easy combination of the existing works [21], [25], our EMFF method can also be seamlessly integrated as a plug-in with various adversarial training strategies into classical architectures to consistently enhance the robustness of the model under all the individual attacks and ensemble-based attack. This proves the practicability and effectiveness of our novel multi-evidential fusion.

2) Specifically, the robustness of ResNet-18 under the attack of FGSM on SVHN can be enhanced by our method with a maximum margin of **11.62%** when combined with AT, and a similar uptrend about **9%** occurs on the ensemble attack. Under other attacks, our robustness promotions present slighter increases but at least around **2%**. Note that these comparisons are made with only the use of vanilla adversarial training techniques. Even compared to the enhanced versions of FSR [21], FPCM [25], and TAAT [31], our robustness improvement can also outperform them under all attacks.

3) An important observation consistent with FSR [21] and FPCM [25] is that our classification accuracy occasionally

drops on natural images. We attribute this case to two sides: on the one hand, conventional DNNs, which are based on classification probability, could be overconfident [27], and thus evidential deep learning is introduced to solve this problem; On the other hand, our KL regularization attempts to shrink the evidence for incorrect classification caused by perturbations, while there are no such adversarial features in natural images. Consequently, this effort may cause to potential loss of some distinctive knowledge. Nevertheless, the occasional drop is only within a slight amount and may depend on the dataset or the picked backbone, which does not reduce our practicability.

Transformers. To further demonstrate our effectiveness and generalization, we also integrate our easy plug-in into different transformer-based backbones, including ConViT [26] and DeiT [4] to make comparisons with ATMT [83] on robustness. The experimental results on CIFAR-10 and Imagenette, Tiny-ImageNet of various adversarial training strategies with different backbones are reported in Table 5 and Table 10, respectively.

It is clear that we can also consistently improve the robustness of the transformer-based architectures with remarkable margins (even up to **45%** occurring at ConViT under the attack of AutoAttack [37] (AA) on Imagenette for MART) under different attacks. Even with ATMT [83], an adversarial training strategy dedicated to Transformer architecture, we still maintain the significant advantages on robustness from **5%** to **30%**. These clear leads, partially attributing to the strengths of Transformers not only in representation learning but also in evidence collection, reaffirm the superiority of our novel EMFF plug-in. Similarly, the classification accuracy on clean images also shows an occasional drop for the same reasons, but we don't think these slight margins of decline will neutralize our clear gains on robustness promotion against adversarial attacks.

It is worth noting that our EMFF mehod focuses on eliminating the uncertainty of different blocks in various backbones, thus improving their robustness, rather than designing some dedicated and intricate strategies like [55], [56]. Therefore, our obvious effectiveness in boosting robustness, clear superiority to various adversarial training, easy integration into distinct backbones cannot be ignored, albeit our slight inferior performance to the latest works [54], [57]. More importantly, our method can still be easily integrated into VAD [54] and ReiT [57] for further improving the robustness when they publish their corresponding codes, which is confirmed by the clear boosts in Table 1, Table 2, Table 3, Table 4, Table 5, and Table 10.

4.3 Defense against Black-box Attacks

Although we believe that the performance under white-box attacks can better reflect the advantages and disadvantages of various defense strategies, we still verify our effectiveness and practicability in the scenario of black-box attacks.

Specifically, we select several recently published black-box attacks, including two transfer-based attacks of TI-FGSM [39] and DI-FGSM [40], a scale-invariant and nesterov iterative attack (SI-NI-FGSM) [44], and variance tuning (VMI-FGSM) [45] to craft adversarial examples based on a naturally trained ResNet-50 for our defense performance evaluation.

The results in Figure 3(a) once again highlight the superiority, effectiveness, and easy applicability of our multi-evidential fusion method in improving robustness against diverse strong black-box attacks.

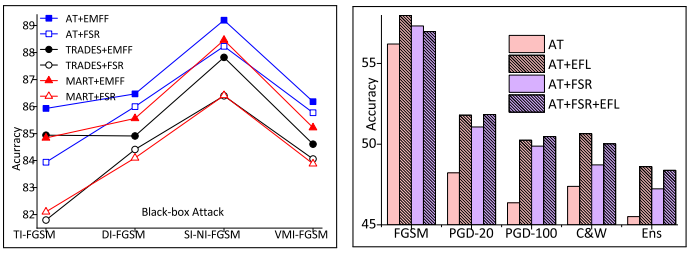

Fig. 3. (a) Robustness comparisons among different defense strategies under diverse black-box attacks on SVHN using ResNet-18. (b) Ablation studies of EFL with ResNet-18 [5] on CIFAR-10 [79], the performance of AT+FSR is our reproduced.

4.4 Ablation Study

To prove our effectiveness, we take AT [12], ResNet-18 [5], and CIFAR-10 [79] as the adversarial training technique, baseline backbone, and dataset respectively to conduct ablation studies. Next, more analysis can be presented gradually.

Effectiveness of Evidential Feature Learning. Firstly, we should demonstrate the effectiveness of introducing evidential feature learning (EFL), which is an important cornerstone of our approach. To this end, we set up two kinds of experiments: (1) directly introduce EFL into the final feature of ResNet-18; (2) EFL is integrated on the robust and non-robust features disentangled by FSR [21].

TABLE 6
Ablation studies of multi-block denoised feature fusion.

Method	FGSM	PGD-20	PGD-100	C&W	Ens
AT + FD_2	57.77	50.20	48.30	49.39	47.12
AT + FD_3	57.82	50.36	48.58	49.33	47.27
AT + FD_4	**58.24**	**50.78**	**49.09**	**49.79**	**47.77**
AT + FD_5	57.64	50.35	48.95	49.61	47.58

The detailed results presented in Figure 3(b) could clearly tell us that both the vanilla adversarial training and another recent strong defense strategy of FSR [21] by re-calibrating non-robust features to obtain more robust cues can further help the backbone to better defend adversarial attacks after introducing our EFL module.

Effectiveness of Multi-Block Fusion. As aforementioned, we also claim that with only the adoption of the feature from the final block of the backbone, the classifier is easily affected by some local prominent features caused by adversarial perturbations to reversely alter its decisions and thus output error classifications. Therefore, a comprehensive understanding of features, that is, the effectiveness of multi-block feature fusion, is also our foundation.

To do so, we integrate feature denoising (FD) [14] into ResNet-18, and verify the effectiveness of different multi-block fusion with AT. Specifically, FD_M indicates that we employ the simple concatenation of the denoised features output by last M blocks as the final feature. The results in Table 6 confirm our argument that multi-block feature fusion can improve robustness.

Effectiveness of Multi-feature Evidential Fusion. Based on the analysis of the previous two parts, we continue to verify the effectiveness of multi-feature evidential fusion, which is an essential step in our method. The results in Table 7 firstly emphasize that our multi-feature evidential fusion can improve the robustness of the backbone with a larger margin than a sole evidential feature learning. Secondly, we can also observe that multi-feature evidential fusion with different blocks can promote robustness with different margins, which re-emphasizes our superiority and effectiveness.

TABLE 7
Ablation studies of multi-feature evidential fusion.

Method	FGSM	PGD-20	PGD-100	C&W	Ens
AT + EMFF_2	58.07	52.04	50.37	50.71	48.96
AT + EMFF_3	58.47	53.19	51.85	51.62	49.95
AT + EMFF_4	**58.50**	**53.59**	**52.35**	**51.98**	**50.04**
AT + EMFF_5	58.28	53.37	52.12	51.87	50.02
AT + EFL	57.98	51.80	50.24	50.65	48.61

How many blocks of evidential feature fusion can maximize the robustness of the selected baseline? The experimental results in both Table 6 and Table 7 reach an agreement that when $M = 4$, our novel EMFF method can help the baseline to obtain optimal robustness. However, another important observation must be made clear, that is, the adoption of more or shallower block features can neutralize or lower the fusion performance. This case may be ascribed to that the features learned by shallow blocks are still elementary, and thus the feature fusion at different levels of semantic abstraction goes in the opposite direction of evidence collection, which leads to the decline of robustness.

4.5 Further Analysis

Visualization. Additionally, we also visualize the Top-4 classification performance of different blocks and the final output before and after our novel EMFF integration. Practically, we present 4 examples on ResNet-18 and ConViT in Figure 4(a) and Figure 4(b), respectively.

From the visualizations, we can make two conclusions: 1) Each block in a backbone can indeed capture somewhat ground-truth features. If the classifier only with the perturbed output from the final block, it is thus easily induced by contaminated features to make incorrect predictions. This observation is illustrated in our motivation. 2) After integrating our multi-evidential fusion, almost all blocks from both ResNet-18 and ConViT can improve the classification of ground truth and the final output is also significantly improved, which helps the backbones prevent the induction of adversarial perturbations and thus enhance their robustness.

Failure Cases Discussion. To further illuminate our proposed EMFF method, a discussion of some failure cases is warranted. These failure cases are illustrated in Figure 5. The failure cases can be divided into two types: 1) the output of some block is still dominated by adversarial perturbation but the final output of the backbone is correct. This indirect confirmation lends further credence to the efficacy of our evidential fusion based on Dempster-Shamfer theory, illustrated in the upper part of Figure 5. 2) The uncertainty measurement of our EMFF method only takes into account the consistency of prediction results but overlooks the conflicting aspects, which may lead to counter-intuitive results as our method may produce excessively low uncertainty to achieve compromising robust results, as displayed in the lower part of Figure 5. These observations are employed to facilitate further improvements in the future, especially some uncertainty-aware belief integration strategies.

TABLE 8
Computation cost comparisons (Params and MACs) between a baseline model and the version with our EMFF plugin.

Method	*ResNet-18*		*WideResNet-34*	
	Params (M)	MACs (G)	Params (M)	MACs (G)
Vanilla	11.174	27.33589	46.160	326.85129
+ EMFF	11.182	27.33590	46.172	326.85130
	DeiT		*ConViT*	
Vanilla	5.488	1.07844	5.481	1.07861
+ EMFF	5.494	1.07845	5.490	1.07862

Computation Analysis. To verify the practicability of our novel EMFF plugin, we also provide the computation analysis between our method and several baselines adopted in this work in terms of the number of parameters (Params (M)) and theoretical amount of multiply-add operations (MACs (G)). Specifically, we use the OpCounter tool[1] to obtain these results on the image size of Imagenette. Displayed as Table 8, we only take a slightly larger number of parameters and almost the same computations yet to promote the robustness with a remarkable margin. Additionally, the training time of one epoch for the vanilla model and its improved version with our EMFF plugin is also very close. These good characteristics guarantee the wide applicability of our novel method.

Furthermore, we also provide the extra cost caused by our EMFF method based on ResNet-18 in CIFAR-100. Figure 6 restates our advantages of negligible computation yet promoting the robustness with a remarkable margin.

Difference from FPN-like methods: The readers may think that our EMFF plugin is the same as Feature Pyramid Networks (FPN) [85], [86], [87]. In fact, we are fundamentally:

1) Different target. The FPN-like focuses on small-scale object detection, while we enhance adversarial robustness.

2) Different insight. The former is a multi-scale fusion at a fixed view, while we dynamically aggregate opinions from various agents (blocks) at an evidence level.

3) Different performance. We also plant our EMFF into a FPN-like method (FPD [84] based on ResNet-50) on adversarial robustness to verify our benefits against white-box attack on SVHN. The specific improvements of our EMFF method over FPD [84] can be referred to Figure 7. The significant improvements once again present our effectiveness, superiority and generalization.

Theoretical Justification: Our work mainly introduces Dempster-Shafer theory to comprehensively fuse evidential features from diverse blocks within the architecture for robust enhancement. The main advantages over simple methods like weighted average or major voting can be summarized as follows: 1) The softmax function employed in these methods only provides a point estimate for the class probabilities of an adversarial example, without providing the associated uncertainty. In addition, our multinomial

1. https://github.com/Lyken17/pytorch-OpCounter

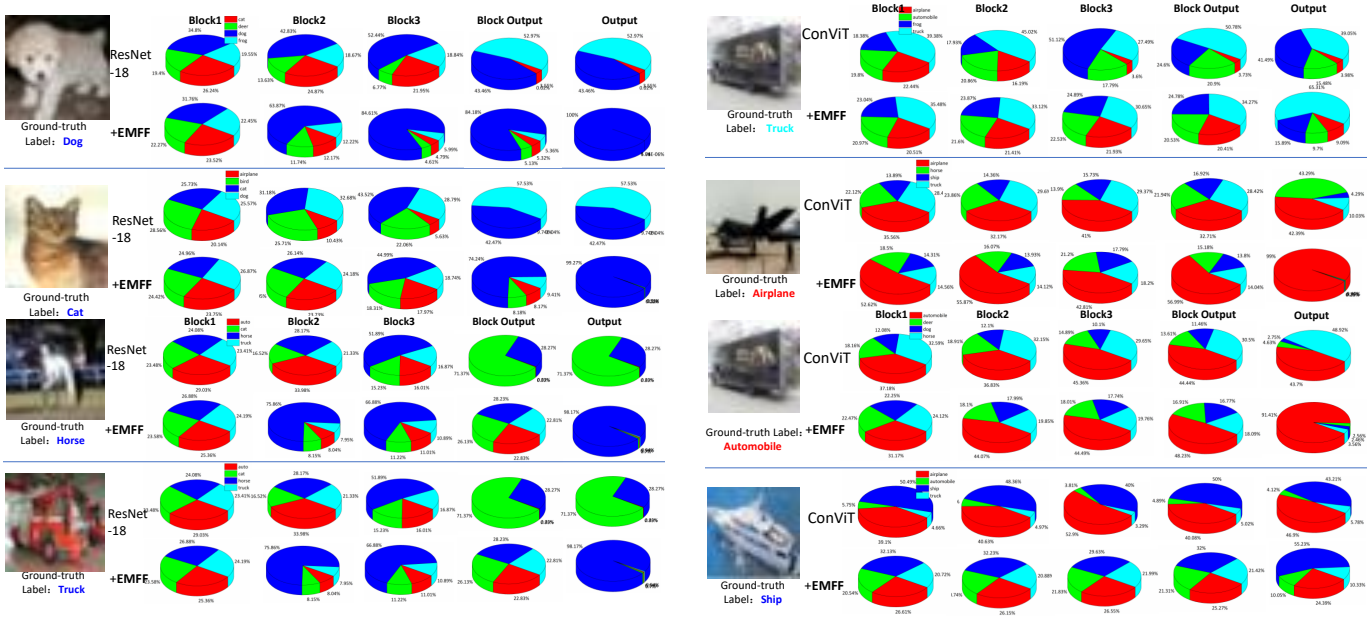

Fig. 4. Visualization. (a):Top-4 classification on CIFAR-10 of different blocks from ResNet-18 w/ and w/o our EMFF; (b) The same case for ConViT.

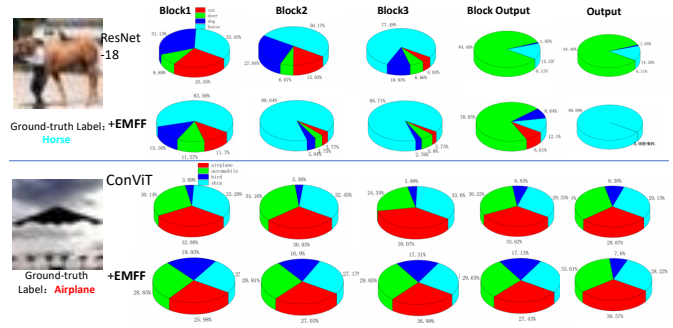

Fig. 5. Failure case visualization on CIFAR10 for (top) ResNet-18 w/ and w/o our EMFF; (down) The same case for ConViT.

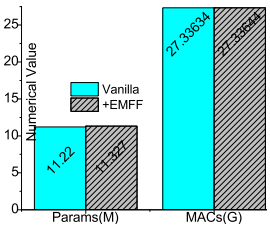

Fig. 6. Additional cost of our method on CIFAR-100.

Fig. 7. Boosting FPD [84] by our method on SVHN.

Method	FGSM	PGD
FPD	33.0	30.0
+ **EMFF**	**35.3**	**32.1**
FPD_{PGD}	39.0	35.0
+ **EMFF**	**40.8**	**36.5**

opinions or equivalently Dirichlet distributions can be used to model a probability distribution for the class probabilities. Consequently, our design represents the predictions of the learner as a distribution over possible softmax outputs, rather than the point estimate of a softmax output. Finally, based on the uncertainty estimation and the Dempster-Shafer theory, we can easily quantify the feature confidence and aggregate various opinions to defend against adversarial attacks.

2) The Dempster–Shafer Theory of Evidence (DST) is a generalization of the Bayesian theory to subjective probabilities [68], [69] by assigning belief masses to subsets of a discernment frame. Orthogonally to Bayesian neural nets that indirectly infer prediction uncertainty through weight uncertainties, our EMFF explicitly models the same using the theory of subjective logic by placing a Dirichlet distribution on the class probabilities.

To provide a basis for comparison with the aforementioned methods, including major voting, weighted average and Bayesian uncertainty estimation [88], we have selected ResNet-18 on CIFAR10 as an experimental example. The results presented in Table 9 highlight our superiority and provide a theoretical justification for our method.

TABLE 9
Our theoretical justification with different methods based on ResNet-18 on CIFAR10.

Method	FGSM	PGD-20	PGD-100	C&W	Ens
AT	56.21	48.22	46.37	47.38	45.51
AT + Voting	57.67	51.62	49.36	50.02	48.82
AT + Weighted Avg.	57.94	51.54	49.99	50.11	48.45
AT + BUE [88]	58.02	51.83	50.25	50.48	49.02
AT + **EMFF**	**58.50**	**53.59**	**52.35**	**51.98**	**50.04**

Integration with More Recent Methods. Follow ADML [89], we still conduct verification on CIFAR-100 and Tiny-ImageNet by integrating our EMFF into ResNet-18 and DeiT-S [4] based on AWP and AT, and then select three attacks only in ADML to evaluate the performance. The results in Table 10 re-confirm our superiority and generalization.

Note that we can plug our EMFF in ADML [89] + AWP [90] to further boost the performance, albeit only with a slight gain compared to directly combined with AWP [90], which also shows our generalization and easy integration.

Moreover, we utilize the recently developed method B-MTARD [91] as a case study to validate our seamless integration on CIFAR10. After plugin with our EMFF designn, the selected baseline B-MTARD demonstrates a performance enhancement with expected margins, thereby further substantiating our efficacy and easy application. The comprehensive outcomes are delineated in Table 11.

TABLE 10
Comparison against various defenses whether to the inclusion of our EMFF method on CIFAR-100 and Tiny-ImageNet.

Method	CIFAR-100/ ResNet-18			Tiny-ImageNet/ DeiT-S		
	AP	DLR	AA	AP	DLR	AA
AT [12]	25.3	25.3	25.1	33.1	33.0	33.0
+ ADML [89]	28.4	28.2	27.8	40.8	39.5	39.4
+ **EMFF**	**32.2**	**32.2**	**34.8**	**41.3**	**40.3**	**40.2**
AWP [90]	30.0	30.0	29.8	39.0	39.1	39.0
+ **EMFF**	**35.5**	**35.4**	**35.3**	**40.2**	**39.8**	**39.8**
AWP + ADML [89]	31.1	30.5	30.3	45.7	42.8	42.8
+ **EMFF**	**35.6**	**35.6**	**35.4**	**45.2**	**43.1**	**43.3**

TABLE 11
Integrating our EMFF method into B-MTARD [91] on CIFAR10.

Method	FGSM	PGD_{sat}	PGD_{trades}	C&W	AA
Natural	56.59	47.29	47.29	47.29	47.29
SAT [12]	69.90	65.08	66.16	65.09	63.85
TRADES [46]	70.68	67.68	68.42	66.62	66.02
B-MTARD [91]	74.81	69.94	71.30	69.04	67.82
+ **EMFF**	**75.27**	**70.38**	**71.62**	**70.29**	**69.14**

TABLE 12
Comparisons among DE [92], FSR [21], and our EMFF on CIFAR100.

Method	FGSM	PGD-20	PGD-100	C&W	Ens
AT	28.80	24.39	23.43	23.92	22.46
AT + FSR	29.58	25.33	24.30	24.54	22.95
AT + DE_4	31.65	26.75	25.53	23.49	22.88
AT + **EMFF**$_4$	**32.27**	**29.84**	**29.29**	**30.21**	**27.84**

Comparisons with uncertainty methods: We take a non-Bayesian method DE [92] as an example to verify the effectiveness of uncertainty measuring on adversarial robustness in CIFAR-100, and then better demonstrate our superiority. The results are displayed in Table 12. When the backbone is fused with 4 blocks configured by DE [92] or Bayesian uncertainty estimation [88] reported as Table 9, its performance can be greatly improved compared with the vanilla version, and even exceed that of the FSR method [21]. However, our EMFF method can win it easily.

Comparisons with randomized smoothing methods: To scale up to practical networks, randomized smoothing [93], [94] has been proposed as a probabilistically certified defense. Therefore, we only follow the settings in [94] to make fair comparisons on CIFAR10 (ResNet-110) and Tiny-ImageNet (ResNet-50). The results are reported in Table 13, which once again highlight the effectiveness, superiority, and generalization of our novel EMFF plugin. It is noteworthy that the efficacy of the EMFF plugin is further enhanced by its integration with modern techniques like random smoothing and certified robustness, underscoring its potential for enhancing the defense capabilities against adversarial attacks.

TABLE 13
Certificed top-1 accuracy comparisons with randomized smoothing [93], [94] on CIFAR10 and Tiny-ImageNet at various l_2 radii.

Methods \ l_2 radius	Tiny-ImageNet			CIFAR10		
	0.5	1.0	1.5	0.5	1.0	1.5
Cohen et al. [93]	49.23	37.56	30.11	43.00	22.00	13.00
+ **EMFF**	**50.85**	**39.42**	**33.16**	**45.28**	**25.61**	**17.72**
SmoothADV [94]	56.12	45.41	38.72	58.00	38.00	29.00
+ **EMFF**	**57.11**	**48.32**	**41.05**	**58.27**	**40.45**	**32.17**

5 Limitation and Discussion

Our EMFF method overcomes the limitations of existing adversarial defenses by easily integrating our approach into existing and future defense strategies across various architectures with evidential fusion and minimal cost. However, our EMFF method also comes with limitations:

(1) An attack specifically targeted the evidence collector or the multi-block fusion process will weaken our defense. In view of this limitation, we can first reasonably assume that only one structure can be attacked at a time, and then in the subsection 4.4 of Ablation Study, we have bypassed the attacked structure to simulate this attack. The results reported in Figure 3(b), Table 6, and Table 7 jointly demonstrate that the remaining structure can still improve the defense capability to a certain extent. Secondly, because the evidence collector network adopted in this method is very simple, it only consists of full connection layers and a non-negative activation function. Therefore, some preprocessing-based defenses such as JPEG compression [95] and pixel deflection [96] can be employed to strengthen the defense together. Thirdly, if it's an attack on multi-block fusion, we can design other strategies in the future to further improve the defense capability. For example, as our fusion strategy only average the uncertainty and beliefs, displayed as Eq 12, a uncertainty-weighted belief fusion can be introduced to eliminate blocks that are too uncertain (they may have been attacked) and then fuse the safe blocks. In addition, some metrics about attack intensity can be proposed to quantitatively evaluate the robustness of different blocks, which can also help to remove the attacked block.

(2) We do not explore the trade-off between robustness and accuracy, which may be another fruitful research topic focused by [91], [97] or potentially solved by randomized smoothing [93], [94]. Moreover, our EMFF method can be seamlessly integrated with these approaches to enhance robustness when necessary.

Note that, the mentioned research directions are beyond this work, and we gratefully thanks the valuable comments of the anonymous reviewers and encourage the readers to further strengthen adversarial robustness in theory and practice by addressing the aforementioned limitations based on our work.

6 Conclusion

In this paper, we have proposed a novel Evidence-based Multi-Feature Fusion (EMFF) method to prevent DNNs from being cheated by the contaminated features only from a single block view. To this end, our EMFF method first introduced evidential deep learning to produce a stable and reasonable uncertainty estimation of features from different blocks within an architecture. It then elegantly integrated multi-block features at an evidence level based on Demster-Shapfer's theory for trusted prediction instead of only adopting the feature of the final block as done previously. With the help of our novel fusion mechanism, the robustness and classification reliability of multiple backbones based on traditional CNNs and recent vision transformers could be remarkably promoted only with a few extra parameters and almost the same cost. Experimentally, we have verified the superiority, effectiveness, and practicability of our EMFF

method under the scenarios of white-box and black-box attacks when applied to different adversarial training strategies across several widely used datasets, including CIFAR-10, CIFAR-100, SVHN, Imagenette, and Tiny-ImageNet.

REFERENCES

[1] E. Vahdani and Y. Tian, "Deep learning-based action detection in untrimmed videos: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4302–4320, 2023.

[2] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[3] Z. Wang, Z. Gao, Y. Yang, G. Wang, C. Jiao, and H. T. Shen, "Geometric matching for cross-modal retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024.

[4] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 10 347–10 357.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] B. Tang, Z. Wang, Y. Bin, Q. Dou, Y. Yang, and H. T. Shen, "Ensemble diversity facilitates adversarial transferability," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 24 377–24 386.

[7] Z. Wang, B. Tang, Y. Bin, L. Zhu, G. Wang, and Y. Yang, "Shapley ensemble adversarial attack," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[8] G. Ian J, S. Jonathon, and S. Christian, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–13.

[9] J. Peck, B. Goossens, and Y. Saeys, "An introduction to adversarially robust deep learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2071–2090, 2024.

[10] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *International Conference on Learning Representations (ICLR)*, 2017, pp. 1–14.

[11] C. Zhang, P. Benz, A. Karjauv, and I. S. Kweon, "Data-free universal adversarial perturbation and black-box attack," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 7848–7857.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018, pp. 1–14.

[13] ——, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018, pp. 1–15.

[14] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 501–509.

[15] X. Jia, X. Wei, X. Cao, and H. Foroosh, "Comdefend: An efficient image compression model to defend adversarial examples," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6077–6085.

[16] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 501–509.

[17] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 51–59.

[18] P. Benz, C. Zhang, and I. S. Kweon, "Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7798–7807.

[19] H. Wang, A. Zhang, S. Zheng, X. Shi, M. Li, and Z. Wang, "Removing batch normalization boosts adversarial training," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2022, pp. 23 433–23 445.

[20] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[21] W. J. Kim, Y. Cho, J. Jung, and S.-E. Yoon, "Feature separation and recalibration for adversarial robustness," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8183–8192.

[22] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1778–1787.

[23] Y. Bai, Y. Zeng, Y. Jiang, S.-T. Xia, X. Ma, and Y. Wang, "Improving adversarial robustness via channel-wise activation suppressing," in *International Conference on Learning Representations (ICLR)*, 2021, pp. 1–15.

[24] H. Yan, J. Zhang, G. Niu, J. Feng, V. Tan, and M. Sugiyama, "Cifs: Improving adversarial robustness of cnns via channel-wise importance-based feature selection," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 11 693–11 703.

[25] Q. Bu, D. Huang, and H. Cui, "Towards building more robust models with frequency bias," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4379–4388.

[26] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 2286–2296.

[27] J. Gao, M. Chen, and C. Xu, "Collecting cross-modal presence-absence evidence for weakly-supervised audio- visual event perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 827–18 836.

[28] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[29] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[30] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 1976, vol. 42.

[31] H. Kuang, H. Liu, X. Lin, and R. Ji, "Defense against adversarial attacks using topology aligning adversarial training," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3659–3673, 2024.

[32] K. Alexey, J. G. Ian, and B. Samy, "Adversarial examples in the physical world," in *International Conference on Learning Representations (ICLR), Workshop*, 2017, pp. 1–12.

[33] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9185–9193.

[34] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2020, pp. 1–15.

[35] Y. Shi, Y. Han, Q. Hu, Y. Yang, and Q. Tian, "Query-efficient black-box adversarial attack with customized iteration and sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2226–2245, 2023.

[36] Z. Zhao, Z. Li, F. Zhang, Z. Yang, S. Luo, T. Li, R. Zhang, and K. Ren, "Sage: Steering the adversarial generation of examples with accelerations," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 789–803, 2023.

[37] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 23–34.

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

[39] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4307–4316.

[40] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with

input diversity," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2725–2734.

[41] Z. Yuan, J. Zhang, Z. Jiang, L. Li, and S. Shan, "Adaptive perturbation for adversarial attack," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5663–5676, 2024.

[42] X. Luo, Z. Chen, M. Tao, and F. Yang, "Encrypted semantic communication using adversarial training for privacy preserving," *IEEE Communications Letters*, vol. 27, no. 6, pp. 23–34, 2023.

[43] R. Imam, I. Almakky, S. Alrashdi, B. Alrashdi, and M. Yaqub, "Seda: Self-ensembling vit with defensive distillation and adversarial training for robust chest x-rays classification," in *MICCAI Workshop on Domain Adaptation and Representation Transfer*, 2024.

[44] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2020, pp. 1–16.

[45] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1924–1933.

[46] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 7472–7482.

[47] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *International Conference on Learning Representations (ICLR)*, 2019.

[48] G. Jin, X. Yi, D. Wu, R. Mu, and X. Huang, "Randomized adversarial training via taylor expansion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 16 447–16 457.

[49] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao, "Las-at: Adversarial training with learnable attack strategy," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 13 388–13 398.

[50] G. Jin, X. Yi, W. Huang, S. Schewe, and X. Huang, "Enhancing adversarial training with second-order statistics of weights," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 252–15 262.

[51] M. Zhao, L. Zhang, Y. Kong, and B. Yin, "Fast adversarial training with smooth convergence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4697–4706.

[52] T. Li, Y. Wu, S. Chen, K. Fang, and X. Huang, "Subspace adversarial training," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 13 399–13 408.

[53] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "Stylized adversarial defense," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6403–6414, 2023.

[54] C. Zhao, S. Mei, B. Ni, S. Yuan, Z. Yu, and J. Wang, "Variational adversarial defense: A bayes perspective for adversarial training," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3047–3063, 2024.

[55] S. Jain and T. Dutta, "Towards understanding and improving adversarial robustness of vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 736–24 745.

[56] Z. Li, D. Yu, L. Wei, C. Jin, Y. Zhang, and S. Chan, "Soften to defend: Towards adversarial robustness via self-guided label refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 776–24 785.

[57] H. Gong, M. Dong, S. Ma, S. Camtepe, S. Nepal, and C. Xu, "Random entangled tokens for adversarially robust vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 554–24 563.

[58] G. S. Dhillon, K. Azizzadenesheli, Z. C. Lipton, J. Bernstein, J. Kossaifi, A. Khanna, and A. Anandkumar, "Stochastic activation pruning for robust adversarial defense," in *International Conference on Learning Representations (ICLR)*, 2018, pp. 12–22.

[59] D. Madaan, J. Shin, and S. J. Hwang, "Adversarial neural pruning with latent vulnerability suppression," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 6575–6585.

[60] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "Adversarial defense by restricting the hidden space of deep neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3384–3393.

[61] K. Xu, S. Liu, G. Zhang, M. Sun, P. Zhao, Q. Fan, C. Gan, and X. Lin, "Interpreting adversarial examples by activation promotion and suppression," *arXiv preprint arXiv:1904.02057*, 2019.

[62] C. Xiao, P. Zhong, and C. Zheng, "Enhancing adversarial defense by k-winners-take-all," in *ICLR*, 2020.

[63] Z. Allen-Zhu and Y. Li, "Feature purification: How adversarial training performs robust deep learning," in *IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 2022, pp. 977–988.

[64] M. Picot, F. Messina, M. Boudiaf, F. Labeau, I. B. Ayed, and P. Piantanida, "Adversarial robustness via fisher-rao regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2698–2710, 2023.

[65] H. Kuang, H. Liu, Y. Wu, and R. Ji, "Semantically consistent visual representation for adversarial robustness," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 5608–5622, 2023.

[66] S.-Y. Lo, P. Oza, and V. M. Patel, "Adversarially robust one-class novelty detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4167–4179, 2023.

[67] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330.

[68] L. Liu and R. R. Yager, *Classic Works of the Dempster-Shafer Theory of Belief Functions: An Introduction*, 2008, pp. 1–34.

[69] A. Josang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*, 2016.

[70] W. Bao, Q. Yu, and Y. Kong, "Evidential deep learning for open set action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 329–13 338.

[71] C. Zhao, D. Du, A. Hoogs, and C. Funk, "Open set action recognition via multi-label evidential learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 982–22 991.

[72] J. Lou, W. Liu, Z. Chen, F. Liu, and J. Cheng, "Elfnet: Evidential local-global fusion for stereo matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 738–17 747.

[73] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 14 927–14 937.

[74] B. Li, Z. Han, H. Li, H. Fu, and C. Zhang, "Trustworthy long-tailed classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6960–6969.

[75] W. Bao, Q. Yu, and Y. Kong, "Opental: Towards open set temporal action localization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2969–2979.

[76] M. Chen, J. Gao, and C. Xu, "Uncertainty-aware dual-evidential learning for weakly-supervised temporal action localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 896–15 911, 2023.

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

[78] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the dirichlet distribution and related processes," *Department of Electrical Engineering, University of Washington, UWEETR-2010-0006*, 2010.

[79] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[80] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[81] J. Howard, "Imagenette," https://github.com/fastai/imagenette/.

[82] J. Cui, S. Liu, L. Wang, and J. Jia, "Learnable boundary guided adversarial training," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 701–15 710.

[83] Y. Mo, D. Wu, Y. Wang, Y. Guo, and Y. Wang, "When adversarial training meets vision transformers: Recipes from training to architecture," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, 2022, pp. 18 599–18 611.

[84] G. Li, S. Ding, J. Luo, and C. Liu, "Enhancing intrinsic adversarial robustness via feature pyramid decoder," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 797–805.

[85] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944.

[86] H. Han, Q. Zhang, F. Li, and Y. Du, "Foreground capture feature pyramid network-oriented object detection in complex backgrounds," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2024.

[87] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoomnext: A unified collaborative pyramid network for camouflaged object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9205–9220, 2024.

[88] M. Teye, H. Azizpour, and K. Smith, "Bayesian uncertainty estimation for batch normalized deep networks," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 4907–4916.

[89] B.-K. Lee, J. Kim, and Y. M. Ro, "Mitigating adversarial vulnerability through causal parameter estimation by adversarial double machine learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4499–4509.

[90] D. Wu, S.-T. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 2958–2969.

[91] S. Zhao, X. Wang, and X. Wei, "Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9338–9352, 2024.

[92] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, p. 6405–6416.

[93] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 1310–1320.

[94] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, "Provably robust deep learning via adversarially trained smoothed classifiers," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[95] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: Dnn-oriented jpeg compression against adversarial examples," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 860–868.

[96] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8571–8580.

[97] X. Wei, S. Zhao, and B. Li, "Revisiting the trade-off between accuracy and robustness via weight distribution of filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8870–8882, 2024.

Zheng Wang is currently a professor with Tongji University, China. He received the B.E. and Ph.D. degrees both from Zhejiang University, China, in 2011 and 2017 respectively. He has co-/authored about 50 peer-reviewed papers, such as CVPR, ICCV, ACM MM, IEEE TPAMI, IEEE TIP, IEEE TNNLS, IEEE TMM, and IEEE TCSVT. His current research interests mainly focus on embodied AI, multimedia understanding and computer vision.

Xing Xu received the Ph.D. degree from Kyushu University, Japan, in 2015. He is currently with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China, China. He is the recipient of six academic awards, including the IEEE Multimedia Prize Paper 2020, Best Paper Award from ACM Multimedia 2017, and the World's FIRST 10K Best Paper Award-Platinum Award from IEEE ICME 2017. His current research interests focus on multimedia information retrieval and computer vision.

Lei Zhu is currently a professor with the School of Electronic and Information Engineering, Tongji University. He received his B.Eng. and Ph.D. degrees from Wuhan University of Technology in 2009 and Huazhong University Science and Technology in 2015, respectively. He was a Research Fellow at the University of Queensland (2016-2017). His research interests are in the area of large-scale multimedia content analysis and retrieval. Zhu has co-/authored more than 100 peer-reviewed papers, such as ACM SIGIR, ACM MM, IEEE TPAMI, IEEE TIP, IEEE TKDE, and ACM TOIS. His publications have attracted more than 8,600 Google citations. At present, he serves as the Associate Editor of IEEE TBD and ACM TOMM He has served as the Area Chair of ACM Multimedia, Senior Program Committee for AAAI and SIGIR. He won ACM SIGIR 2019 Best Paper Honorable Mention Award, ADMA 2020 Best Paper Award, ChinaMM 2022 Best Student Paper Award, ACM China SIGMM Rising Star Award.

Yi Bin is currently with the Tongji University, Shanghai, China. He received the Ph.D. degree from UESTC in 2020. His research interests include multimedia analysis, vision understanding and deep learning.

Guoqing Wang received the B.E. and M.E. degrees from the China University of Mining and Technology, in 2014 and 2017, respectively, and the Ph.D. degree from The University of New South Wales, Australia, in 2021. He is currently with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China. His research interests include computer vision, image analysis, and machine learning

Yang Yang received the Ph.D. degree in computer science from the University of Queensland, Brisbane, QLD, Australia, 2012. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. He was a Research Fellow with the National University of Singapore, Singapore, from 2012 to 2014. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Heng Tao Shen (Fellow'21) received the B.Sc.(Hons.) and Ph.D. degrees from the Departmentof Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively. His research interests mainly include multimedia search, computer vision, artificial intelligence, and big data management. Prof. Shen is a member of Academia Europaea, fellow of ACM, IEEE and OSA.