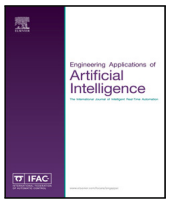
Engineering Applications of Artificial Intelligence 159 (2025) 111619

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

An innovative contrastive learning approach to improve image recognition robustness and interpretability via simulated environmental perturbations

Leijun Cheng [a],[1], Xihe Qiu [a],[*],[1], Xiaoyu Tan [b],[1], Haoyu Wang [a], Yujie Xiong [a]

[a] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China
[b] INFLY TECH(Shanghai)Co.,Ltd, Shanghai, 200030, China

ARTICLE INFO

Keywords:
Simulated environments
Model Robustness
Image noise
Contrastive learning
Image Processing

ABSTRACT

In the field of pattern recognition, the noise inherent in real-world images poses a significant challenge to traditional image processing methodologies. While existing approaches have made progress in addressing this issue, they often struggle with limited model generalization, data distribution shifts, and domain adaptability discrepancies between simulated environments and real-world contexts, compromising efficiency and robustness. In this paper, we propose a novel contrastive learning strategy for Enhancing Robustness and Interpretability in Image Recognition through Environmental Perturbations (ERIEP) of clear-featured image data. ERIEP meticulously identifies a set of core visual features, termed "invariant features", which can offer optimal explanations for image predictions. Concurrently, it emphasizes learning noise-resistant strategies to amplify the model's interpretability. Through ERIEP's contrastive learning approach, we address complex images, enabling the model to progressively refine its understanding of both the invariant features and noise mitigation technique. Our extensive experiments on CIFAR-10, CIFAR-100, and ImageNet-1K demonstrate that ERIEP significantly outperforms several state-of-the-art image-processing baselines, showing robust performance under various noise intensities and environmental perturbations.

1. Introduction

Currently, robustness against various forms of input perturbations remains a crucial challenge (Zhang et al., 2017; Niu et al., 2020). In real-world scenarios, images (Torralba and Efros, 2011; Yang et al., 2024b) exhibit extensive variations attributed to diverse capture conditions, imaging devices, and inherent scene dynamics. To ensure consistent performance across various scenarios, models need to be resilient against both natural and adversarial perturbations. Real-world images are often subject to a much broader range of perturbations than represented in curated datasets, such as varying lighting conditions, image degradation, occlusion, etc. Moreover, the diversity and complexity of objects in real scenarios (Recht et al., 2019; Yang et al., 2024a) far exceed those in dataset images. A promising approach is to simulate various environmental conditions like illumination, motion blur, noise, Tobin et al. (2017), Qiu et al. (2021) etc. during training, thereby exposing the model to richer and more realistic image variations. This can enhance the model's adaptability to complex real-world scenarios and improve its robustness (Sun et al., 2019).

Data augmentation is a commonly utilized technique to improve model generalization. Traditional data augmentation techniques, such as rotation, scaling, and cropping, have been employed to artificially increase dataset size and variability. More recently, advanced augmentation strategies, like AutoAugment (Cubuk et al., 2018), have demonstrated the potential of automated, learned augmentation policies. However, the computational demands of AutoAugment inspired the development of more efficient strategies, such as Faster AutoAugment (FasterAA) (Hataya et al., 2020), which leverages proxy tasks to accelerate the search for optimal augmentations.

In recent years, contrastive learning has also been widely applied in the realm of data augmentation (Chen et al., 2020b; Grill et al., 2020). By constructing contrasts between positive and negative sample pairs, contrastive learning methods can learn beneficial feature representations, thus enhancing model robustness against complex transformations (Sheng et al., 2022; Wang and Qi, 2022). Some studies employ contrastive learning frameworks to design self-supervised data augmentation strategies, obtaining positive and negative pairs by artificially introducing transformations or perturbations to the training data, thereby improving generalization (Chen et al., 2020a). Combining contrastive learning frameworks with data augmentation techniques is a promising direction for enhancing the robustness of models (Qiu et al., 2023; Zhang and Ma, 2022).

* Corresponding author.
E-mail address: qiuxihe@sues.edu.cn (X. Qiu).
[1] These authors have contributed equally to this work.

https://doi.org/10.1016/j.engappai.2025.111619
Received 11 October 2023; Received in revised form 20 May 2025; Accepted 23 June 2025
Available online 8 July 2025

In this study, we introduce a novel training paradigm. By feeding images into an environment simulator, we generate perturbed versions of these images. These simulated images, in conjunction with their unperturbed counterparts, are then processed by a neural network. The subsequent loss calculations serve a dual function: optimizing the model for precise prediction and enhancing its robustness against image variations. Our approach also incorporates a reference model with the same architecture as the primary model, which remains static during the initial training stages. After several iterations, the reference model undergoes optimization, ensuring that our system remains adaptable and up-to-date with the ever-evolving image landscape.

Our contributions are three-fold:

1. We propose a novel approach called ERIEP, which involves training an environment simulator capable of comprehensive simulated perturbations and adopting a dual-input training strategy leveraging both original and perturbed images, enhancing model robustness and interpretability.
2. We prove that by introducing contrastive learning as a perturbation simulator of the external environment, the model is able to acquire more critical features, which enhances the generalization capability of the model.
3. Extensive experiments demonstrate that our approach achieves state-of-the-art performance compared to conventional baseline methods, showing improved model accuracy on perturbed images and robustness.

2. Related work

Models such as Visual Geometry Group Network (VGG) (Simonyan and Zisserman, 2014), and Residual Network (ResNet) (He et al., 2016), trained on clear-featured image dataset (Krizhevsky et al., 2012; Lin et al., 2014), have been effectively deployed for a variety of visual tasks. However, these datasets have limitations. First, their scale is limited without covering the full distribution of images (Recht et al., 2019). Second, their environments are homogeneous without various variations like illumination, occlusion, and noise. Third, the label information is incomplete with only limited semantics. Therefore, studies (Torralba and Efros, 2011; Marcus, 2018) have shown that relying solely on clear-featured image datasets may limit models' generalization to broader applications.

The usage of simulated environments for training models has emerged as a promising approach (Dosovitskiy et al., 2017; Johnson-Roberson et al., 2016). These environments enable the introduction of controlled perturbations to test the robustness of models. However, effective simulation methods that truly encapsulate the complexity of natural images remain an open research question (Shah et al., 2019). Contrastive learning has seen a surge of interest due to its potential in representation learning (Chen et al., 2020b). This unsupervised learning paradigm focuses on pushing apart dissimilar data points while bringing similar ones closer in the latent space (Wang and Isola, 2020). The effectiveness of contrastive learning for noise-resistant representations remains underexplored.

Understanding invariant features or core visual features is a crucial yet often under-addressed topic (Tan et al., 2023). Conventional models frequently exhibit limitations in comprehending the inherent visual semantics, leading to non-robust and less interpretable solutions (Zhang et al., 2021). With the increasing complexity of machine learning models, interpretability has become a key concern. Various techniques have been introduced to improve model interpretability, such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), but these generally focus on the post-hoc explanations and do not inherently make the model more understandable.

Several strategies have been proposed to make models more noise-resistant. Adversarial training (Zhang et al., 2019; Wang et al., 2019; Xie et al., 2019; Wu et al., 2020) and curriculum learning (Bengio et al., 2009; Vaswani et al., 2017) are a few notable methods. However, these approaches usually do not provide insights into why the model has become more robust. Recent work has focused on interpretable robustness, such as generative adversarial networks can be utilized to analyze the noise distribution (Krull et al., 2019), representation learning can reveal feature stability (Arnab et al., 2018), and example-based methods can detect the causes of model misclassifications (Corbett-Davies et al., 2017). These preliminary attempts suggest that by analyzing how noise affects the model, and which patterns in the training examples are critical for model robustness, we can gain a deeper understanding and guide the design of better adversarial noise mechanisms. Our work is also inspired by this idea, trying to obtain explanatory insights into model robustness by analyzing the invariant features and noise patterns. The proposed ERIEP approach innovatively combines elements from several existing research areas. It leverages the strengths of clear-featured image data and simulated environments while employing contrastive learning techniques. This multi-faceted approach aims to address the challenges of noise, interpretability, and the identification of core visual features in a unified framework. ERIEP offers valuable insights for refining environmental simulations and further enhancing the robustness and interpretability of models.

3. Methodology

3.1. Overview

ERIEP aims to enhance the robustness and interpretability of models in image classification tasks. ERIEP integrates contrastive learning with controlled environmental perturbations to improve model generalization across diverse conditions. The method includes three primary components: the Environmental Perturbation Module (EPM), which simulates environmental disturbances on input images; the Neural Network Model ($\mathcal{N}$), which processes the original images for classification; and the Reference Model ($\mathcal{R}$), which assesses the perturbed images and is periodically synchronized with $\mathcal{N}$. As shown in Fig. 1, the framework is divided into two parts: (a) the Environmental Perturbation Module (EPM), which generates perturbed images from the original inputs using a set of sub-policies; and (b) the feature learning and model optimization process, where invariant features are extracted and the main model $\mathcal{N}$ is trained with feedback from the reference model $\mathcal{R}$. The workflow, as illustrated in Fig. 1, involves generating perturbed images using EPM, processing these through both $\mathcal{N}$ and $\mathcal{R}$, and updating $\mathcal{N}$ based on a novel objective function that includes the absolute difference in cross-entropy (Murphy, 2012) losses between the two models. The losses $\mathcal{L}_{\mathcal{N}}$, $\mathcal{L}_{\mathcal{R}}$, and the combined $\mathcal{L}_{\text{final}}$ are computed to guide the optimization of $\mathcal{N}$. In addition, a periodic synchronization step ensures that $\mathcal{R}$ stays aligned with the evolving parameters of $\mathcal{N}$. This setup ensures consistent model performance under varied environmental influences and promotes the learning of invariant features. Detailed discussions on the implementation and mathematical formulations of each component will follow in subsequent sections, elucidating their contributions to the overall goal of improving model robustness and interpretability.

3.2. EPM

To simulate realistic environmental perturbations in images, we design EPM based on the Faster AutoAugment framework (Hataya et al., 2020). This module aims to ensure that the perturbed image retains essential features for correct classification, similar to the original image.

Module Architecture: The core of our EPM is a set of sub-policies $S = \{S_1, S_2, \ldots, S_L\}$, where each S_i represents a unique augmentation strategy. Each sub-policy S_i consists of K consecutive image-processing operations $(O_1^{(i)}, \ldots, O_K^{(i)})$, such as shear, rotate, or color adjustments. Table 1 provides a comprehensive list of all 16 candidate operations

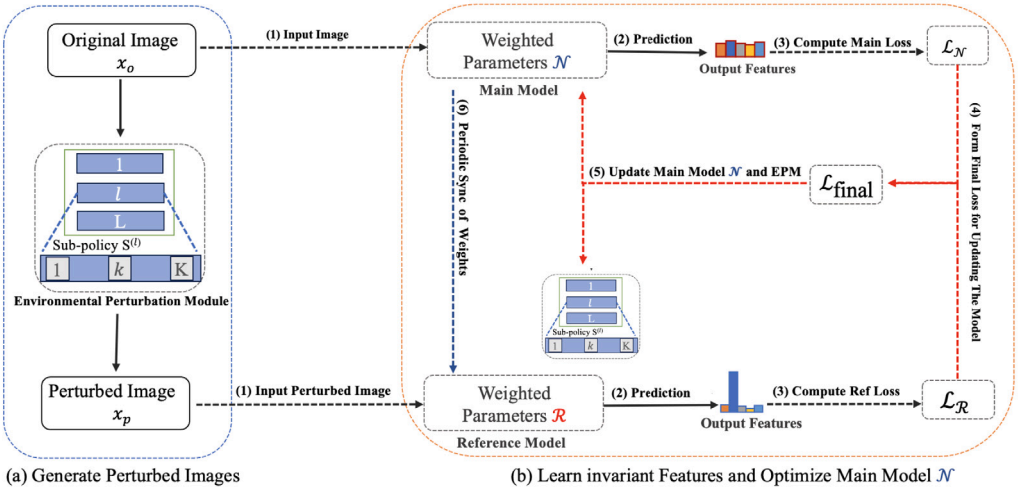


Fig. 1. Overview of the proposed ERIEP framework. Part (a) shows the EPM that generates perturbed images. Part (b) illustrates the learning process, where invariant features are learned and the main model $\mathcal{N}$ is optimized with guidance from the reference model $\mathcal{R}$ through loss computation and periodic synchronization. Here, L denotes the number of sub-policies, and K represents the number of operations in each sub-policy

available to the EPM, grouped by category together with the type of learnable magnitude μ (continuous, discrete or none). Each operation $O_k^{(i)}$ is applied with a probability $p_k^{(i)}$ and magnitude $\mu_k^{(i)}$. These parameters $(p_k^{(i)}, \mu_k^{(i)})$ are learned during the optimization process. During training, a differentiable controller assigns soft weights $w_{k,n}$ to every candidate operation $O_k^{(n)}$ via a temperature-controlled softmax; the perturbed image at stage k is the weighted sum of all candidate outputs. This design allows gradients to flow simultaneously to the operation-selection weights as well as to the probability and magnitude parameters. Once the search converges, we discretely sample the top-1 operation per stage according to $w_{k,n}$ to form the final policy used in inference. The sub-policies are therefore able to create diverse yet realistic environmental perturbations, enhancing the robustness of our model to various image transformations.

$$\mathbf{x}_p = S(\mathbf{x}_o; M, P, W) \tag{1}$$

$$\mathbf{p}(\mathbf{y}|\mathbf{x}_p) = \mathcal{F}_{\mathrm{NL}}(\mathbf{x}_p) \tag{2}$$

Where $\mathbf{x}_o$ represents the input original image, $S(\cdot; M, P, W)$ is the sub-policy function composed of L sub-policies, each with learnable parameters M, P, W. Here, $M = (\mu_1, \ldots, \mu_K)$ represents the magnitude parameters, $P = (p_1, \ldots, p_K)$ represents the probability parameters, and W represents the weights for selecting operations. $\mathbf{x}_p$ is the perturbed image output. $\mathcal{F}_{\mathrm{NL}}$ is the newly added network layer that processes the perturbed image. $\mathbf{y}$ represents the class label, and $\mathbf{p}(\mathbf{y}|\mathbf{x}_p)$ is the probability distribution over class labels given the perturbed image. Each sub-policy S_i in S consists of K consecutive image transformation operations, optimized to create diverse yet realistic environmental perturbations.

Objective and Loss Formulation: The primary objective during the training of the EPM is to minimize the distance between distributions of the augmented and the original images, while ensuring consistent classification results. The total loss function for the EPM is formulated as:

$$\mathcal{L}_{\mathrm{EPM}} = d_\theta(\mathcal{A}, \mathcal{B}') + \epsilon l \tag{3}$$

Where $\mathcal{B}$ is a mini-batch of training samples, $\mathcal{A} = \{S(\mathbf{x}_o; M, P, W); (\mathbf{x}_o, \cdot) \in \mathcal{B}\}$ is the set of augmented images, $\mathcal{B}'$ is another batch sampled from the training set, and ϵ is a weighting parameter. $d_\theta(\cdot, \cdot)$ is the Wasserstein distance with learnable parameters θ, implemented using Wasserstein GAN with gradient penalty. The classification loss l is defined as:

$$l = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{A}}\mathcal{L}(f(\mathbf{x}), y) + \mathbb{E}_{(\mathbf{x}',y')\sim\mathcal{B}'}\mathcal{L}(f(\mathbf{x}'), y') \tag{4}$$

Where $\mathcal{L}$ is the cross-entropy loss function, and f is the image classifier.

Table 1
Augmentation operations in EPM and the type of learnable magnitude parameter μ.

Category	Operation	Magnitude type
Affine	shear_x	continuous
	shear_y	continuous
	translate_x	continuous
	translate_y	continuous
	rotate	continuous
	flip	none
Color	solarize	discrete
	posterize	discrete
	invert	none
	contrast	continuous
	color	continuous
	brightness	continuous
	sharpness	none
	auto_contrast	none
	equalize	none
Other	cutout	discrete
	sample_pairing	continuous

3.3. Processing the original image through the neural network model

To achieve a comprehensive understanding of the impact of the perturbation module, it is essential to first analyze the response of the neural network model to the original image. This step establishes a baseline for comparison with the results after applying environmental perturbations.

Model Processing and Loss Computation: The original image, $\mathbf{x}_o$, serves as the input to our trained neural network model $\mathcal{N}$. The image undergoes the typical transformations and convolutions defined by the network architecture:

$$\hat{\mathbf{y}} = \mathcal{N}(\mathbf{x}_o) \tag{5}$$

Here, $\hat{\mathbf{y}}$ represents the predicted probability distribution over classes obtained after processing $\mathbf{x}_o$ through $\mathcal{N}$.

Cross-Entropy Loss Computation: We use cross-entropy loss as the standard metric for our classification task. Given the true label of the image as $\mathbf{y}$ (a one-hot encoded vector) and the predicted output $\hat{\mathbf{y}}$, the cross-entropy loss is defined as:

$$\mathcal{L}_{\mathcal{N}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{6}$$

Where y_i is the actual label for class i (either 0 or 1 in the one-hot encoding), and $\hat{y}_i$ is the predicted probability for class i.

By measuring the network's response to $\mathbf{x}_o$ and quantifying it via $\mathcal{L}_{\mathcal{N}}$, we establish a baseline against which the effect of environmental perturbations can be evaluated in subsequent steps of our analysis.

3.4. Processing the perturbed image through the reference model

To measure the effectiveness and potential influence of the environmental perturbation, we process the perturbed image through a separate reference model. This step allows us to compare the outputs of the original and perturbed images, providing insights into the impact of our perturbation module.

3.4.1. Reference model forward pass

The perturbed image, $\mathbf{x}_p$, serves as input to our reference neural network model $\mathcal{R}$. This model has an identical structure to the main model $\mathcal{N}$, but its weights are periodically synchronized with $\mathcal{N}$ and remain static during certain phases of training, acting as a benchmark. When $\mathbf{x}_p$ is processed, we have:

$$\hat{\mathbf{y}}_p = \mathcal{R}(\mathbf{x}_p) \tag{7}$$

Here, $\hat{\mathbf{y}}_p$ represents the predicted probability distribution over classes obtained after processing $\mathbf{x}_p$ through $\mathcal{R}$.

3.4.2. Cross-entropy loss computation for perturbed image

We use cross-entropy loss to quantify the difference between the predicted outputs from the perturbed image and the actual labels. Given the true label $\mathbf{y}$ (a one-hot encoded vector) and the predicted output $\hat{\mathbf{y}}_p$, the cross-entropy loss is represented as:

$$\mathcal{L}_{\mathcal{R}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_{p,i}) \tag{8}$$

Where N is the total number of classes, y_i is the actual label for class i (either 0 or 1 in the one-hot encoding), and $\hat{y}_{p,i}$ indicates the predicted probability for class i from the perturbed image.

3.5. Final objective function formation

Our ultimate objective is to minimize the loss of the original image while ensuring that the difference between the losses of the original and perturbed images remains small. This approach encourages the model to be robust to perturbations while maintaining high accuracy on unperturbed images.

Let $\mathcal{L}_{\mathcal{N}}$ be the cross-entropy loss obtained from the main neural network model when processing the original image, and $\mathcal{L}_{\mathcal{R}}$ be the cross-entropy loss obtained from the reference model when processing the perturbed image. Our final objective function $\mathcal{L}_{\text{final}}$ is formulated as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\mathcal{N}} + \alpha \cdot |\mathcal{L}_{\mathcal{N}} - \mathcal{L}_{\mathcal{R}}| \tag{9}$$

Where α is a hyperparameter that controls the trade-off between model accuracy and robustness to perturbations. A larger α places more emphasis on making the model robust to perturbations, while a smaller α prioritizes accuracy on unperturbed images. The optimal value of α can be determined through cross-validation or other hyperparameter tuning techniques.

During training, we update the weights of the main model $\mathcal{N}$ and the perturbation module to minimize $\mathcal{L}_{\text{final}}$. The weights of the reference model $\mathcal{R}$ are periodically synchronized with $\mathcal{N}$ but remain static during the optimization process, serving as a stable target for the perturbation module. This design not only improves robustness but also enhances interpretability, as the consistency between $\mathcal{N}$ and $\mathcal{R}$ under perturbations encourages the model to rely on stable, semantically relevant features rather than spurious patterns.

3.6. Optimizing neural networks for robustness against environmental perturbations

In Algorithm 1, we present the two-stage Optimized Environmental Perturbation Training Strategy, a novel approach for training neural network models.

The training is divided into two distinct stages. In the first stage, the EPM is independently optimized to generate meaningful and diverse perturbations based on specific objectives. Once the EPM reaches convergence, it is then used in the second stage to guide the training of the main model. The second stage incorporates a Reference Model $\mathcal{R}$ alongside the main Neural Network Model $\mathcal{N}$. Here, the EPM perturbs the input images, and both clean and perturbed samples are used to compute a combined loss that promotes robustness.

The key feature of this algorithm is the synchronization of weights between $\mathcal{N}$ and $\mathcal{R}$ at regular intervals, defined by a parameter $sync\_freq$. This delayed synchronization maintains a stable reference signal and enhances the contrastive training effect, avoiding premature convergence between $\mathcal{N}$ and $\mathcal{R}$.

Algorithm 1 Optimized Environmental Perturbation Training Strategy

Require: Original image $\mathbf{x}_o$, Neural Network Model $\mathcal{N}$, Reference Model $\mathcal{R}$, Synchronization Frequency $sync\_freq$, Learning Rate
Ensure: Trained $\mathcal{N}$ and Environmental Perturbation Module EPM

Perturbation Parameter Optimization
procedure OPTIMIZE EPM
  Initialize perturbation parameters M, P, W
  Set learning rate.
  for each epoch **do**
    Fetch original image $\mathbf{x}_o$.
    Generate perturbed image $\mathbf{x}_p = S(\mathbf{x}_o; M, P, W)$.
    Compute the loss $\mathcal{L}_{\text{EPM}} = d_\theta(\mathcal{A}, \mathcal{B}') + \epsilon l$.
    Update EPM parameters M, P, W using gradient descent.
  end for
  Stop if convergence criteria are met.
end procedure

Network Robustness Optimization with Adaptive Perturbation
procedure TRAIN NETWORK
  Initialize $\mathcal{N}$ and $\mathcal{R}$ with random weights
  for each iteration **do**
    $\mathbf{x}_p = EPM(\mathbf{x}_o)$
    $\hat{\mathbf{y}} = \mathcal{N}(\mathbf{x}_o)$
    $\mathcal{L}_{\mathcal{N}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$
    $\hat{\mathbf{y}}_p = \mathcal{R}(\mathbf{x}_p)$
    $\mathcal{L}_{\mathcal{R}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_{p,i})$
    $\mathcal{L}_{\text{final}} = \mathcal{L}_{\mathcal{N}} + \alpha \cdot |\mathcal{L}_{\mathcal{N}} - \mathcal{L}_{\mathcal{R}}|$
    Update $\mathcal{N}$ using $\mathcal{L}_{\text{final}}$
    Update EPM using $\mathcal{L}_{\text{final}}$
    if iteration mod $sync\_freq$ == 0 **then**
      $\mathcal{R}$.weights = $\mathcal{N}$.weights
    end if
  end for
end procedure

4. Experiments

In this section, we present our experimental setup and results on multiple datasets, with a focus on evaluating our proposed training method's effectiveness in improving model robustness and generalization.

4.1. Datasets and perturbation techniques

Our primary experiments were conducted on the CIFAR-10 (Krizhevsky, 2009) image classification dataset, which contains 10 classes with 50,000 32×32 training images and 10,000 test images.

Table 2
Detailed perturbation methods.

Perturbation	Operations	Parameters
1	Random erasing	Scale = (0.01,0.05)
	RandomCrop	size = 32, padding = 4
2	Random rotation	Degree=15
	Random horizontal flip	–
	Color jittering	Brightness = 0.4, Contrast = 0.4
3	Gaussian blur	Kernel size = 3, Sigma = 0.5

We evaluated the top-1 accuracy on both the original test set and under three types of perturbations.

Table 2 provides a detailed overview of the perturbation techniques employed in our experiments. These techniques were carefully designed to simulate various changes and challenges in real-world image capture processes.

We utilized three main perturbation methods:

1. The first combines random erasing (simulating partial object occlusion) and random cropping (mimicking different framing);
2. The second integrates random rotation, random horizontal flipping, and color jittering, simulating camera angle variations, object orientation changes, and lighting condition alterations respectively;
3. The third applies Gaussian blur to emulate camera focus issues or motion blur.

Each perturbation is implemented with specific parameter settings to ensure reasonable and effective distortion effects. By incorporating these diverse perturbations, we aim to enhance our model's robustness and generalization capabilities, better preparing it for the complexities of real-world scenarios.

4.2. Experimental settings

To evaluate the effectiveness of our proposed method, we conducted experiments with various Convolutional Neural Network (CNN) architectures, including 8-Layer CNN, VGG-16 (Simonyan and Zisserman, 2014), ResNet-18 (He et al., 2016), and ConvNeXt (Liu et al., 2022). Each model was trained under two settings: conventional training, which follows the standard procedure of optimizing cross-entropy loss on the training set, and our proposed method, as described in Section 3, incorporating our novel training techniques.

For all models on CIFAR-10, we employed the Adam optimizer and trained for 50 epochs, while keeping other hyperparameters at their default values. Regarding specific architectural details, VGG-16 utilized three linear layers (512-512, 512-256, 256-10) for classification. The 8-Layer CNN had channel configurations of [64, 'M', 128, 'M', 256, 256, 'M', 512, 512, 512, 512] with a 3×3 kernel size, and its linear layers were identical to those of VGG-16. For ConvNeXt, we resized CIFAR-10 images to 224×224 before training to fit the input dimension.

We further extended our experiments to CIFAR-100 and ImageNet-1K datasets using ResNet-50 and ResNet-101 architectures. In these additional experiments, we compared our proposed ERIEP with CrossEntropy, Supervised Contrastive Learning (SupCon) (Khosla et al., 2021), and Unified Contrastive Learning (UniCL) (Yang et al., 2022). The results of these experiments are presented in Table 4. To ensure reliability, all experiments were repeated independently three times, and we report the averaged results.

For CIFAR-10 experiments, we used a learning rate of 0.001 for the Adam optimizer. For CIFAR-100 and ImageNet-1K, we trained the models for 200 and 90 epochs, respectively. We employed the Adam optimizer with a momentum of 0.9 and an initial learning rate of 0.1, which was decreased by a factor of 10 at 60% and 80% of the total epochs. All experiments were conducted on a system equipped with two NVIDIA GeForce RTX 3090 GPUs (24 GB VRAM each), using PyTorch 1.13.1 and CUDA 11.7. The system also featured an Intel Core i9 12900KS CPU and 64 GB of RAM.

Table 3
Model accuracy on original and perturbed CIFAR-10. The term 'Standard' denotes the conventional training methodologies.

Perturbation method	Method	8-Layer-CNN	VGG-16	ResNet-18	ConvNext
None	Standard	76.74	80.43	78.13	80.71
	ERIEP(Our)	**80.01**	**84.21**	**80.21**	**82.21**
perturb-1	Standard	39.97	49.52	31.34	38.98
	ERIEP(Our)	**63.17**	**67.21**	**35.41**	**42.01**
perturb-2	Standard	40.98	49.10	34.43	38.80
	ERIEP(Our)	**66.07**	**68.52**	**39.07**	**40.51**
perturb-3	Standard	55.43	61.86	48.69	46.23
	ERIEP(Our)	**77.60**	**80.91**	**50.11**	**52.14**

Table 4
Comparison of image classification performance on CIFAR-100 and ImageNet-1K using CrossEntropy, SupCon, UniCL, and our Enhanced Robust Image Encoding Pipeline (ERIEP). The results are reported for both ResNet-50 and ResNet-101 architectures.

Dataset	Method	ResNet-50	ResNet-101
CIFAR-100	CrossEntropy	75.3	78.8
	SupCon	76.5	79.6
	UniCL	78.4	81.4
	ERIEP (Ours)	**80.1**	**83.2**
ImageNet-1K	CrossEntropy	78.2	79.8
	SupCon	78.7	80.2
	UniCL	78.1	79.9
	ERIEP (Ours)	**81.7**	**83.6**

4.3. Analysis and results

Table 3 illustrates an evaluation of the proposed ERIEP methodology across a variety of models including 8-Layer-CNN, VGG-16, ResNet-18, and ConvNext, against standard methodologies on both original and perturbed CIFAR-10 datasets. The analysis explains the performance of ERIEP under both unperturbed and various perturbed conditions in detail. All results in this section are reported on a completely independent test set, which was not used for training or validation at any stage. To better reflect real-world noise conditions, we selected perturbation techniques such as Gaussian blur, color jitter, and occlusion. These techniques are intended to simulate common artifacts in natural image acquisition, including motion blur, lighting variation, and sensor noise. Although synthetic, these perturbations are intended to serve as controlled proxies for complex environmental disturbances that models may encounter in practical scenarios.

Under unperturbed conditions, ERIEP consistently manifests superior accuracy across all employed models compared to standard training methodologies. Specifically, in the VGG-16 model, ERIEP achieved an accuracy of 84.21%, surpassing the 80.43% achieved by the standard method. This enhanced performance underscores ERIEP's superior capability in comprehending and processing core visual features of images, culminating in heightened classification accuracy.

Under varied perturbed conditions, ERIEP continues to exemplify eminent performance relative to the conventional methods. For instance, under the 'perturb-1' condition, ERIEP in the VGG-16 model actualized an accuracy rate of 67.21% while the conventional counterpart reached only 49.52%. This insight accentuates the robustness and adaptability of ERIEP in complex, noise-induced environments, particularly when subjected to diverse environmental perturbations.

When analyzing performance disparities amongst different models, it is discernible that ERIEP enhances performance universally across diverse models, indicating a prolific applicability of the methodology. Even in models like ResNet-18, where the relative advantage is somewhat moderate, ERIEP still portrays its applicative potential and reliability across diversified model architectures and operational environments (Huang et al., 2023).

Through the scrutiny of results under multitudinous perturbed conditions, ERIEP delineates comprehensive superiority and exceptional noise resistance, thereby fortifying the interpretative capabilities of the models. This robustness suggests that ERIEP can help ensure stable predictions even in field deployments where environmental conditions may vary unpredictably. This not only augments the explanatory power of models but also facilitates the deployment of models in environments characterized by a plethora of complexities and variabilities (de Jong and Bosman, 2019).

To further validate the effectiveness of our proposed method, we extended our experiments to more challenging datasets: CIFAR-100 and ImageNet-1K. Table 4 presents the results of these experiments, comparing ERIEP with CrossEntropy, SupCon, and UniCL methods on ResNet-50 and ResNet-101 architectures.

As evident from Table 4, ERIEP consistently outperforms CrossEntropy, SupCon, and UniCL methods across different architectures and datasets. On CIFAR-100, ERIEP achieves an 4.8% and 4.4% improvement over CrossEntropy for ResNet-50 and ResNet-101, respectively. Similarly, on the more challenging ImageNet-1K dataset, ERIEP demonstrates substantial improvements, with a 3.5% increase in accuracy for ResNet-50 and a 3.8% increase for ResNet-101 compared to CrossEntropy. Furthermore, ERIEP surpasses the performance of UniCL by 1.7% and 1.8% on CIFAR-100 and 3.6% and 3.7% on ImageNet-1K for ResNet-50 and ResNet-101, respectively.

These results on larger and more complex datasets further corroborate the effectiveness of ERIEP in enhancing model performance and generalization capabilities. The consistent superiority of ERIEP across different model architectures and datasets, including its advantage over the state-of-the-art UniCL method, underscores its potential as a robust and versatile approach for improving image classification tasks.

In conclusion, the experimental outcomes lucidly demonstrate that ERIEP consistently achieves significant performance enhancements across diverse models under both unperturbed and multiple perturbed conditions. Moreover, its effectiveness extends to more challenging datasets and complex model architectures, outperforming even the state-of-the-art UniCL method. This proves that ERIEP not only excels on standard datasets but also demonstrates high adaptability and robustness against various perturbations likely encountered in real-world environments. This detailed analysis corroborates the methodological innovations posited by ERIEP, namely providing enriched interpretative insights and robust performances in noise-pervaded environments (Rawat and Wang, 2017).

5. Ablation

5.1. Quantitative analysis on robustness against Gaussian noise

This subsection aims to provide a rigorous quantitative assessment of the model's robustness to Gaussian noise perturbations. A comprehensive comparison is conducted between the test accuracies achieved using our proposed training methodology (ERIEP) and those obtained through traditional training approaches. We employ various degrees of Gaussian noise as an intentionally manipulated parameter, facilitating a comprehensive examination of the efficacy of our training approach in enhancing the model's resilience to stochastic image degradation. Here, the noise level refers to zero-mean Gaussian noise with varying standard deviations added to each pixel of the image.

As delineated in Fig. 2, we observe the trend of test accuracy on CIFAR-10 when subjected to increasing levels of Gaussian noise. Notably, both the model trained with our proposed ERIEP method and the one trained with standard techniques demonstrate a decrement in accuracy as the noise level intensifies. However, what deserves attention is the relative robustness exhibited by our method, consistently maintaining a higher degree of accuracy under noisier conditions.

At the baseline (noise level 0), ERIEP achieves a slightly higher accuracy (approximately 87%) compared to the standard method (around 84%). As the noise level increases, the performance gap between ERIEP and the standard method becomes more pronounced. For instance, at a noise level of 0.2, ERIEP maintains an accuracy of about 52%, while the standard method's accuracy drops to approximately 30%, showcasing a significant 22% difference.

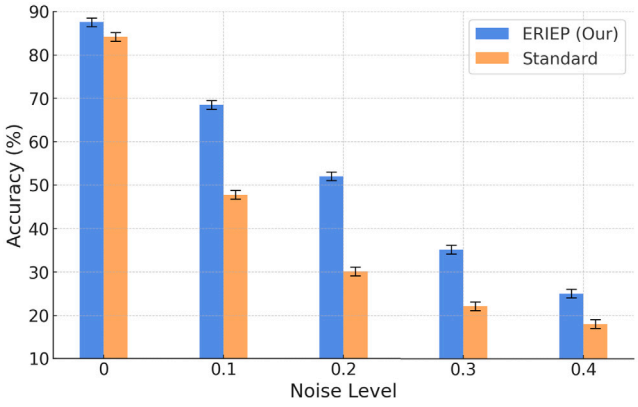


Fig. 2. Test accuracy under Gaussian noise perturbations.

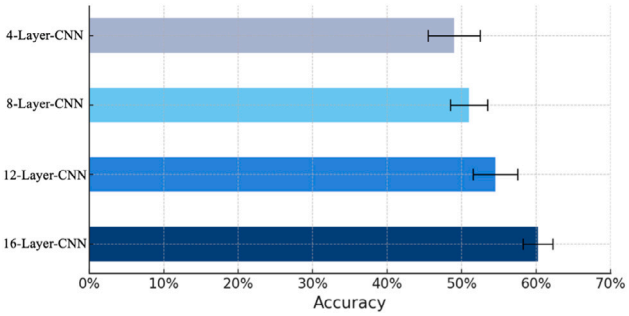


Fig. 3. Model invariance across architectural depths.

The robustness of ERIEP is particularly evident in high-noise scenarios. At a noise level of 0.4, where image quality is severely degraded, ERIEP still manages to maintain an accuracy of about 25%, while the standard method's performance declines to approximately 18%. This demonstrates ERIEP's superior ability to extract meaningful features even under extreme noise conditions.

The experimental results validate the efficacy of our proposed training technique in enhancing model robustness. By incorporating the Environmental Perturbation Module and contrastive learning, the model learns useful feature representations that are invariant to random perturbations, imparting a degree of fault tolerance as evidenced by maintaining high test accuracy under noisy conditions. The addition of the environment disturbance module during training constructs an adversarial setting for the model, compelling it to grasp key visual concepts despite noise interference.

Furthermore, the consistent performance advantage of ERIEP across all noise levels suggests that our method not only improves robustness but also enhances the overall feature learning capability of the model. This is reflected in the higher baseline accuracy and the sustained performance under increasing noise levels.

In summary, both qualitative and quantitative analyses substantiate that our ERIEP approach significantly boosts model robustness, enabling maintenance of strong performance under complex noisy conditions. The results underscore the potential of ERIEP in real-world applications where image quality may be compromised or unpredictable, offering a more reliable solution for image classification tasks in challenging environments.

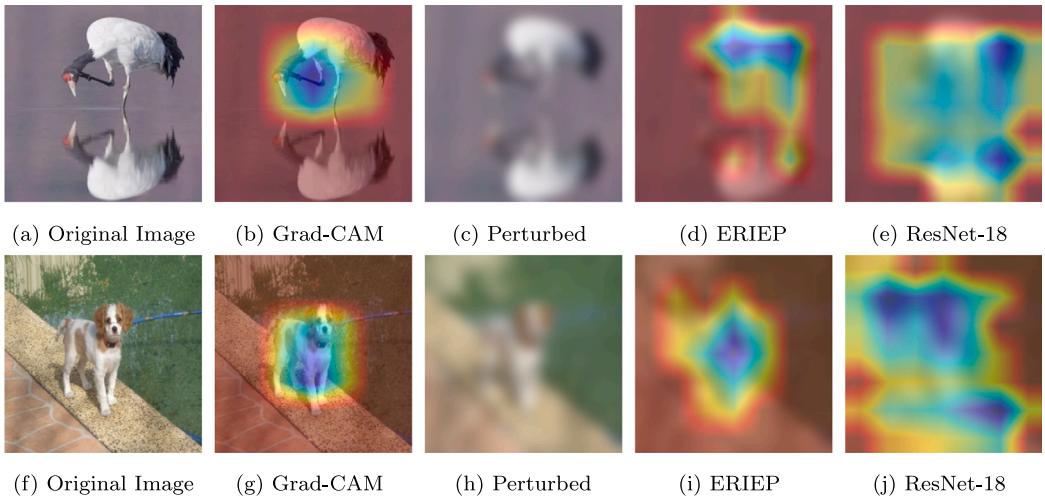


Fig. 4. (a, f) Original images from ImageNet-1K; (b, g) Grad-CAM from ResNet-18 on original images; (c, h) Perturbed images with Gaussian blur; (d, i) Grad-CAM from ERIEP on perturbed images; (e, j) Grad-CAM from ResNet-18 on perturbed images.

5.2. Ablation study on CNN architectural depth

To investigate the role of architectural complexity in the effectiveness of our proposed training methodology, we conducted an ablation study using CNNs of varying depths. Specifically, we trained models with architectures consisting of 4, 8, 12, and 16 layers on the CIFAR-10 training set (see Fig. 3). Subsequently, to evaluate the robustness of the trained models, we introduced Gaussian noise (mean=0, standard deviation=0.2) to the CIFAR-10 test set and measured the performance variations across the different architectures.

Firstly, the experimental results substantiate that our training approach consistently and significantly enhances performance across a range of architectural depths. This underlines the scalability and potential applicability of our methodology, particularly in contexts involving deeper Convolutional Neural Network architectures.

However, an intriguing observation is the deceleration of performance gains as the depth of the model increases. Multiple factors could account for this phenomenon. For instance, deeper models may be susceptible to overfitting, especially when training data are limited. Such overfitting could impede the model's generalization capabilities on the perturbed test set.

In summary, this ablation study provides valuable insights into the relationship between model depth and the efficacy of our training approach. While consistently boosting performance across architectural scales, the benefits appear to taper off in very deep models, likely due to overfitting as well as other factors like vanishing gradients. Further research into regularization techniques and expanded datasets may help address these challenges. Overall, our training methodology demonstrates strong potential, particularly for moderately deep CNN architectures.

5.3. Qualitative analysis of interpretability under perturbation

To support the core claim that ERIEP enhances interpretability by learning invariant features, we conduct a Grad-CAM (Selvaraju et al., 2019) based qualitative analysis on two representative ImageNet-1K samples (a bird and a dog). Each image is processed in the following manner: (1) the original image is passed through both a standard ResNet-18 and our ERIEP-enhanced model to obtain Grad-CAM maps; (2) a perturbed version of the same image is generated using Gaussian blur with a radius of 9, simulating strong environmental noise; (3) both models are evaluated again on the perturbed image using Grad-CAM.

As illustrated in Fig. 4, the Grad-CAM visualizations show that ERIEP consistently focuses on semantically meaningful regions (e.g., the head or torso of the object), even under heavy blur. In contrast, the standard ResNet-18 exhibits attention drift or diffused focus under the same perturbations. These results qualitatively demonstrate ERIEP's improved interpretability and robustness through noise-invariant feature learning.

6. Conclusion and future work

This paper presents ERIEP, an innovative contrastive learning approach designed to enhance model robustness and interpretability in the face of real-world environmental disturbances. ERIEP integrates an EPM to simulate realistic disturbances and adopts a dual-input training strategy that leverages both original and perturbed images. This approach enables the model to learn invariant features and noise-resistant strategies, thereby improving its generalization capabilities and interpretability. Our extensive experiments on CIFAR-10, CIFAR-100, and ImageNet-1K demonstrate ERIEP's versatility across different model architectures and its superior performance under various levels of image degradation. ERIEP consistently outperforms several state-of-the-art baselines, including CrossEntropy, SupCon, and UniCL, showcasing its effectiveness in enhancing model robustness. Furthermore, our ablation studies reveal ERIEP's ability to achieve more stable and optimized learning processes, highlighting its potential for real-world applications.

In the future, we plan to extend ERIEP to more diverse tasks such as object detection and semantic segmentation, investigating its applicability in challenging real-world deployment scenarios, and exploring its potential in transfer learning and few-shot learning contexts. By addressing the challenges posed by environmental disturbances, our work contributes to the broader goal of creating more reliable and adaptable systems for real-world applications.

CRediT authorship contribution statement

Leijun Cheng: Visualization, Investigation, Writing – original draft, Methodology. **Xihe Qiu:** Methodology, Funding acquisition, Supervision. **Xiaoyu Tan:** Writing – review & editing, Methodology, Investigation. **Haoyu Wang:** Validation. **Yujie Xiong:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Shanghai Municipal Natural Science Foundation, China (23ZR1425400).

Data availability

The authors do not have permission to share data.

References

Arnab, A., Miksik, O., Torr, P.H., 2018. On the robustness of semantic segmentation models to adversarial attacks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 888–897. http://dx.doi.org/10.1109/cvpr.2018.00099.

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 41–48. http://dx.doi.org/10.1145/1553374.1553380.

Chen, X., Fan, H., Girshick, R., He, K., 2020a. Improved baselines with momentum contrastive learning. http://dx.doi.org/10.48550/arXiv.2003.04297, arXiv preprint arXiv:2003.04297.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607, URL https://proceedings.mlr.press/v119/chen20j.html.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A., 2017. Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 797–806. http://dx.doi.org/10.1145/3097983.3098095.

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2018. Autoaugment: Learning augmentation policies from data. http://dx.doi.org/10.1109/cvpr.2019.00020, arXiv preprint arXiv:1805.09501.

de Jong, K.L., Bosman, A.S., 2019. Unsupervised change detection in satellite images using convolutional neural networks. In: 2019 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–8. http://dx.doi.org/10.1109/ijcnn.2019.8851762.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V., 2017. CARLA: An open urban driving simulator. In: Conference on Robot Learning. PMLR, pp. 1–16, URL https://proceedings.mlr.press/v78/dosovitskiy17a.html.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 33, 21271–21284, URL https://papers.nips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Hataya, R., Zdenek, J., Yoshizoe, K., Nakayama, H., 2020. Faster autoaugment: Learning augmentation strategies using backpropagation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. Springer, pp. 1–16. http://dx.doi.org/10.1007/978-3-030-58595-2_1.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Huang, S.-Y., An, W.-J., Zhang, D.-S., Zhou, N.-R., 2023. Image classification and adversarial robustness analysis based on hybrid quantum–classical convolutional neural network. Opt. Commun. 533, 129287. http://dx.doi.org/10.1016/j.optcom.2023.129287.

Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R., 2016. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?. http://dx.doi.org/10.1109/icra.2017.7989092, arXiv preprint arXiv:1610.01983.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2021. Supervised contrastive learning. arXiv:2004.11362, URL https://arxiv.org/abs/2004.11362.

Krizhevsky, 2009. Learning multiple layers of features from tiny images. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, http://dx.doi.org/10.1145/3065386.

Krull, A., Buchholz, T.-O., Jug, F., 2019. Noise2void-learning denoising from single noisy images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2129–2137. http://dx.doi.org/10.1109/CVPR.2019.00223.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755. http://dx.doi.org/10.1007/978-3-319-10602-1_48.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986. http://dx.doi.org/10.1109/cvpr52688.2022.01167.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30, URL https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Marcus, G., 2018. Deep learning: A critical appraisal. http://dx.doi.org/10.48550/arXiv.1801.00631, arXiv preprint arXiv:1801.00631.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. Cambridge, MA, URL https://www.semanticscholar.org/paper/Machine-learning-a-probabilistic-perspective-Murphy/360ca02e6f5a5e1af3dce4866a257aafc2d6d6f5.

Niu, X., Mathur, P., Dinu, G., Al-Onaizan, Y., 2020. Evaluating robustness to input perturbations for neural machine translation. pp. 8538–8544. http://dx.doi.org/10.18653/v1/2020.acl-main.755.

Qiu, X., Shi, S., Tan, X., Qu, C., Fang, Z., Wang, H., Gao, Y., Wu, P., Li, H., 2023. Gram-based attentive neural ordinary differential equations network for video nystagmography classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 21339–21348, URL https://openaccess.thecvf.com/content/ICCV2023/papers/Qiu_Gram-based_Attentive_Neural_Ordinary_Differential_Equations_Network_for_Video_Nystagmography_ICCV_2023_paper.pdf.

Qiu, X., Tan, X., Yan, F., Su, Q., Chen, J., Jiang, X., 2021. Semi-supervised recommender system for bone implant ratio recommendation. J. Ambient. Intell. Humaniz. Comput. 1–10, URL https://link.springer.com/article/10.1007/s12652-021-03156-2.

Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. 29 (9), 2352–2449. http://dx.doi.org/10.1162/neco_a_00990.

Recht, B., Roelofs, R., Schmidt, L., Shankar, V., 2019. Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning. PMLR, pp. 5389–5400. http://dx.doi.org/10.48550/arXiv.1902.10811.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. http://dx.doi.org/10.1145/2939672.2939778.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2019. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. 128 (2), 336–359. http://dx.doi.org/10.1007/s11263-019-01228-7.

Shah, M., Chen, X., Rohrbach, M., Parikh, D., 2019. Cycle-consistency for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6649–6658. http://dx.doi.org/10.1109/cvpr.2019.00681.

Sheng, G., Wang, Q., Pei, C., Gao, Q., 2022. Contrastive deep embedded clustering. Neurocomputing 514, 13–20. http://dx.doi.org/10.1016/j.neucom.2022.09.116, URL https://www.sciencedirect.com/science/article/pii/S0925231222012085.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. http://dx.doi.org/10.48550/arXiv.1409.1556, arXiv preprint arXiv:1409.1556.

Sun, Q., Liu, Y., Chua, T.-S., Schiele, B., 2019. Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 403–412. http://dx.doi.org/10.1109/cvpr.2019.00049.

Tan, X., Yong, L., Zhu, S., Qu, C., Qiu, X., Yinghui, X., Cui, P., Qi, Y., 2023. Provably invariant learning without domain information. URL https://dl.acm.org/doi/abs/10.5555/3618408.3619805.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P., 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 23–30. http://dx.doi.org/10.1109/iros.2017.8202133.

Torralba, A., Efros, A.A., 2011. Unbiased look at dataset bias. In: CVPR 2011. IEEE, pp. 1521–1528. http://dx.doi.org/10.1109/cvpr.2011.5995347.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, T., Isola, P., 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. PMLR, pp. 9929–9939, URL https://proceedings.mlr.press/v119/wang20k.html.

Wang, X., Qi, G.-J., 2022. Contrastive learning with stronger augmentations. IEEE Trans. Pattern Anal. Mach. Intell. PP, 1–12. http://dx.doi.org/10.1109/TPAMI.2022.3203630.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q., 2019. Improving adversarial robustness requires revisiting misclassified examples. In: International Conference on Learning Representations. URL https://openreview.net/pdf?id=rklOg6EFwS.

Wu, D., Xia, S.-T., Wang, Y., 2020. Adversarial weight perturbation helps robust generalization. Adv. Neural Inf. Process. Syst. 33, 2958–2969, URL https://proceedings.neurips.cc/paper/2020/hash/1ef91c212e30e14bf125e9374262401f-Abstract.html.

Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K., 2019. Feature denoising for improving adversarial robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 501–509. http://dx.doi.org/10.1109/cvpr.2019.00059.

Yang, Z., Guan, W., Xiao, L., Chen, H., 2024a. Few-shot object detection in remote sensing images via data clearing and stationary meta-learning. Sensors 24 (12), http://dx.doi.org/10.3390/s24123882, URL https://www.mdpi.com/1424-8220/24/12/3882.

Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J., 2022. Unified contrastive learning in image-text-label space. arXiv:2204.03610.

Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., Ghassemi, M., 2024b. The limits of fair medical imaging AI in real-world generalization. Nature Med. http://dx.doi.org/10.1038/s41591-024-03113-4.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64 (3), 107–115. http://dx.doi.org/10.1145/3446776.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: Beyond empirical risk minimization. http://dx.doi.org/10.48550/arXiv.1710.09412, arXiv preprint arXiv:1710.09412.

Zhang, J., Ma, K., 2022. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, Los Alamitos, CA, USA, pp. 16629–16638. http://dx.doi.org/10.1109/CVPR52688.2022.01615, URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01615.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M., 2019. Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning. PMLR, pp. 7472–7482, URL https://proceedings.mlr.press/v97/zhang19p.html.