Non-uniform Illumination Attack for Fooling Convolutional Neural Networks

Akshay Jain, Shiv Ram Dubey, *Senior Member, IEEE*, Satish Kumar Singh, *Senior Member, IEEE*, KC Santosh, *Senior Member, IEEE*, and Bidyut Baran Chaudhuri, *Life Fellow, IEEE*

***Abstract*—Convolutional neural networks (CNNs) have made remarkable strides; however, they remain susceptible to vulnerabilities, particularly to image perturbations that humans can easily recognize. This weakness, often termed as "attacks," underscores the limited robustness of CNNs and the need for research into fortifying their resistance against such manipulations. This study introduces a novel nonuniform illumination (NUI) attack technique, where images are subtly altered using varying NUI masks. Extensive experiments are conducted on widely accepted datasets including CIFAR10, TinyImageNet, CalTech256, and NWPU-RESISC45 focusing on image classification with 12 different NUI masks. The resilience of VGG, ResNet, MobilenetV3-small, InceptionV3, and EfficientNet_b0 models against NUI attacks are evaluated. Our results show a substantial decline in the CNN models' classification accuracy when subjected to NUI attacks, due to changes in the image pixel value distribution, indicating their vulnerability under NUI. To mitigate this, a defense strategy is proposed, including NUI-attacked images, generated through the new NUI transformation, into the training set. The results demonstrate a significant enhancement in CNN model performance when confronted with perturbed images affected by NUI attacks. This strategy seeks to bolster CNN models' resilience against NUI attacks. A comparative study with other attack techniques shows the effectiveness of the NUI attack and defense technique.[1]**

***Impact Statement*—While CNN models demonstrate strong performance on controlled data, their susceptibility to manipulation raises significant concerns about their robustness and suitability for real-world applications, as they can potentially fooled by data perturbation. In this context, we explore NUI masks that manipulate images to deceive CNN models while preserving their semantic content. Additionally, we introduce a key defense strategy involving NUI augmentation during training to enhance CNN model robustness. Given the prevalence of illumination variations in practical computer vision applications, our NUI masks offer a crucial means of bolstering model resilience.**

***Index Terms*—Convolutional neural network (CNN), deep learning, fooling deep models, image categorization, nonuniform illumination (NUI), robustness.**

I. Introduction

DEEP learning, a subfield of artificial intelligence, known for neural networks with multiple interconnected layers, enables the automated extraction of progressively abstract features from input data [1]. It is resurgence in the 2010s was catalyzed by ample data availability, enhanced computational resources, and novel architectures such as convolutional and recurrent networks. Ongoing research in optimization, interpretability, and robustness continues to refine deep learning's efficacy and broaden its applicability across intricate real-world problem domains. The convolutional and recurrent networks made significant advancements in diverse domains including computer vision [2], [3], natural language processing [4], health informatics [5], and sentiment analysis [6]. The convolutional neural networks (CNNs) are utilized for computer vision applications [7], such as image recognition [8], [9], COVID-19 grading [10], image quality assessment [11], image super-resolution [12], and human action recognition [13]. CNN models employ backpropagation to learn the weights [14], [15]. However, if a CNN model is more complex than the dataset and appropriate regularization techniques are not utilized, they are susceptible to overfitting the training data. Common regularization approaches include dropout [16], batch normalization [17], and data augmentation [18].

Recent studies uncovered that the CNN models can be deceived via data perturbation in multiple different ways [19], [20], [21], [22], [23], [24]. To address this issue, many defense methods and network robustness aspects were studied [25], [26], [27]. However, none of them studied the robustness of CNN models against nonuniform illumination (NUI). In this article, we propose mask-based NUI variations as depicted in Fig. 1 to fool the CNN models. Existing methods for adversarial attacks and defense techniques depend on data and the model's

Received 12 September 2023; revised 6 July 2024 and 21 January 2025; accepted 16 February 2025. Date of publication 7 March 2025; date of current version 29 August 2025. This article was recommended for publication by Associate Editor Alexander Gegov upon evaluation of the reviewers' comments. (*Corresponding author: KC Santosh.*)

Akshay Jain, Shiv Ram Dubey, and Satish Kumar Singh are with the Computer Vision and Biometrics Lab, Department of Information Technology, Indian Institute of Information Technology Allahabad, Prayagraj, Uttar Pradesh 211015, India (e-mail: jainvakshay97@gmail.com; srdubey@iiita.ac.in; sk.singh@iiita.ac.in).

KC Santosh is with AI Research Lab, Department of Computer Science, University of South Dakota, Vermillion, SD57069, USA (e-mail: kc.santosh@usd.edu).

Bidyut Baran Chaudhuri, retired, was with Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, 700108, India (e-mail: bidyutbaranchaudhuri@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TAI.2025.3549396, provided by the authors.

Digital Object Identifier 10.1109/TAI.2025.3549396

[1]The code is available at https://github.com/Akshayjain97/Non-Uniform_Illumination

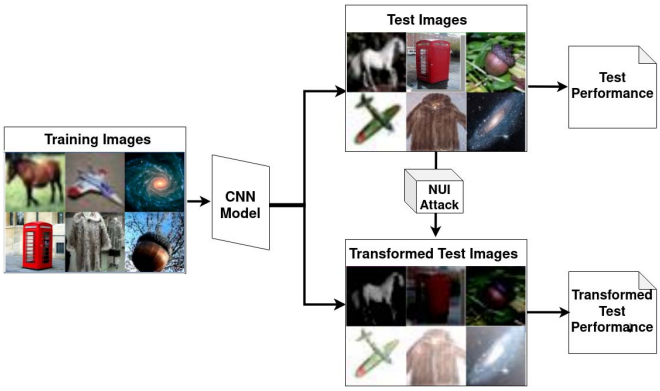

Fig. 1. Overview of the NUI attack to fool the CNN models (image classification): test images are transformed through NUI attacks and their corresponding performance.

gradient. The proposed NUI attack is data-independent and utilizes varying weights of brightness and darkness.

The majority of the techniques to perturb test images have a few drawbacks: prior knowledge of the model and dataset limits their applications in unfamiliar scenarios, and the inability to add NUI variations in the brightness of the images, whereas the NUI attack technique adds nonuniform brightness to the image while keeping the semantic meaning intact. The following are the contributions of this article:

1) The proposed NUI attack produces the attacked images by combining the input image with a NUI mask. Specifically, 12 NUI attack masks are presented.
2) The NUI attack mask is created using several nonlinear transformations generating nonuniform variations of brightness and darkness exploiting the spatial structure of the image.
3) We analyze the robustness of the CNN models including VGG, ResNet, MobilenetV3, InceptionV3, and EfficientNet_b0 over the proposed NUI attack on various benchmark datasets, including CIFAR10, CalTech256, TinyImageNet, and NWPU-RESISC45.
4) We also train the CNN models on the NUI-attacked images to evaluate the robustness of the models when the NUI attack is used as a data augmentation technique.

The remaining article is structured as follows: Section II describes the related work; Section III describes the proposed NUI attack; Section IV describes the experimental settings, datasets, and training settings used; Section V illustrates the experimental results with observations; and Section VI concludes the article.

II. Related Work

This section briefs about the adversarial attacks using brightness and defense mechanisms to such attacks.

A. Adversarial Attacks Using Brightness

Several works have focused on attacking the neural network models by perturbing the intensity values of the image pixels. Nguyen et al. [28] have explored the possibility and practicality of performing real-time physical attacks on face recognition systems using adversarial light projections. Singh et al. [29] have generated adversarial examples using curriculum learning. The natural adversarial lighting conditions are generated by utilizing a physical lighting model proposed by Zhang et al. [30] for conducting an adversarial relighting attack. Given an image, Yang et al. [31] have generated the adversarial examples by applying a brightness transformation to an image and feeding it into a CNN. Hsiung et al. [32] have utilized the componentwise projected gradient descent and automatic attack-order scheduling to find the optimal attack composition for creating the composite adversarial examples.

Most existing methods require a neural network to generate adversarial examples. The color channel perturbation (CCP) attack perturbs the channels of images to generate the mixed color channels randomly [33]. The impact of color is also studied in [34] on the robustness of deep learning models. The article aims to judge the robustness of CNN models against various NUI variations generated through different masks. The proposed method is data-independent, does not require any neural network and gives a high attack success rate.

B. Defense Against Brightness Attacks

The primary defense mechanism employed by most methods includes the attacked samples in the training set through data augmentation and retrains the model. A survey of defense strategies is presented in [35]. Agarwal et al. [36] have exploited the image transformations, including discrete wavelet transform and discrete sine transform, against adversarial perturbation using deep models. The performance of CNN models on CCP-attacked images greatly improved when the models were trained on the training set containing the CCP-attacked samples [33]. The adversarial examples generated in [29] are designed to be resilient against variations in real-world brightness conditions. Agarwal et al. [37] have developed an adversarial perturbation detector agnostic to databases, attacks, and models. Adversarial visual reconstruction is used against DeepFakes in [38]. Hsiung et al. [32] have performed the generalized adversarial training (GAT) to enhance the robustness of the model against composite semantic perturbations, including combinations of hue, saturation, brightness, contrast, and rotation. Recently, a self-supervised defense mechanism has been utilized in [39] against adversarial face images. Premakumara et al. [40] have systematically investigated the amount of artificial perturbation needed to enhance the models' generalization by augmenting the data for object detection using neural networks. We propose a primary defense mechanism against the NUI attack by employing data augmentation through NUI attack in the training set and retraining the CNN models for the image classification task. The proposed defense technique can be useful in common use cases where the input image gets distorted due to exposure to sunlight or part of the image becomes relatively darker because of reflection.

TABLE I
LIST OF ALL MASKS AND THEIR REGION OF PERTURBATION

Mask ID	Mask	Region of Perturbation
Mask 1	$a = ((u-x)\times\frac{30}{u}) + ((v-y)\times\frac{30}{v}) + ((u-y)\times\frac{20}{u}) + ((v-y)\times\frac{20}{v})$	Focused more on the left side
Mask 2	$a = (x\times\frac{30}{u}) + ((v-y)\times\frac{30}{v}) + (y\times\frac{20}{u}) + ((v-x)\times\frac{20}{v})$	Distributed throughout
Mask 3	$a = ((u-x)\times\frac{30}{u}) + (y\times\frac{30}{v}) + ((u-y)\times\frac{20}{u}) + (y\times\frac{20}{v})$	Focused on the top right corner
Mask 4	$a = (x\times\frac{30}{u}) + (y\times\frac{30}{v}) + (x\times\frac{20}{u}) + (y\times\frac{20}{v})$	Focused on the bottom right corner
Mask 5	$a = \text{abs}(16-x)\times\text{abs}(16-y)$	The curved diamond shape
Mask 6 Mask 7 Mask 8	$a = 144 - \text{abs}(16-x)\times\text{abs}(16-y)$ $a = 100 - \text{abs}(16-x)\times\text{abs}(16-y)$ $a = 50 - \text{abs}(16-x)\times\text{abs}(16-y)$	Circular perturbation with different radius at centre
Mask 9	**if**$(0\le y\le 5\ or\ 10\le y\le 15\ or\ 20\le y\le 25\ or\ 30\le y\le 32) : a =$ Mask 1 **else** : $a = -$Mask 2	A pattern of vertical lines
Mask 10	**if**$(0\le x\le 5\ or\ 10\le x\le 15\ or\ 20\le x\le 25\ or\ 30\le x\le 32) : a =$ Mask 1 **else** : $a = -$Mask 2	A pattern of horizontal lines
Mask 11	**if**$(x\le 16\ and\ y\le 16) : a =$ Mask 1 **if**$(x\le 16\ and\ y> 16) : a =$ Mask 2 **if**$(x> 16\ and\ y\le 16) : a =$ Mask 3 **if**$(x> 16\ and\ y> 16) : a =$ Mask 4	Differs for different quadrants of the image
Mask 12	**if**$(x\le 16\ and\ y\le 16) : a =$ Mask 1 **if**$(x\le 16\ and\ y> 16) : a =$ Mask 2 **if**$(x> 16\ and\ y\le 16) : a = -$Mask 3 **if**$(x> 16\ and\ y> 16) : a = -$Mask 4	Differs for different quadrants of the image and produces a pattern effect of vertical lines

Note: x and y are the horizontal and vertical coordinate axis variables, respectively, and u and v represent the image size which remained constant (i.e., 32) throughout the experiment. a is the amount of brightness to be added to the input image.

III. PROPOSED NUI ATTACK

In recent years, various attack methods have been investigated to judge the robustness of CNN models. However, the conventional attack methods do not take advantage of creating NUI variations with different brightness and darkness levels.

A. Proposed NUI Attacks

We propose a simple yet effective NUI attack on test image data. The rationale behind developing this attack technique stemmed from a desire to investigate perturbation methods applicable to CNN models which can give a high attack success rate and do not require any neural network to model such attack. Specifically, the aim is to explore how illumination variations could be utilized to attack these models. In the earlier stages of the experiments we considered only Masks 1 to 4, but later to experiment with the region of attack, we added Masks 5 to 12 given in Table I. The proposed NUI attack brightens or darkens the image pixels nonuniformly to generate the synthesized test images to fool the CNN models. The core of the proposed attack is the weight of image brightness and darkness. The weight (k) value controls the brightness or darkness added to the test image based on certain patterns. The proposed attack technique uses several masking strategies to generate different masks (a) for the images of size $h \times w$, where h and w are image height and width, respectively. The created masks are applied to the test images to generate the synthesized test images to fool the CNN models. In this article, we experiment using 12 different masks. We analyzed the robustness of CNN models on the Attacks caused by different NUI masks. The formulas utilized to create these masks (a) are given in Table I with its region of perturbation in the image. There are a total of 23 different weight values k used in this article, ranging from -2.2 to $+2.2$ with a gap of 0.2. It leads to $23 \times 12 = 276$ experiments for a given model on any dataset. The extremes of k were selected to the limit where images remain recognizable to humans, with a 0.2 gap to capture accuracy differences. Higher k values increase image perturbation, leading to greater attack success.

The masking function, Mask 1, is considered from [41]. Masks 2–4 are the variations of Mask 1 and are formulated by considering the exploitation of spatial locality. Mask 5 perturbs the image center up to the centers of each side in the shape of a curved diamond. The effect of Masks 6–8 is similar, but with different severity. These masks create a circular perturbation effect in the images. The amount of perturbation is highest for Mask 6 and lowest for Mask 8. Masks 9 and 10 use Mask 1 and negative of the Mask 2 in specific conditions leading to perturbation of the pattern of vertical and horizontal lines, respectively. Mask 11 adds perturbations of Masks 1–4 in different quadrants. The effect of the Mask 12 is similar to Mask

Algorithm 1: Proposed NUI Attack Algorithm.

Input: Image data I, Mask ID i and weight value k

Image $I \in \mathbb{R}^{u \times v}$, Mask $i \in \{1, 2, \ldots, 12\}$
weight $k \in \{-2.2, -2.0, \ldots, 2.0, 2.2\}$

Output: NUI Attacked Image $I_{M_i,k}$.

1: Generate the i^{th} Mask using Table 1 as $M_i = a(x, y), \forall x \in \{1, u\}$ and $\forall y \in \{1, v\}$.
2: Perform the mask weighting with weight k as $M_{i,k} = M_i \times k$.
3: Attack the image (I) to generate the perturbed image $(I_{M_i,k})$ as $I_{M_i,k} = I + M_{i,k}$.

11, except for the right part of the image which becomes darker instead of brighter.

The algorithm for the proposed NUI attack is illustrated in Algorithm 1. The input image (I) is attacked to purturbed image ($I_{M_i,k}$) using the ith Mask and weight value k. As shown in Table I, 12 NUI Masks are used in this article. Based on the chosen Mask and weight, the final Mask is computed and added in the input image to generate the attacked image.

B. Effect of NUI Attacks

The effect of different NUI attacks is illustrated in Fig. 2 using the sample images as to how the brightness, color, details, appearance, and so on change after applying different NUI masks. Here, the perturbation weight (k) value is different for all columns and is positive, because of which all the images look brighter than their original form. The first column contains the original sample images. The second to thirteenth columns correspond to the images generated using Masks 1 to 12, respectively. As mentioned, the perturbed image is brighter on the left side and the perturbation drops when it goes to the right for Mask 1. The image is bright in general for Mask 2. The images appear bright in the top right corner for Mask 3. The perturbations are focused more in the bottom right corner for the masking function 4. These masking functions are simple and do not change the underlying semantic meaning of the input image, but can provide a good attack success rate. The effect of a curved diamond can be observed for the Mask 5. The perturbations for Masks function 6–8, respectively, produce samples like the reverse of the Mask 5. The images produced using Mask 6 are perturbed with higher intensity values. However, the amount of perturbation is reduced for Mask 7 which is further reduced for Mask 8. Moreover, the attack success increases for Mask 8 without losing the visual perceptibility of the image. The perturbations caused by Masks 9 and 10, respectively, have vertical and horizontal patterns of alternate brightness and darkness. Masks 11 and 12 perturb the images using different masks in different quadrants. Mask 11 adds mask value in each quadrant, while Mask 12 adds mask value in the left quadrants and subtracts in the right quadrants. The effect on the histogram is shown in Supplementary Material.

Fig. 2. First column contains original images, second to thirteenth columns contain the images perturbed using mask first to twelfth, respectively. Images are taken from CIFAR10, TinyImageNet, and CalTech256 datasets.

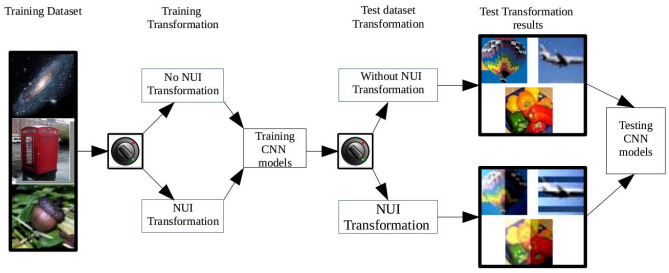

Fig. 3. Workflow of the proposed method and the experimental settings used for the training and testing of the CNN models using NUI attack.

C. Proposed Workflow Using NUI Attacks

The workflow of the proposed method is illustrated in Fig. 3. To analyze the robustness of the CNN models against NUI attacks, we trained models on the original datasets and tested them for all NUI masks for all values of (k). Further to analyze the defense capability, the CNN models are trained on the NUI-attacked datasets and again tested. For training models on perturbed datasets, the NUI perturbation is added to 80% of the training set. We limit the weight factor (k) in the training part to 12 different settings to avoid high bias in the training set toward severe perturbation, i.e., from -1.2 to $+1.2$ with a gap of 0.2 excluding 0.0 as it is already included in the 20% part of the training set. The number of masks for perturbation during training is reduced to 10 only, excluding Masks 6 and 7 as these are similar to Mask 8. Mask 12 is replaced with the following mask for training:

$$\begin{aligned}
&\textbf{if}(x \leq 16\, and\, y \leq 16)\colon a = +\ \text{Mask } 1\\
&\textbf{if}(x \leq 16\, and\, y > 16)\colon a = -\ \text{Mask } 2\\
&\textbf{if}(x > 16\, and\, y \leq 16)\colon a = +\ \text{Mask } 3\\
&\textbf{if}(x > 16\, and\, y > 16)\colon a = -\ \text{Mask } 4
\end{aligned}$$

which subtracts Masks 2 and 4 in the leading diagonal quadrants, respectively, and adds Masks 1 and 3 in the other two quadrants, respectively. This represents the general case for quadrant perturbation. After being trained on perturbed images, the CNN models not only preserved the original accuracy on unperturbed data but also became robust to NUI attacks.

IV. Experimental Settings

A. Datasets

To examine the impact of the proposed NUI attacks, we conduct the image classification experiments on three benchmark datasets, including CIFAR10 [42], CalTech256 [43], TinyImageNet [44], and NWPU-RESISC45 [45]. The 60 000 images in the CIFAR10 dataset are equally divided into 10 different categories. Out of 60 000 images 10 000 images are marked as the test set and the rest as the training set. The 30 607 images in the CalTech256 dataset represent 257 different object categories. 20% of the CalTech256 dataset is utilized for testing, while the rest for training. The CalTech256 dataset exhibits a high level of complexity due to several categories and more instances within each category, it also exhibits high interclass similarity. The training set of the TinyImageNet dataset contains 100 000 images and the validation set consists of 10 000 images. The dataset comprises 200 categories which have 500 training images and 50 validation images for each category. It consists of a subset of images from ImageNet, specifically curated for small-scale experiments.

B. CNN Architectures Used

We used VGG [46], ResNet [47], MobilenetV3 [48], InceptionV3 [49], and EfficientNet_b0 [50] to demonstrate the effects of the proposed NUI attack. The VGG network is a deep CNN model containing 16 or 19 trainable layers. The principal thought behind the VGG network is to utilize a series of convolutional layers with small filter sizes (3×3) and stack them together to create a deeper network. For experiments on the CIFAR10 and TinyImageNet datasets, VGG16 is used and for experiments on the CalTech256 dataset, VGG19 is used. The ResNet model includes the residual connections that allow the flow of gradients during backpropagation effectively. Deep CNNs utilizing the residual model demonstrate improved convergence, leading to enhanced performance. The ResNet18 model is used with all the datasets for experiments. MobileNetV3 is a CNN specifically optimized for mobile phone CPUs through a combination of hardware-aware network architecture search (NAS). This network has been further refined through several innovative architectural improvements, including integrating complementary search methodologies, developing new efficient nonlinearities suitable for mobile environments and creating efficient network design tailored for mobile applications. Inception-v3 represents an advanced CNN architecture within the Inception series, incorporating several enhancements. These include Label Smoothing, factorized 7×7 convolutions, and the integration of an auxiliary classifier to propagate label information to earlier network layers with the implementation of batch normalization within the auxiliary head layers. EfficientNet, a CNN family for image classification, uses compound scaling to balance depth, width, and resolution. This design ensures computational efficiency and adaptability, allowing scalability while preserving accuracy. The Cifar10 dataset has been used for experimentation with MobilenetV3-small and InceptionV3.

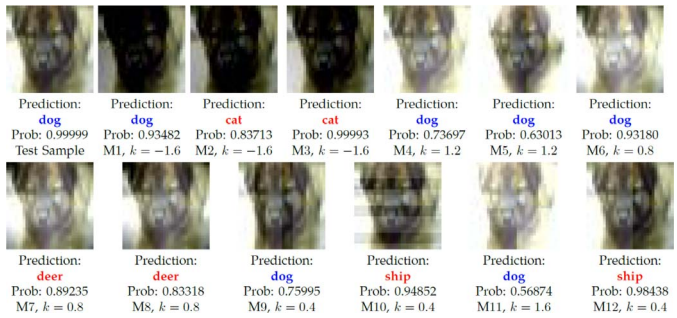

Fig. 4. Predictions of ResNet18 for an original test image and NUI-attacked images using different masks with varying weight (k). "Prob" refers to probability and Mi refers to ith mask.

C. Training Settings

All the experiments are performed using the PyTorch framework [51]. The batch size of 64 is used for VGG and ResNet models, 256 for MobileNet model and 128 for the Inception model. Using the Adam optimizer, the models are trained for 100 epochs. For the first 80 epochs, the learning rate is set at 10^{-3} for CIFAR10 and TinyImageNet and 10^{-4} for the CalTech256 dataset, and for the final 20 epochs, it is reduced by a factor of 10. The categorical cross-entropy loss function is used as an objective function to measure the dissimilarity between predicted and actual class labels. Batch normalization is used for regularization. The following data augmentation is used during training: random cropping of size 32, random horizontal flipping, and normalization to zero mean and unit standard deviation. The images are also resized to 32×32 resolution for VGG and ResNet models, whereas the MobileNet and Inception models accept images of size 224×224 and 299×299, respectively.

V. Experimental Results and Analysis

In this section, the qualitative and quantitative results are presented for image classification using VGG and ResNet models on CIFAR10, TinyImageNet and CalTech256 datasets as well as MobileNet and InceptionV3 models on CIFAR10 dataset.

A. Qualitative Results

The visual results for a sample image from the CIFAR10 dataset under different NUI attacks are shown using the ResNet18 model in Fig. 4 following the predicted category with the probability of classification. The first image in the first row is an original dog image taken from the CIFAR10 dataset, and the model predicts it as a dog with very high probability. The second to seventh images in the first row and the first to sixth images in the second row represent samples generated using the first to twelfth mask in the same order along with its predicted category with probability. Different values of NUI weight (k) are used with different masks. Note that when for negative k, the resultant image becomes darker and vice-versa. The images are misclassified with high probability under NUI attacks with second, third, seventh, eighth, tenth, and twelfth masks. Whereas, the probability of classification to correct class

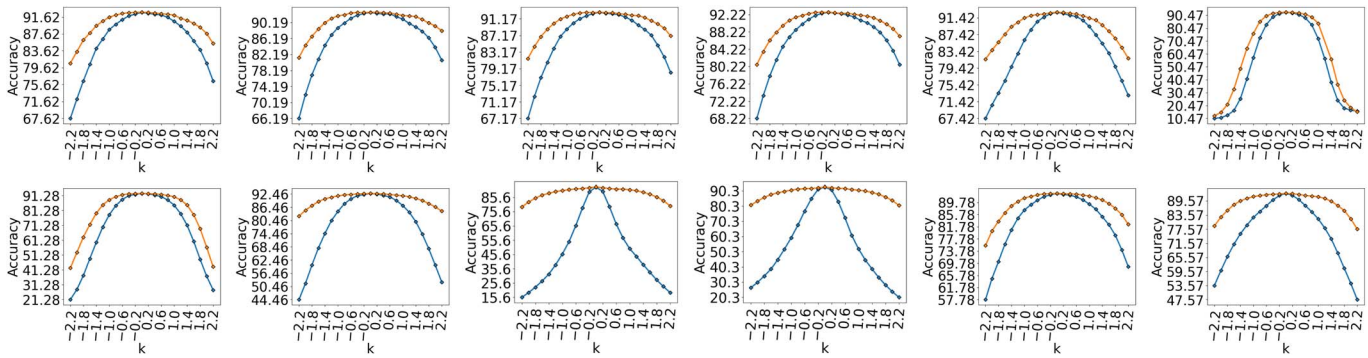

Fig. 5. Results of VGG16 on CIFAR10 dataset under different NUI attacks (test set). Blue and orange curves show the performance of the model trained on the original training set and the NUI perturbed training set, respectively.

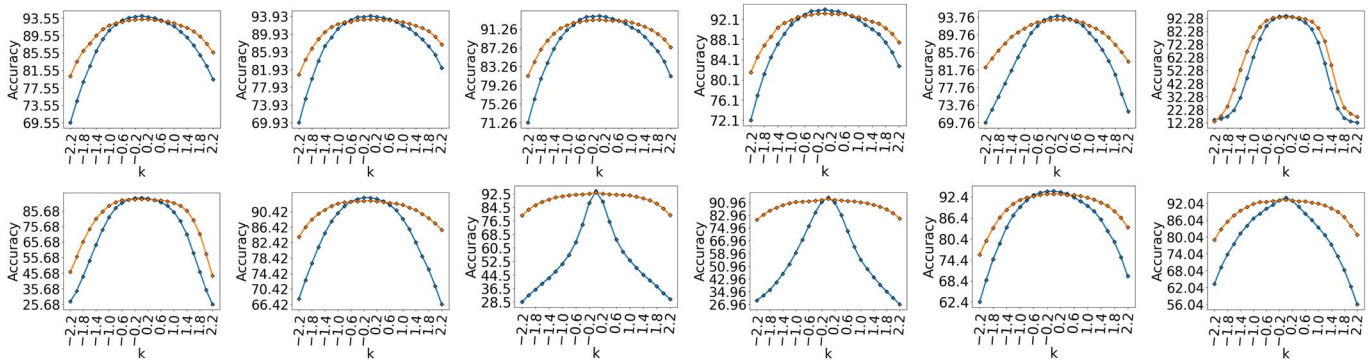

Fig. 6. Results of ResNet18 on CIFAR10 dataset under different NUI attacks (test set). Blue and orange curves show the performance of the model trained on the original training set and the NUI perturbed training set, respectively.

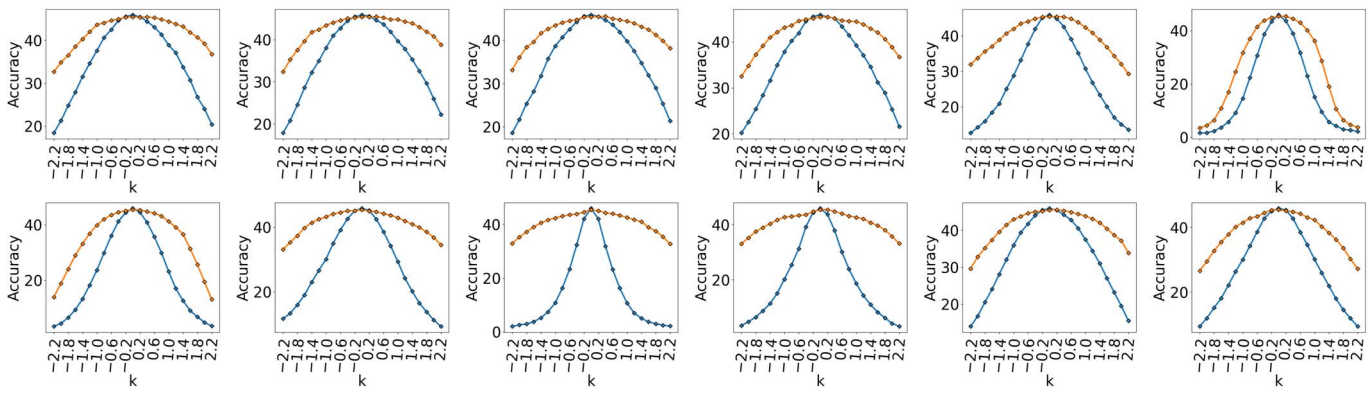

Fig. 7. Results of VGG16 on TinyImageNet dataset under different NUI attacks (test set). Blue and orange curves show the results of the model trained on the original training set and the NUI perturbed training set, respectively.

is decreased under other NUI attacks. It is evident from these results that almost all the images are visually perceptible to the original image with some amount of brightness or darkness, however, these images are either misclassified by a trained CNN model or confidence of classification decreases. We refer to the Supplementary Materials to observe the impact of the NUI attack on image pixel value distributions.

B. Quantitative Results

The goal of this study was to evaluate the robustness of CNN models under NUI attacks, on different datasets. After conducting several experiments, we have recorded a substantial drop in the accuracy of CNNs on all datasets. Fig. 5 shows the performance of VGG16 over CIFAR10 under different NUI attacks on the test set, similarly, Figs. 6–10 show the performance curve for different CNNs over different datasets. The plots are reported in blue color when the models are trained on the original training set and in orange color on the augmented training set with NUI transformations. Each figure contains 12 subfigures corresponding to NUI attacks with first to twelfth masks in the order of first row from left to right for first to sixth masks and second row from left to right for seventh to twelfth masks, respectively. The x-axis and y-axis represent

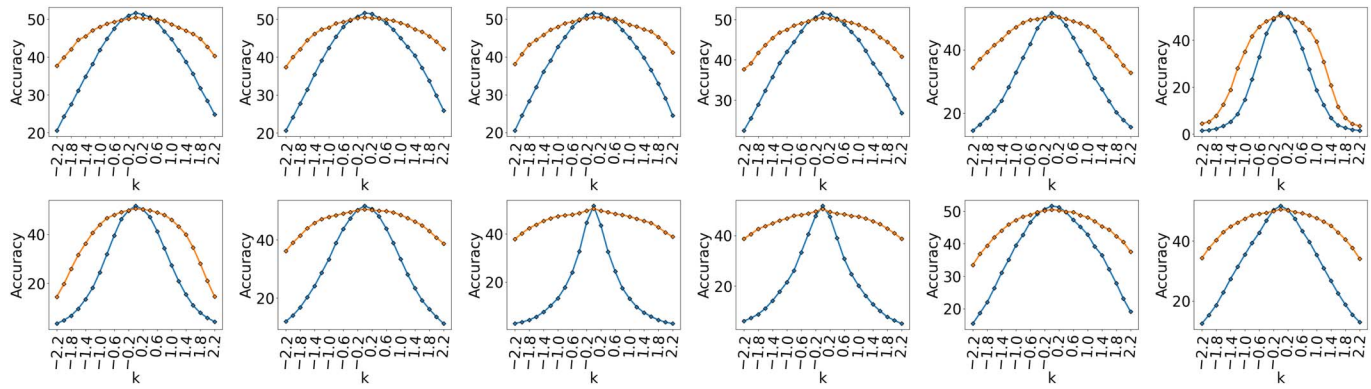

Fig. 8. Results of ResNet18 on TinyImageNet dataset under different NUI attacks (test set). Blue and orange curves show the results of the model trained on original training set and NUI perturbed training set, respectively.

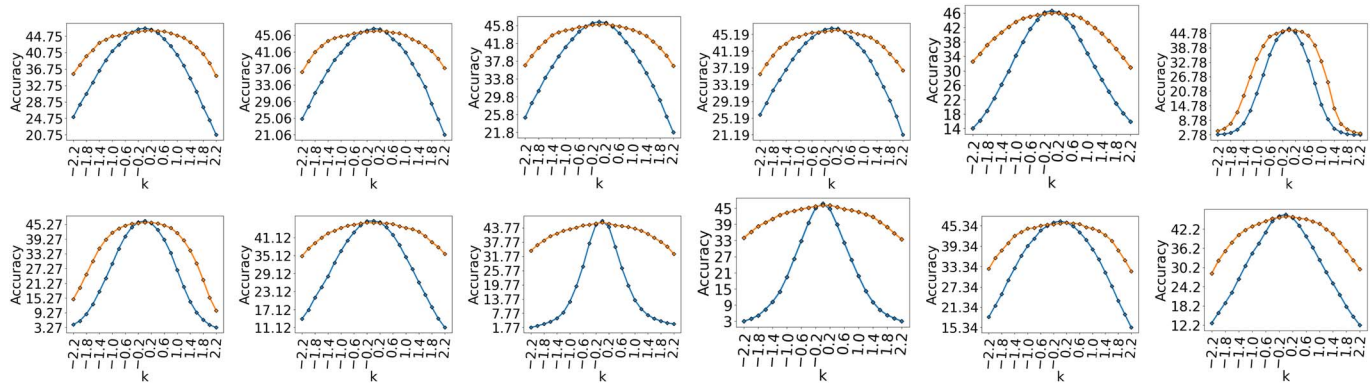

Fig. 9. Results of VGG19 on CalTech256 dataset under different NUI attacks (test set). Blue and orange curves show the results of the model trained on the original training set and the NUI perturbed training set, respectively.

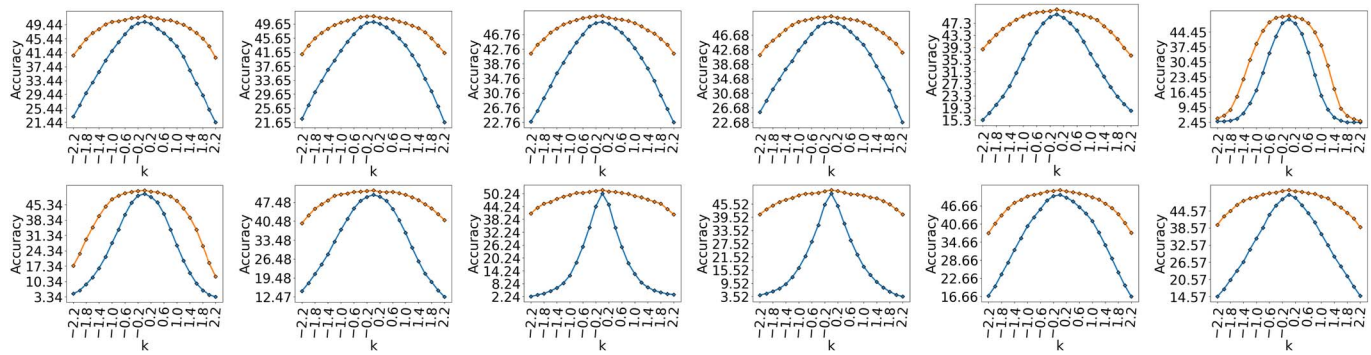

Fig. 10. Results of ResNet18 on CalTech256 dataset under different NUI attacks (test set). Blue and orange curves show the results of the model trained on the original training set and NUI perturbed training set, respectively.

different NUI weights (k) and accuracy (%), respectively. Note that $k = 0$ indicates no attack. From these plots, it is clear that the performance of the CNN models decreases on the NUI-attacked test sets. However, the performance is enhanced by including the NUI attack-based augmentation during training. It is also observed that the accuracy of the CNN models decreases as the weight (k) of the NUI attack moves toward extreme positive or negative values. We can observe that the curve for a mask remains similar for a particular dataset irrespective of the model used. It depicts the generalizability of proposed NUI attacks for different CNN models. The performance of a particular mask on a dataset also depends on the number of classes. If the number of classes is less, the probability of correctly classifying a test image is high as compared to a dataset with more classes.

The blue curves show that the CNN models are not robust against the NUI attacks as these models get fooled by the perturbed images. The sixth, seventh, nineth, and tenth masks lead to a very high impact on the performance degradation of the CNN models. The poor performance of the models for Mask 6 is due to severe circular perturbation which leads to

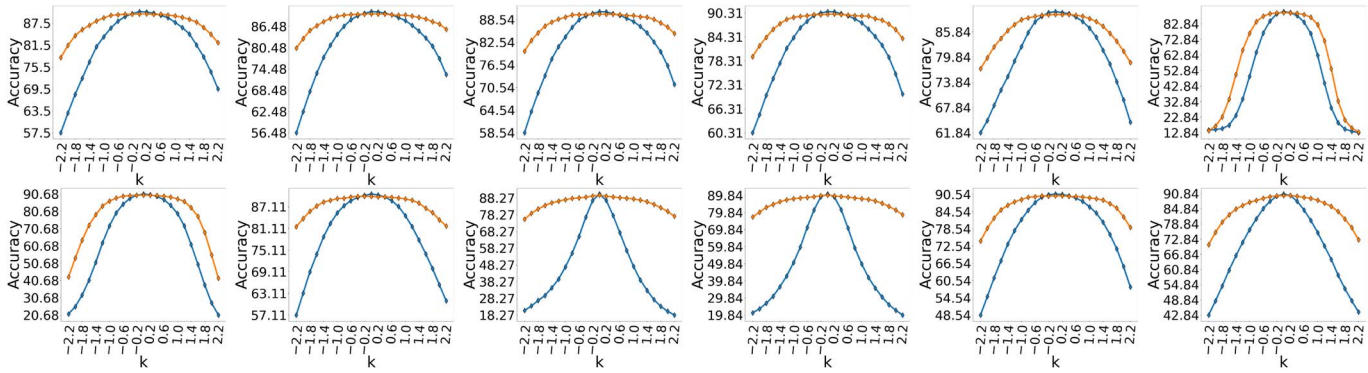

Fig. 11. Results of MobilenetV3-small on CIFAR10 under different NUI attacks (test set). Blue and orange curves show the results of the model trained on the original training set and the NUI perturbed training set, respectively.

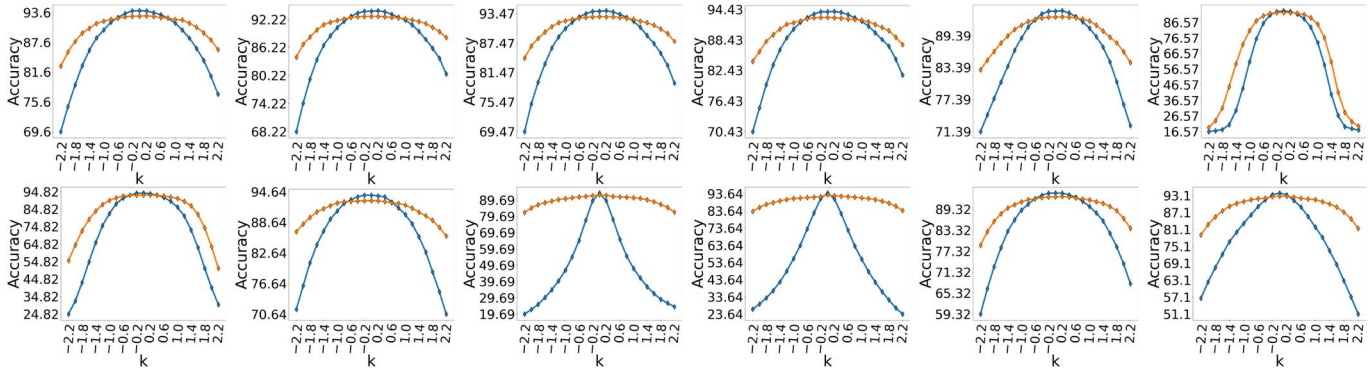

Fig. 12. Results of InceptionV3 on CIFAR10 dataset under different NUI attacks (test set). Blue and orange curves show the results of the model trained on the original training set and the NUI perturbed training set, respectively.

TABLE II
DECREASE (IN PERCENT) IN ACCURACY OF MODELS ON VARIOUS TEST SETS UNDER NUI ATTACK FOR $k = -1.4$

Model	Dataset	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
VGG16	CIFAR10	9.19	8.91	9.33	8.5	13.9	72.29	35.47	19.63	65.63	51.44	12.93	18.49
VGG16	TinyImageNet	31.22	29.74	30.7	31.09	54.4	87.3	71.17	49.9	88.61	75.31	38.72	51.72
VGG19	CalTech256	22.12	21.56	21.66	21.17	43.86	83.8	61.71	39.58	87.9	78.26	29.45	43.02
ResNet18	CIFAR10	8.55	7.23	7.4	7.2	13.21	66.7	32.07	9.75	54.65	50.93	11	13.55
ResNet18	TinyImageNet	32.53	31.4	30.01	30.67	53.67	89.89	73.89	53.29	85.43	72.57	39.08	46.97
ResNet18	CalTech256	28.47	27.09	28.04	25.76	47.17	86.72	67.27	43.75	87.57	81.78	36.17	46.84
Mobilenet	CIFAR10	15.10	14.22	13.84	14.23	17.04	73.59	43.63	13.07	61.91	59.87	19.21	26.87
Inception	CIFAR10	8.45	7.79	8.02	7.99	11.00	68.02	30.10	7.48	63.24	54.98	11.63	18.32

the complex generated images. Mask 7 is similar to Mask 6 but with reduced complexity. Still, the complexity of images generated by Mask 7 is very high to fool the CNN models. Masks 9 and 10 add perturbation as a pattern in the horizontal and vertical directions, respectively. Using these NUI attacks, the images after adding the mask are still visually perceptible, however, the performance of CNN models has significantly dropped. Moreover, only a small value of k can produce a powerful NUI attack with high fooling rate using Masks 9 and 10. Red curves depict there has been considerable improvement in the performance of the CNN models after being trained on NUI-augmented training data. We exclude Masks 6 and 7 in the training set, hence the improvement after NUI augmentation is low under these attacks on the test set.

Table II summarizes the percentage reduction in the accuracy of the CNN models under different NUI attacks for $k = -1.4$ w.r.t. without attack. A high attack success rate is achieved using Masks 6, 7, 9, and 10. TinyImageNet images are more prone to heavy perturbation using NUI attacks as depicted by the highest performance drop among all the datasets. The success rate of attack is higher for datasets for which the number of classes is large as the perturbation creates more confusion in class probabilities. Table III summarizes the percentage increase in the accuracy of the CNN models after being trained on the NUI perturbed dataset. The percentage improvement in the performance is calculated on the model's performance on the NUI attack and the model's performance after being trained on the NUI perturbed dataset. If a model on a particular dataset has a higher percentage reduction in Table II then in most of such cases a higher percentage increase is observed in the model's performance on the same dataset in Table III. Masks 9 and 10 lead to the highest increment when trained on the NUI

TABLE III
INCREASE (IN PERCENT) IN ACCURACY ON VARIOUS TEST SETS FOR $k = -1.4$, AFTER APPLYING THE PROPOSED DEFENSE TECHNIQUE

Model	Dataset	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
VGG16	CIFAR10	6.36	6.73	7.14	5.7	11.9	91	32.65	21.17	177.3	97.66	9.77	18.4
VGG16	TinyImageNet	27.64	29.73	31	19.74	85.83	193.7	150.5	80.66	675.8	256	40.8	71
VGG19	CalTech256	18.81	19.4	19.27	16.8	55.21	149.4	98.25	51.18	632.7	304.2	28.92	51.87
ResNet18	CIFAR10	4.11	3.59	3.75	3.75	9.42	68.7	26.18	7.12	108.8	93.73	6.82	9.97
ResNet18	TinyImageNet	30.4	30.26	26.84	26.92	80.85	262.4	169.2	90	503.4	215.8	41.82	63.62
ResNet18	CalTech256	33.71	31.85	33.81	29	77	237	144.1	70.44	664.1	418.5	48.5	79.98
Mobilenet	CIFAR10	13.04	12.95	12.55	12.65	14.58	109.65	54.09	12.26	147.37	136.82	18.42	27.81
Inception	CIFAR10	5.21	5.13	7.62	4.8	6.86	98.86	27.44	5.02	157.46	111.51	7.99	16.32

perturbed dataset. The readings also indicate that using NUI transformation as data augmentation is an effective technique and results in considerable performance improvements on NUI-attacked test sets.

We also provide the details of Mask parameters, histogram analysis on effect of NUI attacks, t-SNE plot feature analysis, precision, recall and F1-score results, comparison with SOTA, details of computational aspects and analysis on effect of NUI Attack on color channels in Supplementary Material.

VI. CONCLUSION

In this research, we introduce NUI attacks to study the robustness of the CNN models. The proposed NUI attacks can deceive the CNN models for image classification. The attack is simple and data-independent. It leverages the pixel value with spatial information to create different masks that are added to the original image with a weight factor to generate perturbed images. The images generated using NUI attacks retain their semantic significance. Through extensive experimentation using VGG and ResNet models on CIFAR10, TinyImageNet, and CalTech256 datasets as well as MobilenetV3-small and InceptionV3 models on CIFAR10 dataset and EfficientNet_b0 on NWPU-RESISC45, we observe a significant decline in classification performance across all the NUI-attacked test sets. Notably, several samples that were correctly classified with high confidence in the original test set, were incorrectly classified with high confidence after undergoing the NUI attack. With the help of t-SNE graphs and image histograms as depicted in Supplementary Material, we observed the change in pixel value distribution, causing misclassification. The proposed NUI attack is also utilized as a data augmentation during training as a primary defense mechanism and to make the models resilient against such attacks. We have compared the effectiveness of the defense technique using accuracy, precision, recall and $F1-$score evaluation metrics and found the defense technique highly effective against NUI attack. A comparative study of NUI with other attack and defense techniques is also demonstrated in Supplementary Material, documenting the effectiveness of the NUI attack and defense technique such that NUI-based defense can be useful for high exposure and low light surveillance. We have also observed the effects of the NUI attack on different color channels, detailed in Supplementary Material, as a potential future research scope.

CONFLICT OF INTEREST

We declare there is no conflict of interest.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[2] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "Deep learning advances in computer vision with 3D data: A survey," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–38, 2017.
[3] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.
[4] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.
[5] D. Ravì et al., "Deep learning for health informatics," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.
[6] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. 28th Int. Conf. on Mach. Learn.*, 2011, pp. 513–520.
[7] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.
[8] D. Dai, Z. Zhuang, J. Wei, S. Xia, Y. Li, and H. Zhu, "Random sharing parameters in the global region of convolutional neural network," *IEEE Trans. Artif. Intell.*, vol. 3, no. 5, pp. 738–748, Oct. 2022.
[9] S. R. Dubey, S. S. Basha, S. K. Singh, and B. B. Chaudhuri, "Adainject: Injection-based adaptive gradient descent optimizers for convolutional neural networks," *IEEE Trans. Artif. Intell.*, vol. 4, no. 6, pp. 1540–1548, Dec. 2023.
[10] C. de Vente et al., "Automated COVID-19 grading with convolutional neural networks in computed tomography scans: A systematic comparison," *IEEE Trans. Artif. Intell.*, vol. 3, no. 2, pp. 129–138, Apr. 2022.
[11] Z. Pan, F. Yuan, X. Wang, L. Xu, X. Shao, and S. Kwong, "No-reference image quality assessment via multibranch convolutional neural networks," *IEEE Trans. Artif. Intell.*, vol. 4, no. 1, pp. 148–160, Feb. 2023.
[12] A. Esmaeilzehi, M. O. Ahmad, and M. Swamy, "Ultralight-weight three-prior convolutional neural network for single image super resolution," *IEEE Trans. Artif. Intell.*, vol. 4, no. 6, pp. 1724–1738, Dec. 2023.
[13] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph convolutional neural network for human action recognition: A comprehensive survey," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 128–145, Apr. 2021.
[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2014.
[15] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Adanorm: Adaptive gradient norm correction based optimizer for CNNs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5284–5293.
[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. on Mach. Learn.*, 2015, pp. 448–456.
[18] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 113–123.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015.
[20] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.
[21] G. Elsayed et al., "Adversarial examples that fool both computer vision and time-limited humans," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
[22] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.
[23] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2020, pp. 1–10.
[24] I. Singh, T. Araki, and K. Kakizaki, "Powerful physical adversarial examples against practical face recognition systems," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 301–310.
[25] R. Wang, et al., "Fakespotter: A simple yet robust baseline for spotting AI-synthesized fake faces," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2021, pp. 3444–3451.
[26] M. Ren, Y. Zhu, Y. Wang, and Z. Sun, "Perturbation inactivation based adversarial defense for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2947–2962, 2022.
[27] N. Akhtar, J. Liu, and A. Mian, "Defense against universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3389–3398.
[28] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, "Adversarial light projection attacks on face recognition systems: A feasibility study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 814–815.
[29] I. Singh, S. Momiyama, K. Kakizaki, and T. Araki, "On brightness agnostic adversarial examples against face recognition systems," in *Proc. Int. Conf. Biometrics Special Interest Group*, Piscataway, NJ, USA: IEEE, 2021, pp. 1–5.
[30] Q. Zhang, Q. Guo, R. Gao, F. Juefei-Xu, H. Yu, and W. Feng, "Adversarial relighting against face recognition," 2021, *arXiv:2108.07920*.
[31] B. Yang, K. Xu, H. Wang, and H. Zhang, "Random transformation of image brightness for adversarial attack," *J. Intell. Fuzzy Syst.*, vol. 42, no. 3, pp. 1693–1704, 2022.
[32] L. Hsiung, Y.-Y. Tsai, P.-Y. Chen, and T.-Y. Ho, "Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 24658–24667.
[33] J. Kantipudi, S. R. Dubey, and S. Chakraborty, "Color channel perturbation attacks for fooling convolutional neural networks and a defense against such attacks," *IEEE Trans. Artif. Intell.*, vol. 1, no. 2, pp. 181–191, Oct. 2020.
[34] K. De and M. Pedersen, "Impact of colour on robustness of deep neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 21–30.
[35] X. Zhang, X. Zheng, and W. Mao, "Adversarial perturbation defense on deep neural networks," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 8, pp. 1–36, 2021.
[36] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Image transformation-based defense against adversarial perturbation on deep learning models," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 5, pp. 2106–2121, Sep./Oct. 2020.
[37] A. Agarwal, G. Goswami, M. Vatsa, R. Singh, and N. K. Ratha, "DAMAD: Database, attack, and model agnostic adversarial perturbation detector," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3277–3289, Aug. 2022.
[38] Z. He, W. Wang, W. Guan, J. Dong, and T. Tan, "Defeating deepfakes via adversarial visual reconstruction," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 2464–2472.
[39] D. Deb, X. Liu, and A. K. Jain, "Faceguard: A self-supervised defense against adversarial face images," in *Proc. IEEE 17th Int. Conf. Autom. Face Gesture Recognit. (FG)*, Piscataway, NJ, USA: IEEE, 2023, pp. 1–8.
[40] N. Premakumara, B. Jalaian, N. Suri, and H. Samani, "Enhancing object detection robustness: A synthetic and natural perturbation approach," 2023, *arXiv:230410622*.
[41] S. R. Dubey, S. K. Singh, and R. K. Singh, "A multi-channel based illumination compensation mechanism for brightness invariant image retrieval," *Multimedia Tools Appl.*, vol. 74, pp. 11223–11253, Dec. 2015.
[42] A. Krizhevsky, "Learning multiple layers of features from tiny images," Master's thesis, Univ. of Toronto, Toronto, ON, Canada, 2009.
[43] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Pasadena: Technical Report 7694, California Inst. of Technol., vol. 10, Apr. 2007.
[44] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," *CS 231n*, vol. 7, p. 7, Jun. 2015.
[45] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[48] A. Howard et al., "Searching for mobilenetv3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016., pp. 2818–2826.
[50] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946
[51] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.