Adversarial Attack Detection via Fuzzy Predictions

Yi Li, *Member, IEEE,* Plamen Angelov, *Fellow, IEEE,* Neeraj Suri, *Senior Member, IEEE*

***Abstract*—Image processing using neural networks act as a tool to speed up predictions for users, specifically on large-scale image samples. To guarantee the clean data for training accuracy, various deep learning-based adversarial attack detection techniques have been proposed. These crisp set-based detection methods directly determine whether an image is clean or attacked, while, calculating the loss is non-differentiable and hinders training through normal back-propagation. Motivated by the recent success in fuzzy systems, in this work, we present an attack detection method to further improve detection performance, which is suitable for any pre-trained neural network classifier. Subsequently, the fuzzification network is used to obtain feature maps to produce fuzzy sets of difference degree between clean and attacked images. The fuzzy rules control the intelligence that determines the detection boundaries. Different from previous fuzzy systems, we propose a fuzzy mean-intelligence mechanism with new support and confidence functions to improve fuzzy rule's quality. In the defuzzification layer, the fuzzy prediction from the intelligence is mapped back into the crisp model predictions for images. The loss between the prediction and label controls the rules to train the fuzzy detector. We show that the fuzzy rule-based network learns rich feature information than binary outputs and offer to obtain an overall performance gain. Experiment results show that compared to various benchmark fuzzy systems and adversarial attack detection methods, our fuzzy detector achieves better detection performance over a wide range of images.**

***Index Terms*—Neural network, adversarial attack detection, fuzzification, fuzzy mean-intelligence, confidence function**

I. Introduction

With the advent of deep learning, neural network models [1]–[3] have demonstrated revolutionary performance in machine learning tasks, for example, natural language processing (NLP) [4], object detection [5] and audio signal processing [6], of real-world datasets. Nevertheless, the vulnerability of neural networks to image corruptions and adversarial examples has been unveiled [7]. Adversarial attacks are techniques used to manipulate neural networks by introducing small, often imperceptible, perturbations to input images, audio, and videos, causing the model to make incorrect predictions [8]. The impact of adversarial attacks is significant, as they can undermine the reliability and security of AI systems in critical applications, such as autonomous driving [9], cyber security [10] and facial recognition [11]. Addressing these vulnerabilities is crucial for developing robust and trustworthy neural networks. Consequently, research in adversarial defense mechanisms and attack detection [12]–[14] has become a vital area in the field of AI.

In the machine learning, neural networks are trained to re-estimate the input image sample by minimizing the reconstruction loss between the re-constructed and original images. Towards guarantee credibility of input images by attack detection, recent research looks at the distribution of the mean reconstruction error (MRE) for attacked and original image samples [15]. Particularly, image samples with higher MRE is potentially caused by an adversarial attack or perturbation to the clean image which leads to a wrong prediction or poorer reconstruction as the model output. Since adversarial attacks from new diverse sources become increasingly sophisticated, obtaining lebelled images of all possible attack algorithms or building attack detection techniques for each type of attack algorithms are not feasible. There are various deep learning methods [16]–[18] to make models more to unknown attack algorithms. However, some require retraining of the model with adversarial examples [19] or altering loss functions during the training step.

Recently, fuzzy set theory is widely applied in deep learning techniques [20], [21]. Different from crisp set-based techniques, which output 0 or 1, fuzzy logic is a system of many-valued logic where the truth value of variables can be any real number between 0 and 1 [22]. It is applied to process the concept of partial truth, where the truth value may range between completely true and completely false. Recent studies have shown that the fuzzy system offers several advantages in handling problems traditionally addressed by crisp set-based techniques. Firstly, fuzzy sets enable the representation of uncertainty by assigning degrees of membership to elements [20], [22]. In contrast, crisp sets operate under a binary classification, which proves inadequate in situations with vague or uncertain information, particularly in the context of small and imperceptible adversarial attacks. Secondly, Fuzzy sets exhibit greater robustness in the presence of noise or data attacks [23], while crisp sets, which are sensitive to exact values, are adversely affected by small variations. Therefore fuzzy logic can potentially handle tasks at several levels, from low level (e.g., binary classification) to a high level (e.g., model-based structural recognition and scene interpretation). It provides a flexible framework for information fusion as well as powerful tools for reasoning and decision making [24]. In this paper, we show how the use of fuzzy detectors offers significant benefits in adversarial attack detection. Specifically, we propose a fuzzification process with difference degree between clean and attacked images.

The contributions of this paper are summarized as follows:

- A fuzzy rule-based detector, simplified as fuzzy detector, is introduced as a novel approach to address the adversarial attack detection problem. The proposed method addresses the limitations of crisp set-based predictions and offers advantages in imperceptible attack detection.
- Different from previous intelligence in fuzzy logic, we propose a new intelligence mechanism to improve fuzzy rule's quality. The proposed support and confidence functions are

Y. Li, P. Angelov, and N. Suri are with School of Computing and Communications, Lancaster University, U.K.

E-mail for correspondence: y.li154@lancaster.ac.uk

shown to be better adapted to the fitness function than the previous work.

• A comprehensive evaluation of the proposed fuzzy detector with various backbones over a wide range of datasets and attack algorithms is presented. The experimental results confirm the effectiveness of the proposed method. Furthermore, it shows a promising way to apply fuzzy logic in adversarial attack detection task.

II. Related Work

A. Attacks

As one of commonly used single-step adversarial attacks, Fast Gradient Sign Method (FGSM) calculate the gradient by using backpropagation [25]. Assume y and x are the clean image and the attacked image, respectively, the adversarial image is calculated as $y + \epsilon * \text{sign}(\Delta y J(\lambda, x, y))$ with the scale of the distortion ϵ and the cost function $J(\lambda, x, y)$. As an iterative version of the FGSM, Projected Gradient Descent (PGD) introduces a perturbation in each step during the training to improve robustness [26]. The PGD attack motivates the recent success in diffusion models [27]. In [28], the authors assert DeepFool is the first effective method to accurately compute the robustness of state-of-the-art deep classifiers to perturbations on large-scale datasets. Moreover, Basic Iterative Method (BIM) attack is introduced in [29] show that neural networks are vulnerable to adversarial examples by feeding adversarial images obtained from a cell-phone camera to a pre-trained classifier and measuring the classification accuracy of the system. Carlini & Wagner (CW) [30] demonstrate that conventional defensive algorithms cannot guarantee the robustness of neural networks by introducing unseen attack algorithms that are successful on both distilled and undistilled neural networks with 100% probability. In [31], the authors introduce Jacobian-based saliency map attack (JSMA) to generate craft adversarial samples based on a precise understanding of the mapping between inputs and outputs of neural networks. As the most recent adversarial attack, semantic similarity attack on high-frequency components (SSAH) concentrates in semantic similarity on feature representations [7]. To maintain the perceptual similarity between original and adversarial data, the authors introduce a low-frequency constraint to limit perturbations within high-frequency components of input data. The high-frequency components of an image capture minor details and noise, while the low-frequency components convey fundamental information. The authors report that the algorithm is one of the strongest attacks to recent detection and defense techniques [7].

B. Adversarial Attack Detection

In [32], a general framework is introduced to defend object detectors against adversarial attacks by using segment and complete defense (SAC). Each image is segmented into patches by patch masks which provide pixel-level localization of adversarial patches. Then the completion algorithm is trained to remove the adversarial patch from the image if the outputs of the segmenter are within a certain Hamming distance of the ground-truth patch masks. Qi et al. train two neural networks in [33]. Some adversarial attack samples are generated toward the local DL model. Then, the target model is attacked and produces perturbed samples. In the adversarial training, the misclassification probability of all training samples is estimated by the local model to detect and delete perturbed samples from the dataset. Different from these techniques, multiple prediction heads (i.e., detectors) are combined to generate predictions from different depths in deep models and introduce shallow information for inference [34]. The distribution parameter is estimated by moment matching. Then, cognitive uncertainty from the adversarial attacks becomes easier to remove. As a semi-supervised learning network, an adversarial autoencder enables imperceptible attack learning multiclassification tasks for adversarial attacks [35]. The experiments show high detection accuracy of the AAE model with only very limited training samples. In more recent attack detection techniques, center-outward ordering of points is estimated with the data distribution [36], which makes the halfspace-mass (HM) depth a natural choice for adversarial attack detection in the feature space. To improve the performance of the attack detector, Hussian et al. apply naturally occurring noises to generate boundary- and decision-based attacks to attack the neural network [37].

C. Fuzzy Systems

Recent fuzzy system studies demonstrate high performance on classification and detection tasks. In order to enhance the detection performance of adversarial attacks to deep models and boost machine learning robustness, classification boundaries are blurred [38]. The network traffic is set with linear decision trees are wrapped by a one-class-membership scoring algorithm. In [39], a multiple-attribute decision-making (MADM) model is introduced for fuzzy classification. Membership functions of a fuzzy set from training data are constructed to form a decision-making matrix. However, these techniques have some shortcomings, e.g., limited generalization capability, which leads to obtaining extensive uncovered image samples over new unseen samples. Therefore, multiple fuzzy candidate rules to each example [40]. The usage of more rules boosts the generalization capacity of the feature information to further improve the classification accuracy. Furthermore, in order to efficiently address anomaly detection problem, accurate and interpretable rules are extracted [21]. The population of individual rules is evolved in an evolution system. Then, the fuzzy rules in the system are mined with a Michigan cooperative approach.

III. Proposed Method

In this section, we present the fuzzy detector-based adversarial attack detection framework. Data preparation and the encoder in the overall framework of the proposed method is introduced in the first subsection, followed by the description of the fuzzy detector and training losses in the remaining subsections.

A. Data Preparation and Encoder

The overall framework of the proposed fuzzy detector-based adversarial attack detection is presented in Fig. 1. The aim

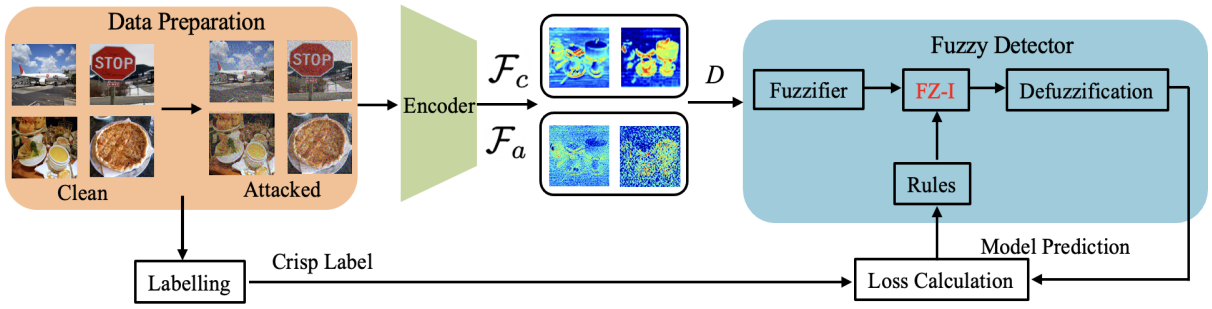


Fig. 1. Proposed training framework of the proposed fuzzy detector-based adversarial attack detection method. The fuzzy detector obtains the feature maps $\mathcal{F}_c$ and $\mathcal{F}_a$ and converts the difference between them into a fuzzy set. Then, a fuzzy prediction is generated by using the fuzzy rule with the proposed support and confidence functions. The defuzzification module converts the fuzzy set back into a crisp model prediction. Furthermore, the loss between the label and model prediction is utilized to fine-tune the fuzzy rule and make more accurate predictions.

of adversarial attack detection is to learn feature information of images and detect the difference between feature maps of clean and attacked images at the pixel level by re-estimating the image sample.

Initially, the noisy data is generated by attacking a clean image sample with a random attack algorithm, e.g., FGSM attack. The feature maps $\mathcal{F}_c$ and $\mathcal{F}_a$ are extracted from clean and attacked images by using a pre-trained ImageNet model, respectively. Particularly, we select the pre-trained EfficientNetV2-XL on the ILSVRC dataset [41] because it achieves the state-of-the-art benchmark on the ILSVRC challenge. The comparison of different encoder backbones will be provided in the experiment section.

B. Fuzzy Detector

The feature maps $\mathcal{F}_c$ and $\mathcal{F}_a$ from the pre-trained encoder are fed into the proposed fuzzy detector with hard labels, i.e., clean or attacked samples. Then, the mean squared error (MSE) loss at the pixel level between the feature maps is calculated as:

$$\mathcal{L}_{\mathcal{F}} = \frac{1}{N}\sum_{n=1}^{N}(\mathcal{F}_a - \mathcal{F}_c)^2 \tag{1}$$

where n refers to index of each pixel in feature maps and N is the total number of pixels. We select the MSE loss because it is more simple, interpretable, and differentiable [42] than other loss algorithms, e.g., cross entropy loss. Then, in order to learn more feature information than the crisp value-based prediction, the loss is converted into the fuzzy set, which describes the difference degrees between clean and attacked feature maps at the pixel level. Particularly, high loss values refer to very different feature information, which leads to a high possibility of attacks to the image, and vice versa. In this work, we use a triangular fuzzifier [43] to design the fuzzification. As a consequence of the fuzzification, we obtain a non-interval type-2 fuzzy system in which sets are characterised by fuzzy subsets of the truth range and the membership function is cropped triangular. The degree of differences in the fuzzy set quantifies the difference levels across the feature maps of clean and attacked images. The triangular membership function is illustrated in Fig. 2.

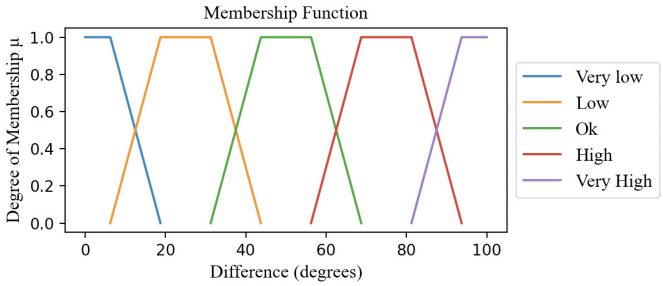


Fig. 2. The membership of the fuzzy set.

The rules of the proposed fuzzifier follow a commonly used fuzzy-rule-based classifier [20], [22] as:

$$\begin{aligned} R^i: \text{IF}\,(o_1\ is\ around\ o_1^{i*}) \quad &\text{AND} \quad (o_2\ is\ around\ o_2^{i*}) \\ \text{AND}\cdots\text{AND}\,(o_n\ is\ around\ o_n^{i*}) \quad &\text{THEN}\,(P^i) \end{aligned} \tag{2}$$

where $o = [o_1, o_2, \ldots, o_n]^T$ is the pixels of feature maps. The prototype of i-th fuzzy rule is denoted as o_n^{i*}. In the intelligence layer, (o_n is around o_n^{i*}) indicates the l-th fuzzy set of the i-th fuzzy rule R^i. To achieve that, we consider the Eucledian Distance d between o_n and o_n^{i*} with a hyper-parameter α_n^i. When the distance d is smaller than α_n^i, the fuzzy prediction with the i-th fuzzy rule is P^i that predicts how much the model trusts the image. The hyperparameter α_n^i is further updated to improve the boundary accuracy in the training stage.

In order to determine the detection boundaries based on the membership function shown in Fig. 2, we propose a fuzzy mean-intelligence (FZ-I) mechanism. Firstly, we define a fitness function $f(\cdot)$ based on a combination of a confidence function $\mathrm{C}(\cdot)$ and a support function $\mathrm{S}(\cdot)$ with the i-th fuzzy rule R_i described in equation (2).

$$f(R_i) = \mathrm{C}(R_i) + \mathrm{S}(R_i) \tag{3}$$

where $\mathrm{C}(R_i)$ measures accuracy of the fuzzy rule with the m-th sample x_m and the p-th class $Class_p$ as:

$$\mathrm{C}(R_i) = \frac{\sum_{m\in Class_p}\sum_{l=1}^{L}\varphi(v_{kl}(x_m))}{\sum_{m=1}\sum_{l=1}^{L}\varphi(v_{kl}(x_m))} \tag{4}$$

where v_{kl} represents the k-th dimension of the membership degree of the l-th antecedent fuzzy set. In this work, we exploit the binary classification (i.e., clean or attacked images) in adversarial attack detection problem. We define the binary function of the membership degree $\varphi(v_{kl})$ with the threshold value T as:

$$\varphi\left(v_{kl}\left(x_m\right)\right) = \left\{ \begin{array}{ll} 1 & \text{if } v_{kl}\left(x_m\right) > T \\ 0 & \text{Otherwise} \end{array} \right\} \tag{5}$$

Fuzzy logic leverages human expertise and intuition in system design [22]. An adaptive threshold allows system designers to incorporate their domain knowledge or intuition into the fuzzy system, adjusting the threshold to align with the task. Moreover, in real-world datasets, input data potentially vary due to different distributions. The adaptability in the threshold helps the system handle different scenarios. Therefore, we empirically set the threshold value from the distribution of the dataset. This adaptive threshold mechanism used in the fitness function aims to adapt the miner system to problems with dynamic training data. The adaptive threshold mechanism is also used in our inference system for attack detection. To find the winner rule, our inference system compute, for each rule in the database, the sum of the membership degrees. If the obtained result exceeds the defined threshold value, the instance is classified as an anomaly; otherwise, it is classified as normal. Moreover, $\mathrm{S}(R_i)$ with the i-th fuzzy rule R_i measures how often the fuzzy rule appears in training images:

$$\mathrm{S}\left(R_i\right) = \frac{\sum_{m \in Class_p} \sum_{j=1}^{J} \varphi\left(v_{kl}\left(x_{mj}\right)\right)}{M \times J} \tag{6}$$

where j is the index of mutant vectors. These mutant vectors are combined sets of linguistic values, e.g., very clean, few clean, medium, very noisy, and extreme noisy. Total numbers of samples and mutant vectors are denoted as M and J, respectively. In the proposed FZ-I mechanism, we maximize the fitness function so as to be more adapted to the fuzziness of the system and thus improve the rule's quality.

A centroid defuzzification method is exploited to convert the fuzzy prediction set into the model prediction [44]. Particularly, the center of gravity of the fuzzy set is calculated along the difference degree as:

$$P = \frac{\sum_i \mu_D\left(P_i\right) P_i}{\sum_i \mu_D\left(P_i\right)} \tag{7}$$

where P is the model prediction, i.e., 0 or 1 for clean or attacked image. The membership function $\mu_D(\cdot)$ is:

$$\mu_D = \begin{cases} 1, & \text{if a} < x < \text{b} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where a and b are both trainable hyper-parameters. The fuzzification and defuzzification are summarized in Fig. 3.

C. Training Losses

The training loss is calculated as follows. Firstly, we calculate the fuzzy loss $\mathcal{L}_{\mathcal{F}}$ between the label and fuzzy prediction.

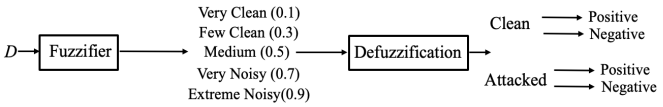


Fig. 3. Proposed fuzzy set-based fuzzy detector. The crisp difference between feature maps from clean and attacked images are converted into a fuzzy set to map into a fuzzy measure between 0 and 1 to describe how noisy is the input image, i.e., higher values indicate more noisy images. The defuzzification module makes a crisp prediction based on the fuzzy set.

Secondly, the overall loss $\mathcal{L}$ is calculated by the loss between the label and model prediction $\mathcal{L}_{\mathcal{M}}$ with λ_1 and λ_2 as:

$$\mathcal{L} = \begin{cases} \lambda_1 \cdot \mathcal{L}_{\mathcal{F}}, & \text{if } \mathcal{L}_{\mathcal{M}} \neq 0 \\ \mathcal{L}_{\mathcal{F}} / \lambda_2, & \text{otherwise} \end{cases} \tag{9}$$

The loss $\mathcal{L}$ updates the parameter α_n to refine the fuzzy rules, making more accurate fuzzy predictions. The pseudo-code of the proposed fuzzy rule-based attack detection method is summarized in Algorithm 1.

Algorithm 1: Fuzzy rule-based detector.

Input: Feature maps $\mathcal{F}_c$ and $\mathcal{F}_a$, label X, pixels of feature maps $o = \{o_1, o_2, ..., o_N\}$, prototypes for i-th fuzzy rule $o^{i*} = \{o_1^{i*}, o_2^{i*}, ..., o_N^{i*}\}$, loss constraints λ_1 and λ_2, epoch $E_{\max}$

Output: Model prediction P

Initialize hyperparameters a, b, α_n^i;
for $E = 1, 2, ..., E_{\max}$ **do**
  $\mathcal{L}_{\mathcal{F}} = \mathrm{MSE}(\mathcal{F}_c, \mathcal{F}_a)$ // Calculate the loss between feature maps;
  for $i = 1, 2, ..., I$ **do**
    $R_i \leftarrow o, o^{i*}, \alpha_n^i$ // Update fuzzy rules;
    $f(R_i) = \mathrm{C}(R_i)$+$\mathrm{S}(R_i)$ // FZ-I;
    if $f(R_i) < f(R_{i-1})$ **then**
      $R_i = R_{i-1}$ // Maximize R with FZ-I ;
    end
    if $d(o_n, o_n^{i*}) < \alpha_n^i$ **then**
      $P_i \leftarrow R_i$ // Fuzzy prediction ;
    end
  end
  $P \leftarrow P_i, a, b$ with Eq. (8) // Defuzzification;
  if $X = P$ **then**
    $\mathcal{L} = \mathcal{L}_{\mathcal{F}} / \lambda_2$;
  else
    $\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\mathcal{F}}$
  end
  $a, b, \alpha_n^i \leftarrow \mathcal{L}$ // Updates parameters with loss ;
end

IV. Experimental Results

A. Datasets and Attacks

We extensively perform experiments on ImageNet-R [45], Canadian Institute For Advanced Research-10 (CIFAR-10) [46], Common Objects in Context (COCO) [47], and ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

[41]. In the training, validation, and test stages, 50,000, 10,000, and 10,000 images are randomly selected from each dataset.

Moreover, we select 7 adversarial attack algorithms to generate attacked images due to their robustness to recent defense and recovery techniques [12], [48]. We summarize the parameters of these attacks in Table I.

TABLE I
PARAMETERS OF SEVEN ADVERSARIAL ATTACKS

Attack	Parameters
FGSM	ϵ=0.008
PGD	ϵ=0.01, α=0.02, Steps=40
SSAH	α=0.01
DeepFool	Steps=20
BIM	ϵ=0.03, α=0.01, Steps=10
CW	C=2, Kappa=2, Steps=500, learning rate=0.01
JSMA	γ=0.02

B. Backbones and Competitors

As aforementioned we use the pre-trained EfficientNetV2-XL [49] as the encoder's backbone. Moreover, we apply the proposed fuzzy detector on several state-of-the-art backbone models, e.g., Res2Net-v1b-101 (ResNet) [50], YOLOX-L (YOLO) [2], and PRB-FPN6-2PY (PRB) [51]. These models are initialized and re-trained with the proposed fuzzy logic.

In addition, the proposed method is evaluated and compared to 7 adversarial attack detection techniques [32]–[37], [48] and 3 fuzzy systems [38]–[40]. It is highlighted that these models are reproduced as the original implementations in the literature but with same data as the proposed method.

C. Performance Measure

To evaluate and compare the adversarial attack detection accuracy, we use the detection rate (DR) [48] as the performance measure. In particular, we define a detection rate for adversarial images ((DR_a)) as:

$$DR_a(\%) = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \qquad (10)$$

where TP and TN are true positive and true negative results, and FP and FN are false positive and false negative results. Moreover, we define a detection rate for clean images ((DR_c)) to evaluate the true positives as:

$$DR_c(\%) = \frac{TP}{TP+TN} \times 100 \qquad (11)$$

D. Implementation Details

We dynamically set hyper-parameters λ_1 and λ_2 between 1 and 10 due to slightly different performance in experiments over datasets and attack algorithms. In each experiment, we aim to find the optimal values of λ_1 and λ_2 that strike a balance between the model's ability to capture the underlying data distribution and its robustness against adversarial attacks. To achieve this, we employ a systematic approach to dynamically tune these hyperparameters based on the specific characteristics of the dataset and the nature of the attacks encountered. By adapting the values of λ_1 and λ_2 according to the experimental conditions, we can effectively tailor the model's behavior to the task at hand, thereby enhancing its performance and generalization capabilities. This dynamic parameter tuning strategy enables us to explore a wide range of parameter configurations and identify the most suitable settings for robust adversarial defense. For example, we find that smaller values of λ_1 and λ_1 provide slightly better performance (89.3% $\rightarrow$ 89.8%). We train the proposed model with the M-SGD optimizer and empirically set the learning rate to 0.0008. We set the batch size to 32. The network is trained for 100 epochs with Tesla V100 GPUs.

E. Number of Fuzzy Rules

We conducted experiments to demonstrate the trade-off between performance improvement and computational cost, specifically, varying the number of fuzzy rules. Fig. 4 (upper left) present these results, with each data point being an average of 70,000 experiments (10,000 images of ImageNet-R $\times$ 7 attacks).

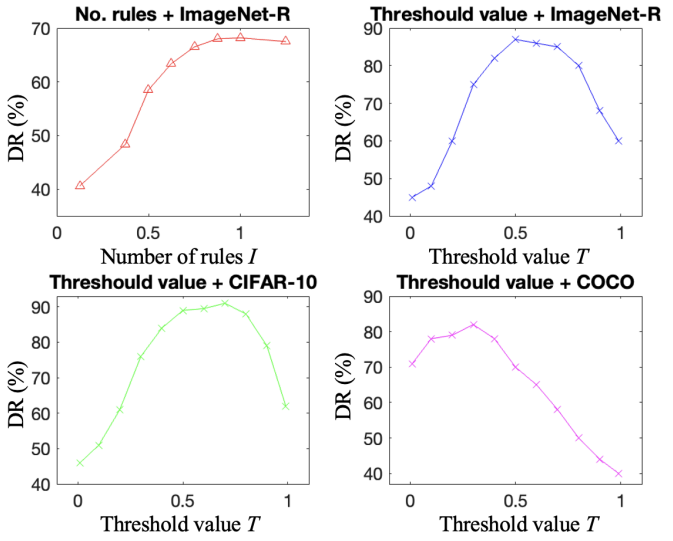


Fig. 4. Detection performance against the number of fuzzy rules (upper left) and threshold value (others).

Fig. 4 (upper left) compares the number of fuzzy rules against detection accuracy on ImageNet-R. The results indicate: (1) $I = 35$ offers the best trade-off, validating the chosen implementation setting. (2) The detection performance is sensitive to the number of fuzzy rules. This is maybe because the number of rules affects the coverage of different input scenarios. Limited rules may result in an insufficient representation of the input space, leading to imprecise or incomplete decision-making.

F. Threshold Value

As aforementioned, we empirically set the threshold value in equation (5) to 0.5. In this section, we aim to confirm the chosen configuration. The results are presented in Fig. 4, with each data point being an average of 70,000 experiments (10,000 images $\times$ 7 attacks).

TABLE II
ATTACK DETECTION RATIO (%) ON THE CIFAR-10 AND IMAGENET-R DATASETS.

	Detection Ratio (%)													
	CIFAR-10							ImageNet-R						
Method	FGSM	PGD	SSAH	DeepFool	BIM	CW	JSMA	FGSM	PGD	SSAH	DeepFool	BIM	CW	JSMA
SAC [32]	60.1	59.7	56.8	21.6	16.1	17.7	23.4	58.9	57.5	52.9	19.5	15.8	17.0	21.2
Sim-DNN [48]	70.5	60.0	49.4	26.7	22.3	22.9	31.1	71.0	66.2	61.4	28.3	26.0	26.1	34.6
DTBA [33]	78.3	75.6	71.7	36.2	30.7	32.3	40.4	78.0	72.4	68.9	34.8	32.7	32.9	36.5
MH-UI [34]	79.2	76.5	74.6	49.1	49.7	52.5	71.8	79.4	74.9	70.6	59.2	46.0	44.7	69.9
AAE [35]	80.5	76.9	75.4	63.7	60.4	60.2	75.8	79.7	75.6	71.8	60.8	55.1	55.0	74.9
HSJ [37]	77.5	75.2	75.6	60.1	55.2	59.4	72.0	78.1	76.8	71.3	59.7	48.9	53.6	71.5
HM [36]	86.9	84.5	84.0	80.6	76.7	77.9	87.8	87.3	85.2	80.7	85.2	79.3	81.1	88.4
FCB [38]	49.8	47.1	43.6	15.1	10.7	11.4	16.8	48.5	46.8	43.9	14.9	10.1	10.7	15.9
MF [40]	51.4	48.0	46.1	16.4	11.1	11.9	18.2	53.6	48.2	48.2	15.0	10.7	11.2	16.5
MADM [39]	62.4	54.2	51.5	19.0	15.8	14.3	23.1	60.7	54.0	54.6	21.6	13.8	13.9	20.1
ResNet+Ours	89.3	86.7	85.9	88.8	86.3	86.5	89.6	89.0	85.8	85.2	88.1	85.6	85.7	88.9
PRB+Ours	89.5	87.0	86.2	87.7	86.1	**86.8**	**90.1**	89.5	85.9	85.6	90.3	86.2	**86.4**	**90.3**
YOLO+Ours	**89.6**	**87.2**	**86.6**	**89.5**	**88.0**	85.9	**90.1**	**89.8**	**86.2**	**85.7**	**90.4**	**87.8**	86.1	89.9

TABLE III
ATTACK DETECTION RATIO (%) ON THE COCO AND ILSVRC DATASETS.

	Detection Ratio (%)													
	COCO							ILSVRC						
Method	FGSM	PGD	SSAH	DeepFool	BIM	CW	JSMA	FGSM	PGD	SSAH	DeepFool	BIM	CW	JSMA
SAC [32]	58.7	58.9	55.0	21.1	16.0	16.8	22.8	56.3	55.2	52.7	16.2	13.9	16.7	21.0
Sim-DNN [48]	63.5	62.5	57.8	24.2	21.0	22.8	30.5	74.1	70.8	66.7	25.7	25.8	28.2	33.1
DTBA [33]	74.6	70.2	67.1	34.0	25.9	31.6	37.8	79.8	79.1	74.6	36.8	35.3	36.5	40.2
MH-UI [34]	76.0	74.9	71.5	47.5	48.8	50.1	37.5	80.3	81.6	75.0	52.4	52.2	55.6	72.0
AAE [35]	77.3	74.6	72.9	56.5	57.7	55.1	69.1	82.0	83.5	75.4	66.6	58.1	60.7	77.3
HSJ [37]	76.6	73.0	71.7	58.8	51.9	57.5	68.3	73.9	74.8	69.5	56.7	43.6	51.0	68.7
HM [36]	85.6	83.7	81.2	78.3	78.8	75.8	84.9	89.6	83.6	79.9	82.8	76.9	80.1	86.7
FCB [38]	46.7	45.0	39.8	14.3	10.1	10.5	14.4	51.1	48.5	47.8	14.6	9.2	10.5	14.6
MF [40]	48.8	47.6	45.2	15.9	11.0	11.5	17.7	56.4	51.7	50.1	15.8	13.3	12.5	20.1
MADM [39]	61.2	52.6	48.9	18.5	15.3	14.0	22.8	64.8	58.3	56.2	26.7	18.0	17.2	23.5
ResNet+Ours	87.3	84.4	81.9	84.2	87.1	86.8	88.7	90.1	87.3	85.5	88.5	86.8	85.7	90.3
PRB+Ours	86.3	85.5	83.8	**85.0**	87.2	87.3	89.4	91.0	86.1	84.5	88.3	87.9	85.8	**91.4**
YOLO+Ours	**86.6**	**85.8**	**85.9**	84.8	**87.6**	**87.4**	**89.7**	**91.3**	**87.0**	**84.6**	**88.6**	**88.5**	**85.9**	**91.4**

As Fig. 4 (upper right) shows, detection accuracy on ImageNet-R starts to increase with T = 0.01 and reaches its peak around T = 0.5, but performance drops after the peak point. Therefore, Fig. 4 (upper right) suggests the threshold value T = 0.5 for ImageNet-R. Moreover, Fig. 4 (lower) confirms that expertise and intuition provide valuable insights when dealing with datasets that differ significantly from the current data distribution.

G. Comparisons to SOTA Methods

We compare the proposed method to state-of-the-art adversarial attack detection methods [32]–[35], [48] and fuzzy systems [38]–[40] with the same attack between the training and test stage. The results are provided in Tables II&III.

Tables II&III shows the averaged attack detection performance of the proposed method as compared with those of the methods using the CIFAR-10, ImageNet-R, COCO, and ILSVRC datasets. From these tables, it can be observed that: (1) In all the evaluated models, the proposed fuzzy prediction-based methods with different backbones offer the best effectiveness. Different from crisp set-based decision-making pipelines, the proposed fuzzy detectors convert the loss between feature maps into fuzzy sets and provide difference scores ('very clean', 'few clean', 'medium', 'very noisy', and 'extreme noisy'). Therefore, the proposed method exploits more feature information than binary decisions. The fuzzy rules are trained with difference scores to help the detector make more accurate decisions. (2) The proposed method offers the best attack detection performance with the YOLO model on all datasets. The reason is likely due to the combined implicit knowledge and explicit knowledge in the YOLOX decoder [2]. (3) Compared to the improvement in FGSM, PGD, and SSAH attacks, the improvement of detection accuracy tends to fall drastically when evaluating the true positives on clean image samples. For example, compared to ESMAF model, the proposed method with the YOLO model obtain 7.9% improvement on PGD attacked CIFAR-10 dataset, while it is only 2.8% on the true positive evaluation.

Furthermore, some qualitative analysis are given in Fig. 5 which are related to the reconstructions after detecting attacks of three randomly selected images from the COCO dataset. After comparing the reconstructed images with the original and attacked images, it can be observed that the reconstructions obtained via the proposed method, i.e., Fig. 5 (d), are closer to original images, which again confirms the efficacy of the proposed method.

Fig. 6 compares confusion matrices of AAE (left) and ours (right) on ImageNet-R. The results indicate: (1) Our model outperforms AAE in both TP and TN. (2) TP and TN are both relatively high, but there is a significant difference

TABLE IV
ATTACK DETECTION RATIO (%) WITH DIFFERENT ENCODER BACKBONES. EACH RESULT IS THE AVERAGE OF 40,000 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULTS. *Italic* SHOWS THE PROPOSED METHODS.

	Detection Ratio (%)								
Method	clean	FGSM	PGD	SSAH	DeepFool	BIM	CW	JSMA	Average
Pre-trained EfficientNetV2-XL + YOLO + Ours	94.2	89.3	86.6	85.7	88.3	88.0	86.3	90.3	**88.6**
Pre-trained VGG-16 + YOLO + Ours	93.8	89.4	86.4	85.8	88.1	88.2	86.0	89.8	88.4
Pre-trained Resnet56 + YOLO + Ours	93.2	88.8	85.5	85.1	87.2	86.9	85.6	89.4	87.7
Reproduced EfficientNetV2-XL + YOLO + Ours	94.0	89.1	86.2	85.6	87.9	87.8	86.0	90.0	88.3
Reproduced VGG-16 + YOLO + Ours	93.5	89.0	86.2	85.1	87.4	87.8	85.5	89.3	88.0
Reproduced ResNet56 + YOLO + Ours	93.0	88.5	85.1	84.7	87.0	86.3	84.9	89.2	87.3
Pre-trained EfficientNetV2-XL + AAE	90.6	80.5	78.0	74.1	62.4	57.5	58.2	74.9	72.0
AAE	90.5	80.0	77.7	73.9	61.9	57.8	57.8	74.3	71.7

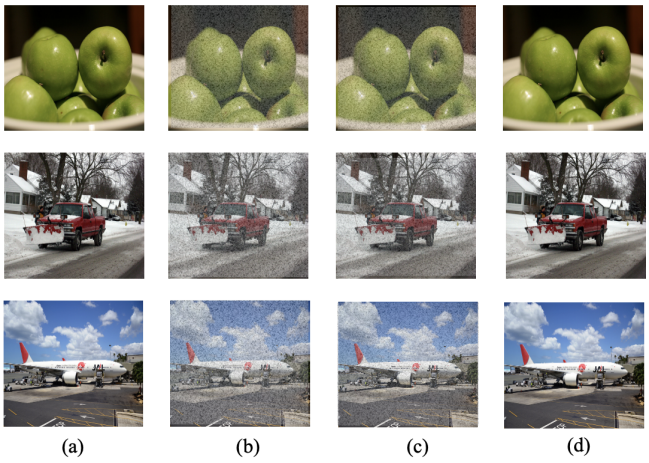


Fig. 5. Attack detection results: (a) original images; (b)&(c) attacked by random attack types and error rates; (d) reconstruction from attacks.

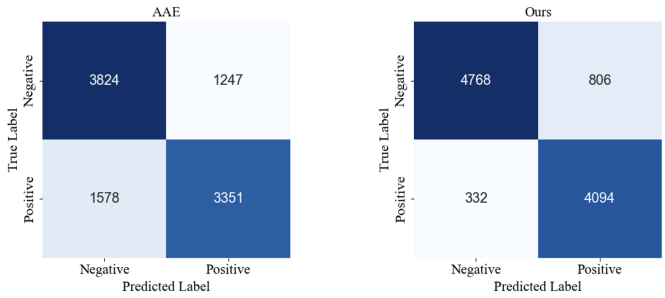


Fig. 6. Confusion matrices of AAE (left) and ours (right).

between them. This suggests that while the model performs well in correctly identifying clean images, it shows even greater proficiency in correctly classifying attacked images.

Moreover, Fig. 7 presents the t-distributed stochastic neighbour embedding (t-SNE) visualisation of the penultimate layer of the baseline YOLO model [2], proposed fuzzy rule-based YOLO model, and fuzzy mean-intelligence based YOLO model. We observe that the feature embeddings by the baseline YOLO model are not quite separable for detection of clean (blue) and attacked (yellow) images. The features representation from both proposed fuzzy detectors is generally better separated than those from the baseline YOLO model because the proposed fuzzy prediction-based methods capture more feature information than crisp value-based predictions. Moreover, the proposed method with the YOLO detector achieves the best separated clusters among clean and attacked images, which means that the learned representations in the embedding space are more distinguishable. These t-SNE visualization results demonstrate that proposed methods are able to learn discriminative feature representations which are better generalized to adversarial attack detection with various attack algorithms.

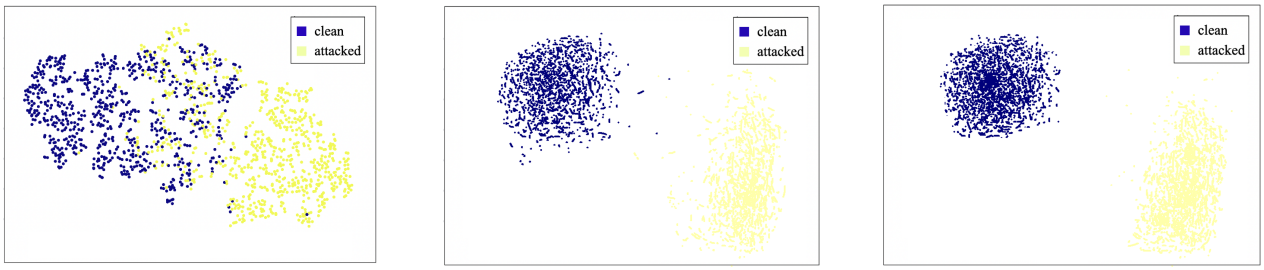
H. Comparison of Encoder Backbones

Moreover, to confirm the fair of experiments, we conduct experiments without pre-trained model. Firstly, we use different pre-trained models, i.e., VGG-16 [52] and Resnet56 [1] because they are commonly used feature extractor [48]. Secondly, we reproduce them with same data. Thirdly, we replace the original encoder in AAE [35] to pre-trained EfficientNetV2-XL [49] to evaluate the performance improvement when comparing to original implementation in [35]. It is highlighted that this experiment share the same experimental setting with Section IV. E, i.e., same attack algorithms and dataset between the training and test stages. Each result in Table IV is the average of 40,000 experiments (10,000 images $\times$ 4 datasets).

From Table IV, it can be observed that the proposed fuzzy detector brings much more performance improvement than the pre-trained encoders, which further confirms the effectiveness of the proposed fuzzy detector. For an example, when the pre-trained EfficientNetV2-XL is implemented as the feature extractor, the proposed method with the YOLO model achieves 16.6% better accuracy than the AAE decoder. However, comparing to original AAE, the combination of the pre-trained EfficientNetV2-XL and AAE only obtains a slight improvement, i.e., 0.3%, which again confirms the effectiveness of the proposed fuzzy detector than the SOTA attack detector. Moreover, the detection accuracy of reproduced models makes a slight difference to pre-trained models, which confirms the implementation.

I. Unseen Domain Study

In this experiment, to evaluate and compare the detection performance in a more challenging case, we use unseen attack algorithms and datasets in domains between the training and test stages. To achieve that, we first randomly select 5000, 1000, and 1000 images from each dataset for the training, validation, and test stages. Then, each image is attacked by using a random attack algorithm. Therefore, there are 140,000,



(a) Baseline YOLO (b) YOLO + Fuzzy rules (c) YOLO + Ours

Fig. 7. t-SNE feature visualization of the model penultimate layer of (a) Baseline YOLO model [2]; (b) YOLO + proposed fuzzy rules; (c) YOLO + proposed fuzzy rules + fuzzy mean-intelligence.

28,000 and 28,000 samples (5,000, 1000 and 1000 images $\times$ 4 datasets $\times$ 7 attack algorithms) images are labelled with the attack algorithm and dataset for the training, validation, and test stages, respectively. In each batch of 5000 training samples with same attack algorithm and dataset, we randomly select samples with different labels from test images. For example, the proposed fuzzy detector and competitor models are trained by using images with unseen attack algorithms and datasets, while evaluated by using images with FGSM attack from the COCO dataset. Therefore, each result in Table V is the average of 18,000 experiments (1000 images $\times$ 3 datasets $\times$ 6 attack algorithms).

TABLE V
ATTACK DETECTION RATIO WITH UNSEEN **ATTACK ALGORITHMS AND DATASETS** IN THE TEST STAGE. EACH RESULT IS THE AVERAGE OF 18,000 EXPERIMENTS.

Method	Detection Ratio (%)
SAC [32]	50.7
DTBA [33]	60.4
MH-UI [34]	63.8
AAE [35]	66.7
FCB [38]	40.9
MF [40]	45.1
MADM [39]	65.6
ResNet + Ours	72.5
PRB + Ours	73.3
YOLO + Ours	**73.4**

We can observe from Table V that the proposed fuzzy rule-based models achieve better performance than previous models. Moreover, compared to Table II, it is possible to note that the proposed DGAD models performance tends to fall less than the previous models when less training data applied and unseen attack type evaluated. The goal of the adversarial training provided by the MH-UI and AAE is to increase the model's robustness, however, they lack generalisation to unseen domains, i.e., datasets and attack algorithms. The proposed fuzzy detector maintains its performance stable even when adversarial or clean images from unknown datasets are presented to the detection model due to its inner fuzzy rules and detection mechanism that was projected for such scenarios.

We further conduct a more challenging experiment with different semantics, e.g., clear weather and foggy weather. The proposed methods and competitors are trained with 50,000 images from the ImageNet-R dataset. Table VI presents the results, each of them is average of 5,000 images of *clear* weather from the Cityspaces dataset or 5,000 images of *foggy* weather from the Foggy Cityscapes dataset [53].

TABLE VI
ATTACK DETECTION RATIO WITH IMAGENET-R $\rightarrow$ CITYSCAPES. F REFERS TO FOGGY WEATHER. β IS THE ATTENUATION COEFFICIENT.

Method	Clear	F (β=0.0005)	F (β=0.001)	F (β=0.002)
SAC [32]	59.9	58.4	57.9	55.0
Sim-DNN [48]	63.4	63.0	62.1	60.4
DTBA [33]	66.3	66.1	65.5	64.6
MH-UI [34]	70.7	69.9	69.5	67.7
AAE [35]	73.1	73.0	72.4	71.1
FCB [38]	56.5	55.9	55.2	52.6
MF [40]	59.8	59.2	58.3	56.0
MADM [39]	69.9	69.8	69.2	68.1
ResNet + Ours	71.7	71.5	71.1	70.5
PRB + Ours	72.9	72.6	72.2	71.7
YOLO + Ours	**73.3**	**73.1**	**73.0**	**72.5**

It can be observed that the proposed method with YOLO outperforms the competitor models for both clear and foggy weathers. The proposed method with PRB and ResNet is also competitive with AAE, the best of the state-of-the-art methods, particularly at high attenuation coefficients. (71.7% VS 71.1% with β=0.002).

J. Ablation Study

In this experiment, we investigate the effectiveness of each contribution based on the ImageNet-R dataset. The cross mark 7 for fuzzy logic means we only use crisp values to train the decoder. Then, we break the fitness function into the support function and confidence function. Then, we study the performance improvement of these two sub-functions individually. The ablation study is presented in Table VII and the setting of adversarial attack parameters is the same as Table I. It is highlighted that this experiment share the same experimental setting with Section IV. G, i.e., different attack algorithms and dataset between the training and test stages. Each result in Table VII is the average of 18,000 experiments (1,000 images $\times$ 3 datasets $\times$ 6 attack algorithms).

Initially, the effectiveness of the fuzzy prediction is studied. Compared to the baseline, the detection performance is significantly improved by the fuzzy detector. The reason is membership scores from fuzzification in the proposed fuzzy detector are added as new features to improve detection performance. As the most influential contribution, the fuzzy

TABLE VII
ABLATION STUDY OF TWO CONTRIBUTIONS IN THE PROPOSED METHOD. EACH RESULT IS THE AVERAGE OF 18,000 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULTS.

Ablation Settings			DR (%)
Fuzzy rule	Support function	Confidence function	
✗	✗	✗	61.6
✓	✗	✗	69.3
✗	✓	✗	-
✗	✗	✓	-
✗	✓	✓	-
✓	✗	✓	70.8
✓	✓	✗	71.1
✓	✓	✓	**73.4**

prediction provides a soft response whereas crisp predictions have discontinuous response at the detection boundary, which enables smoother fits and hence lower bias around the split boundaries.

Moreover, the experiment is performed by adding the proposed fitness function $f(\cdot)$. This is due to the different mechanisms implemented in the learning process, more specifically, the new support and confidence sub-functions used in the fitness function which are more adapted to the fuzziness of the system and the mechanism of sharing information between fuzzy rules through the fuzzy detector. In the proposed framework, the difference degrees are converted into fuzzy sets through a fuzzification process. Different from conventional approaches, the intelligence are tailored for each fuzzy rule as an important step in the framework. The complexity is increased when fuzzy sets shared by all fuzzy rules, but the intelligence of fuzzy system improves and further boost detection performance.

K. Discussion

The above detailed experimental results confirm that the proposed fuzzy detector with a new fuzzy mean-intelligence mechanism can further improve adversarial attack detection performance in different scenarios, i.e., seen or unseen datasets and attack algorithms, compared to the state-of-the-art attack detection methods and fuzzy systems.

YOLO shows better accuracy because it is optimized for object detection tasks, with a well-balanced structure for feature extraction and detection speed. The backbone integrates deep feature hierarchies effectively, making it robust to adversarial perturbations. In contrast, ResNet and PRB may not capture multi-scale features as efficiently in combination with the encoder, which could explain the lower accuracy. YOLOX's architecture may better align with EfficientNetV2-XL in extracting discriminative features from adversarial inputs.

Expertise and intuition may still provide valuable insights when dealing with datasets that differ significantly from the current data distribution, which is even more critical. Our understanding of the underlying data generating processes, potential biases, and domain-specific nuances are leveraged to make most appropriate decisions about adapting the threshold. Intuition, developed through experience and familiarity with the data, can also guide practitioners in identifying patterns, outliers, and anomalies that may impact the detection accuracy. However, it's essential to fine-tune the threshold on the new dataset to ensure the effectiveness and generalizability of the detection approach across different datasets. In future work, incorporating automated techniques, such as cross-validation or model monitoring, can help enhance the adaptability and robustness of the proposed method in diverse data settings.

The major limitation of this paper includes encoder and prototypes. Firstly, the proposed fuzzy logic is implemented on the decoder and further studies on encoder are out of scope of this paper. The fuzzy logic-based encoder is considered by converting feature vectors into fuzzy concepts in the future work. Secondly, prototypes play an important part in recent studies [48], which allows a reasoning process that relies on the similarity (proximity in the feature space) of a data sample to a given prototype. In the further study, we will exploit local peaks of the density as the prototype to help calculate the difference degree between clean and attacked image samples.

V. CONCLUSION

In this paper, we have proposed a fuzzy detector-based adversarial attack detection method, a simple yet effective replacement to the conventional crisp set-based decision-making pipelines. Differing from these pipelines, the difference degrees between clean and attacked feature maps provide rich information to improve the proposed model's ability to detect adversarial attacks. Moreover, we have proposed a fuzzy mean-intelligence mechanism with new support and confidence functions to improve fuzzy rule's quality. Our evaluation with different datasets and attacks has demonstrated the high effectiveness of the proposed method.

ACKNOWLEDGEMENT

Research supported by the UKRI Trustworthy Autonomous Systems Node in Security/EPSRC Grant EP/V026763/1.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[3] H. C. Li, P. F. Xiong, J. An, and L. X. Wang, "Pyramid attention networks," *Proceedings of The British Machine Vision Conference (BMVC)*, 2018.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[5] T. Lin, P. Goyal, R. Girshick, and P. D. K. He, "Focal loss for dense object detection," *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

[6] Y. Li, Y. Sun, W. Wang, and S. M. Naqvi, "U-shaped Transformer with frequency-band aware attention for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, p. 1511–1521, 2023.

[7] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen, "Frequency-driven imperceptible adversarial attack on semantic similarity," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[8] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805 – 2824, 2019.

[9] Y. Li, P. Angelov, Z. Yu, A. L. Pellicer, and N. Suri, "Federated adversarial learning for robust autonomous landing runway detection," *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, 2024.

[10] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems," *Journal of Information Security and Applications*, vol. 58, p. 102717, 2021.

[11] A. L. Pellcier, Y. Li, and P. Angelov, "Pudd: Towards robust multimodal prototype-based deepfake detection," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[12] Y. Li, P. Angelov, and N. Suri, "Rethinking self-supervised learning for cross-domain adversarial sample recovery," *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2024.

[13] J. Chen, T. Yu, C. Wu, H. Zheng, W. Zhao, L. Pang, and H. Li, "Adversarial attack detection based on example semantics and model activation features," *Proceedings of International Conference on Data Science and Information Technology (DSIT)*, 2022.

[14] Z. Ji, B. Yang, P. L. Yeoh, Y. Zhang, Z. He, and Y. Li, "Active attack detection based on interpretable channel fingerprint and adversarial autoencoder," *Proceedings of IEEE International Conference on Communications (ICC)*, 2022.

[15] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, 2021.

[16] Y. Li, P. Angelov, and N. Suri, "Domain generalization and feature fusion for cross-domain imperceptible adversarial attack detection," *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2023.

[17] C. Cintas, S. Speakman, V. Akinwande, W. Ogallo, K. Weldemariam, S. Sridharan1, and E. McFowland, "Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error," *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[18] Y. Li, P. Angelov, and N. Suri, "Self-supervised representation learning for adversarial attack detection," *Proceedings of European Conference on Computer Vision (ECCV)*, 2024.

[19] J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[20] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," *Springer Science Business Media*, 2013.

[21] W. Guendouzi and A. Boukra, "A new differential evolution algorithm for cooperative fuzzy rule mining: application to anomaly detection," *Evolutionary Intelligence*, vol. 15, p. 2667–2678, 2022.

[22] P. Angelov and X. Zhou, "Evolving fuzzy-rule-based classifiers from data streams," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 6, pp. 1462–1475, 2008.

[23] L. A. Zadeh and J. M. Mendel, "Fuzzy logic for tolerance to noise and variation in intelligent systems," *Proceedings of the IEEE*, vol. 83, no. 3, pp. 442–450, 1995.

[24] J.-J. Liu and J.-C. Fan, "A novel fuzzy c-means clustering algorithm based on local density," *Proceedings of International Conference on Intelligent Information Processing (IIP)*, 2020.

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations (ICLR)*, 2015.

[26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *Proceedings of International Conference on Machine Learning (ICML)*, 2017.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[28] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[30] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *Proceedings of IEEE symposium on security and privacy*, 2017.

[31] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," *Proceedings of IEEE symposium on security and privacy*, 2016.

[32] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi, "Segment and complete: defending object detectors against adversarial patch attacks with robust patch detection," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[33] P. Qi, T. Jiang, L. Wang, X. Yuan, and Z. Li, "Detection tolerant black-box adversarial attack against automatic modulation classification with deep learning," *IEEE Transactions on Reliability*, vol. 71, no. 2, pp. 674–686, 2022.

[34] Y. Yang, S. Yang, J. Xie, Z. Si, K. Guo, K. Zhang, and K. Liang, "Multi-head uncertainty inference for adversarial attack detection," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[35] Z. Ji, B. Yang, P. L. Yeoh, Y. Zhang, Z. He, and Y. Li, "Active attack detection based on interpretable channel fingerprint and adversarial autoencoder," *Proceedings of IEEE International Conference on Communications*, 2022.

[36] M. Picot, F. Granese, G. Staerman, M. Romanelli, F. Messina, P. Piantanida, and P. Colombo, "A halfspace-mass depth-based method for adversarial attack detection," *Transactions on Machine Learning Research*, 2023.

[37] M. Hussain and J.-E. Hong, "Reconstruction-based adversarial attack detection in vision-based autonomous driving systems," *Machine Learning and Knowledge Extraction*, vol. 5, p. 1589–1611, 2023.

[38] F. Iglesias, J. Milosevic, and T. Zseby, "Fuzzy classification boundaries against adversarial network attacks," *Fuzzy Sets and Systems*, vol. 368, pp. 20–35, 2019.

[39] M. Ranjbar and S. Effati, "A new approach for fuzzy classification by a multiple-attribute decision-making model," *Soft Computing*, vol. 26, p. 4249–4260, 2022.

[40] L. Jara, A. González, and R. Pérez, "A new multi-rules approach to improve the performance of the Chi fuzzy rule classification algorithm," *Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2022.

[41] A. Howard, E. Park, and W. Kan, "Imagenet object localization challenge," *Proceedings of International Journal of Computer Vision (IJCV)*, 2015.

[42] J. Ren, M. Zhang, C. Yu, and Z. Liu, "Balanced MSE for imbalanced visual regression," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[43] P. Qi, T. Jiang, L. Wang, X. Yuan, and Z. Li, "Triangular fuzzification of random variables and power of distribution tests: Empirical discussion," *Computational Statistics Data Analysis*, vol. 51, no. 9, pp. 4742–4750, 2007.

[44] T. J. Ross, "Fuzzy logic with engineering applications," *John Wiley Sons Ltd*, 2004.

[45] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: a critical analysis of out-of-distribution generalization," *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[46] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Master's thesis*, 2009.

[47] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: common objects in context," *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[48] E. Soares, P. Angelov, and N. Suri, "Similarity-based deep neural network to detect imperceptible adversarial attacks," *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022.

[49] M. Tan and Q. V. Le, "EfficientNetV2: smaller models and faster training," *Proceedings of International Conference on Machine Learning (ICML)*, 2021.

[50] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A New Multi-scale Backbone Architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[51] P.-Y. Chen, M.-C. Chang, J.-W. Hsieh, and Y.-S. Chen, "Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection," *IEEE transactions on Image Processing*, vol. 30, pp. 9099–9111, 2021.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[53] C. Sakaridis, D. Dai, and L. V. Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, p. 973–992, 2018.