Group-wise Inhibition based Feature Regularization for Robust Classification

Haozhe Liu†, Haoqian Wu†, Weicheng Xie*, Feng Liu*, Linlin Shen
[1] Computer Vision Institute, College of Computer Science and Software Engineering,
[2] SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
[3] National Engineering Laboratory for Big Data System Computing Technology
[4]Guangdong Key Laboratory of Intelligent Information Processing,
Shenzhen University, Shenzhen 518060, China
{liuhaozhe2019, wuhaoqian2019}@email.szu.edu.cn, {wcxie,feng.liu,llshen}@szu.edu.cn

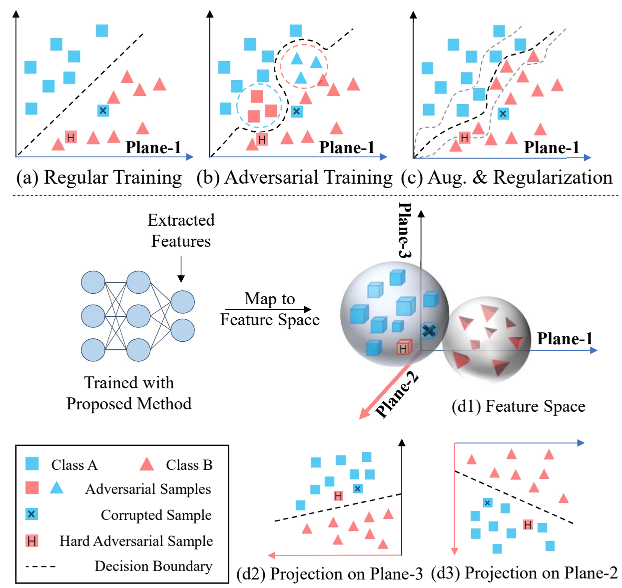
Abstract

The convolutional neural network (CNN) is vulnerable to degraded images with even very small variations (e.g. corrupted and adversarial samples). One of the possible reasons is that CNN pays more attention to the most discriminative regions, but ignores the auxiliary features when learning, leading to the lack of feature diversity for final judgment. In our method, we propose to dynamically suppress significant activation values of CNN by group-wise inhibition, but not fixedly or randomly handle them when training. The feature maps with different activation distribution are then processed separately to take the feature independence into account. CNN is finally guided to learn richer discriminative features hierarchically for robust classification according to the proposed regularization. Our method is comprehensively evaluated under multiple settings, including classification against corruptions, adversarial attacks and low data regime. Extensive experimental results show that the proposed method can achieve significant improvements in terms of both robustness and generalization performances, when compared with the state-of-the-art methods. Code is available at https://github.com/LinusWu/TENET_Training.

1. Introduction

Recent advances in convolutional neural networks (CNNs) have led to far-reaching improvements in computer vision tasks [11, 20]. However, vulnerability of CNNs to image variations, including image corruptions [10] and adversarial samples [8], has not been well resolved yet. Researchers are thus exploring various ways to improve the network robustness against these variations.

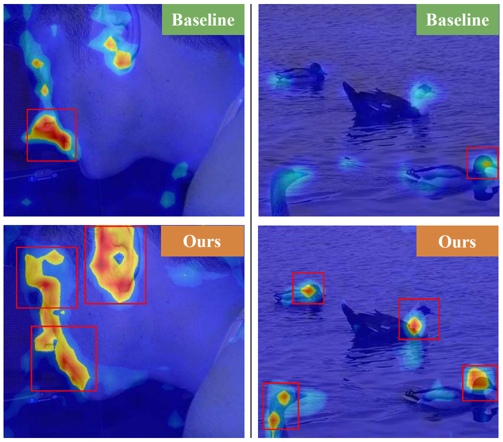
Adversarial training [10, 30, 32] is a typical solution to improve the robustness of CNNs, which includes the attacked samples into the training data, as shown in Fig. 1 (b). Since adversarial training may impair the generalization performance, there is often an inherent trade-off between classification accuracy and adversarial robustness [29, 30]. In order to improve the robustness and generalization simultaneously, data augmentation and regularization methods (e.g. Random Erasing [33], Augmix [14], Cutout[7],



(d) Feature Regularization Motivation of the Proposed Method

Figure 1. Some solutions to improve the robustness of CNN. Unlike with the regular training (a), adversarial training (b) widely utilizes adversarial samples to train a robust CNN. Data augmentation and regularization based method (c) improves the robustness performance by filling up new samples surrounding the decision boundary. The proposed regularization method (d) enables network to increase the representation space (e.g. red auxiliary axis in *d1*) of the features learned by the CNN, and achieves better robustness against corrupted and adversarial samples, with various projections on new planes (e.g. *d2* and *d3*). Best viewed in color.

†Equal Contribution
*Corresponding Author



(a) Single Instance (b) Multiple Instances

Figure 2. The heatmap visualization of feature maps encoded with ResNet-50, based on Grad-CAM [23, 34] with or without the proposed method. Our method locates more diverse discriminative regions (in red boxes) for both single-instance (a) and multiple-instance (b) samples.

Dropout [15] and DeepAugmentation[12]) are proposed. As shown in Fig. 1 (c), these algorithms address data augmentation by randomly generating new samples obeying the same distribution as the training data. Generally, data regularization methods are state-agnostic, which can not be dynamically adjusted during CNN training. Thus, these regularization techniques of CNNs [5, 16, 27, 28] failed to learn features with sufficient diversity. As shown in the first row of Fig. 2, CNNs can locate the most discriminative regions [34] for both single-instance and multi-instance samples with the regularization method, while neglecting other auxiliary features that are critical for the recognition. The lack of auxiliary features may lead to insufficient feature diversity, which consequently results in a feature space with low-dimension for classification and limits the robustness. Meanwhile, current adversarial training and regularization methods concentrate on the global image information by expanding the training set, while the independence of local features is not fully explored. These limitations motivate us to improve the diversity of extracted features by CNNs and devise a non-image-wise regularization strategy to enhance network robustness.

In this paper, we propose a group-wise inhibition based regularization method for improving feature diversity and network robustness, denoted as TENET Training. Fig.1 (d1), (d2) and (d3) show the motivation of the proposed method, where the increase of feature dimension and diversity is beneficial for classification robustness against input variations and adversarial attacks. To increase feature representation space, group-wise feature regularization is proposed to leverage the independence among group-wise features. To improve feature diversity, the proposed algorithm regularizes group-wise features dynamically in each training step. Specifically, based on the grouping of feature maps and their importance evaluation, the group-wise reversed map is proposed to suppress the activation values corresponding to the most significant discriminative regions, and guide the network to learn more auxiliary information in less significant regions. As shown in the second row of Fig. 2, the suppression of most significant discriminative regions is beneficial for exploring more diverse features in CNNs. Experimental results show that the proposed method can improve the top-1 error rate of adversarial training from 36.37% to 31.75%, and outperforms regularization methods significantly in terms of classification accuracy based on small sample. In summary,

- A group-wise inhibition based regularization method is proposed to explore auxiliary features and promote feature diversity.

- Feature maps with different activation distribution are processed separately to learn richer discriminative features hierarchically to better represent images.

- Our proposed method achieves competitive performances in terms of adversarial robustness and generalization compared with related variants and the state of the arts.

2. Related Work

2.1. Robustness against Corruption and Adversarial Attack

The human vision system is robust in ways that CNN based computer vision systems are not [13]. Particularly, a large mount of studies [8, 10, 13, 17] show that CNNs can be easily fooled by small variations in query images, including common corruption [13] and adversarial perturbation [10]. In order to improve the robustness against these variations, studies have been proposed based on various strategies, such as structure modification, adversarial training and regularization. Xie et al. [30] proposed a non-local feature denoising block to suppress the disturbation caused by the malicious perturbation. A Discrete Wavelet Transform (DWT) layer is proposed by Li et al. [21], which disentangles the low- and high-frequency components to yield the noise-robust classification. Different from structure based methods, adversarial training and regularization methods can improve the robustness without the modification of network structure. Adversarial training proposed by Goodfellow et al. [10], in which a network is trained on adversarial examples, is reported to be able to withstand strong attacks [24]. However, there is a trade-off between classification accuracy (generalization) and adversarial robustness. Hence, more and more studies are resorted to the regularization solutions [7, 14, 15, 33] to simultaneously improve generalization and robustness against variations, i.e. common corruption and adversarial attack.

2.2. Regularization for CNNs

Regularization [7, 12, 14–16, 25, 28, 33] has been widely employed in the training of CNNs, where image-wise and feature-wise regularization methods were proposed to improve generalization or robustness. Data augmentation is a typical image-wise solution to regularize the data distribution [7, 12, 14, 33]. Devries et al. [7] proposed a regularization technique to randomly mask out square regions of input during training. Random Erasing proposed by Zhong et al. [33] randomizes the values of pixels in a random rectangle region. Hendrycks et al. [14] proposed Augmix to coordinate simple augmentation operations with a consistency loss. In a nutshell, these image-wise regularization solutions generate images by random operations (e.g. cutout, erasing and mixing), which concentrate on the global information without fully exploring the independence of local features. Meanwhile, the random operations are not dynamically adapted during the training, which limit the feature diversity. These studies motivate us to enhance the feature diversity to improve network robustness and generalization performances.

To explore local information during regularization, feature-wise regularization techniques, including attention based dropout [5], self-erasing [16, 28] and group orthogonal training [4], are proposed. Attention based dropout proposed by Choe et al. [5] utilizes the self-attention mechanism to regularize the feature maps. Self-erasing [16, 28] is an extension method of popular class activation map (CAM) [23, 34], which erases the most discriminative part of CAM, and guides the CNNs to learn classification features from auxiliary regions and activations [27]. However, these methods are proposed for semantic segmentation rather than the classification task. Meanwhile, the steep gradients introduced by the binary mask limit the performances of dropout and erasing operation for classification task. From another aspect, the erasing operation and dropout are global regularizers, which do not fully explore the independence of feature semantics, i.e. different feature groups contain different semantics and should be processed specifically. Group orthogonal training proposed by Chen et al. [4] provides a solution for this problem, which guides CNNs to learn discriminative features from foreground and background separately. Although this group orthogonalization strategy brings improvement of classification performance by enhancing feature diversity, the relied large annotation limits its applicability for general tasks.

In this paper, a regularization method based on group-wise inhibition, namely TENET Training, is proposed to improve network robustness and generalization, which is free of extra annotations. Particularly, a Channel-wise Feature Grouping (CFG) module is proposed to model the channel-wise features in groups. Subsequently, the features in different groups are processed specifically by Group-wise Map Weighting (GMW) module to quantify the importance of each group. Meanwhile, in order to avoid the steep gradients caused by binary mask, a Rectified Reverse Function (RRF) is proposed to smooth group-wise reversed maps. Finally, these reversed maps are used to suppress the activation values to regularize the learned features. Extensive experiments clearly show the significant improvements in terms of robustness and generalization performances.

3. Proposed Method

The overview of the proposed TENET Training is shown in Fig. 3., where CNN is dynamically regularized according to the training step, and significant activation values are suppressed to guide network to explore different features hierarchically. Since the feature maps with the similar activation distribution are prone to contain redundant information, we firstly group the channel-wise feature maps using the proposed CFG module in Section 3.1. In order to further quantify the contribution of each group, the GMW module is introduced in Section 3.2 to evaluate the group importance. Considering the feature groups with negative importance score should contribute less to the classification performance, Rectified Reverse Function (RRF) is proposed to smooth the reversed map of the filtered groups. Following RRF, the group-wise inhibition is devised to suppress the most significant features and explores the less significant auxiliary features, which is introduced in Section 3.3. Finally, we conclude the pipeline of the proposed TENET Training together with the loss design in Section 3.4.

3.1. Channel-wise Feature Grouping Module

According to the pipeline shown in Fig. 3, a feature extraction module $F(\cdot)$ is firstly applied to encode the features set $A = \{a_1, ..., a_j, ..., a_{N_c}\}$ of the input sample x, where a_j is the jth feature map. Since A is prone to contain redundant features, a Channel-wise Feature Grouping module, denoted as CFG module, is introduced to group A to reduce the complexity of feature-wise operation. Given N_c features as input, the corresponding N_G centers are obtained to form the set A_c, which are initialized as a random subset of A. The distance from each feature map of A to the corresponding center is calculated as follows

$$Dist(a_j, A_c[l]) = \frac{1}{H_a \times W_a} \sum_{H_a} \sum_{W_a} (a_j - A_c[l])^2 \quad (1)$$

where $l \in [1, N_G]$ is the index of the center and (H_a, W_a) is the size of a_j. Based on Eq. (1), the centers are updated as similar as k-means clustering. N_G groups are then obtained by grouping the feature maps to the corresponding center. In order to alleviate the influence caused by the random selection, the center searching process is carried out repeatedly in the CFG module. Based on the grouping procedure, the centers are updated according to Center Point Search Function, i.e. CF$(\cdot)$ as follows

$$\mathrm{CF}(\mathit{IDS}) = \{\arg\min_{a_j \in A} dist(a_j, \frac{1}{n_l} \sum_{ID_i = l} a_i) \Big| l \in [1, N_G]\} \quad (2)$$

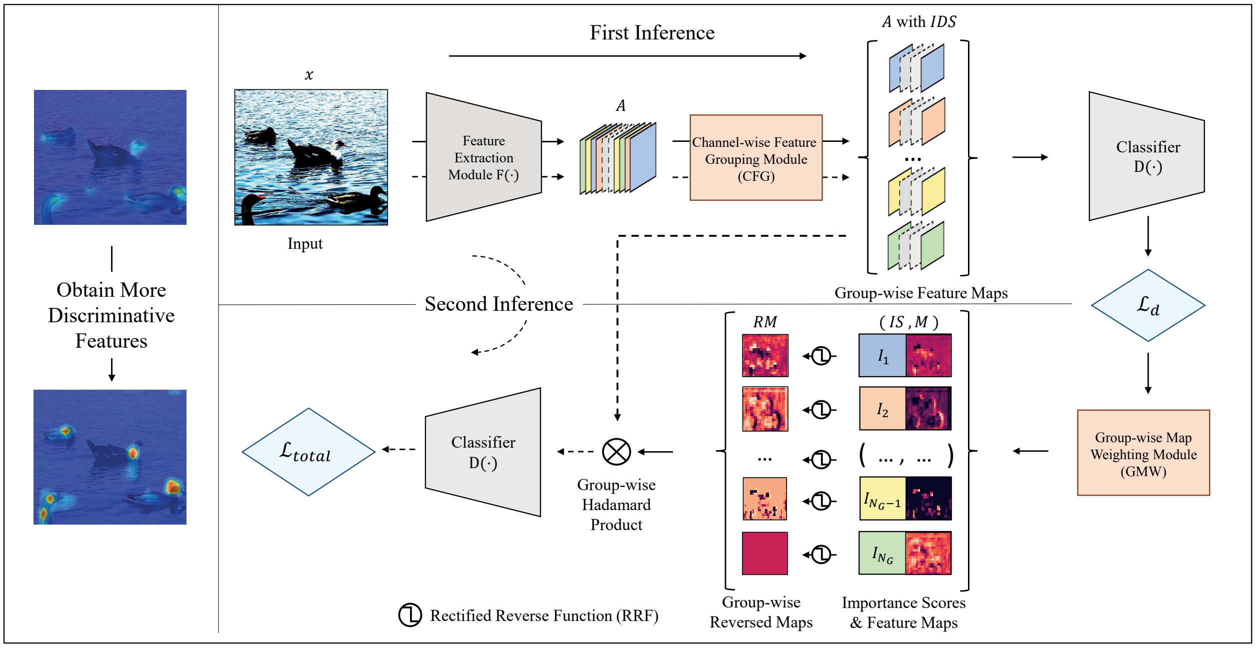


Figure 3. The pipeline of the proposed regularization method (TENET Training). Notice that CNNs consist of the feature extraction module $F(\cdot)$ and the classifier $D(\cdot)$. In the first inference, feature maps A encoded with $F(\cdot)$ are divided into N_G groups by the CFG module, and loss $\mathcal{L}_d$ is calculated based on $D(\cdot)$. Reversed maps RM are then derived using GMW module and RRF. In the second inference, the Hadamard Product of A (with IDS) and RM is fed to $D(\cdot)$ to calculate the loss $\mathcal{L}_{total}$.

where the set $IDS = \{ID_1, ..., ID_j, ..., ID_{N_c}\}$ stands for the set of feature map indices corresponding to each group. ID_j refers to the group index of a_j. n_l is the number of feature maps in the lth group. Based on Eq. (2), A_c can be refined iteratively until CF(·) is stable.

3.2. Group-wise Map Weighting Module

Following feature grouping module, the feature maps are processed in the group-wise mode. To differ the contribution of each group, a Group-wise Map Weighting module, namely GMW module, is proposed to calculate the weight w_j of each a_j as follows

$$w_j = \frac{1}{H_a \times W_a} \sum_{H_a} \sum_{W_a} \frac{\partial \mathcal{L}_d(A)}{\partial a_j}$$
$$\mathcal{L}_d(A) = D(A) \times \text{One-Hot}(D(A)) \tag{3}$$

where $D(\cdot)$ is a classifier, which maps A to the class score. $\mathcal{L}_d(A)$ is the product of prediction and the corresponding one-hot vector of $D(A)$. Since $\frac{\partial \mathcal{L}_d(A)}{\partial a_j}$ is applied to quantify the importance of a_j to the prediction, the group-wise importance scores, i.e. $IS = \{I_1, ..., I_l, ..., I_{N_G}\}$ can be obtained by averaging w_j of each group (ID_j=l) as follows

$$I_l = \frac{1}{N_l} \sum_{ID_j = l} w_j \tag{4}$$

Similar to IS, the group-wise feature maps, i.e. $M = \{m_1, ..., m_l, ..., m_{N_G}\}$ can be obtained by averaging the weighted feature maps as follows

$$m_l = \frac{1}{N_l} \sum_{ID_j = l} w_j \times a_j \tag{5}$$

3.3. Group-wise Inhibition using Rectified Reverse Function

Based on the importance scores, group-wise feature maps are applied to obtain the reversed map set, i.e. $RM = \{rm_1, ..., rm_l, ..., rm_{N_G}\}$. Since the steep gradients introduced by the binary mask may limit the classification performance, the reversed maps are further smoothed. Meanwhile, considering the feature groups with negative importance scores should contribute less to the update of the reversed mask, we therefore propose a Rectified Reverse Function, i.e. RRF(·), to obtain the reversed maps as follows

$$rm_l = \text{RRF}(m_l, I_l) = sgn(I_l > 0) \times \frac{1}{1 + e^{m_l}} \tag{6}$$

where $sgn(\cdot)$ is the sign function. Due to the negative correlation between m_l and rm_l, the computation of RM is deemed as a reversed map. Based on RM, the group-wise inhibition is formulated as follows

$$\hat{y} = D(RM \otimes A) \tag{7}$$

where $D(\cdot)$ is a classifier with the input of A and $\hat{y}$ refers to the predicted label of the group-wise inhibition. $\otimes$ refers to the group-wise Hadamard product.

3.4. Loss Design of TENET Training

While $\hat{y}$ is obtained by group-wise inhibition, $F(\cdot)$ and $D(\cdot)$ can be directly learned based on the loss $\mathcal{L}_c(y, \hat{y})$, i.e. the cross entropy for single-label classification or binary cross entropy for multi-label classification. The group-wise inhibition reduces the variation between groups, while it may introduce invalid activation units in $F(\cdot)$ or $D(\cdot)$. To regularize these activation units, an orthogonal loss $\mathcal{L}_o(A)$ is adopted, which is formulated as follows

$$\mathcal{L}_o(A) = \prod_{l=1}^{N_g}(\sum_{j=1}^{N_c}(sgn(ID_j = l) \times a_j)) \tag{8}$$

From another aspect, by mapping rm_l into the region of [0, 1], the magnitude of back-propagation gradients is suppressed for $F(\cdot)$ and $D(\cdot)$. To alleviate vanishing gradient problem, a general classification loss, i.e. $\mathcal{L}_c(y_i, D(A))$, is employed. Finally, the total loss is formulated as follows

$$\mathcal{L}_{total} = \mathcal{L}_c(y_i, D(A)) + \alpha\mathcal{L}_c(y_i, \hat{y}) + \mu\mathcal{L}_o(A) \tag{9}$$

where α and μ are the hyper parameters. For clarity, TENET Training is summarized in Algo. 1

Algorithm 1 TENET Training

Input:
Training Sample: x
Initialization of $F(\cdot)$ and $D(\cdot)$
Output:
Trained CNNs: $F(\cdot)$ and $D(\cdot)$

1: **for** all training steps **do**
2: Extract A from $F(x)$;
3: Obtain IDS of A using CFG Module according to Eqs. (1) and (2);
4: Derive (IS, M) with GMW Module according to Eqs. (3), (4) and (5);
5: Employ RRF to obtain RM according to Eq. (6);
6: Obtain $\hat{y}$ according to Eq. (7);
7: Calculate $\mathcal{L}_{total}$ according to Eqs. (8) and (9);
8: Update $F(\cdot)$ based on $\frac{\partial \mathcal{L}_{total}}{\partial F}$ and update $D(\cdot)$ based on $\frac{\partial \mathcal{L}_{total}}{\partial D}$;
9: **end for**
10: Return $F(\cdot)$ and $D(\cdot)$.

4. Experimental Results and Analysis

As listed in Table 1, to evaluate the performance of the proposed method, extensive experiments are carried on publicly-available data sets, including PASCAL VOC 2012 [9], CIFAR-10/100 [18], ImageNet-C [13] and CUB-200 [26]. We firstly introduce the employed data sets and the corresponding implementation details. The performance of the proposed method on standard image classification task is evaluated, and the encoded feature maps are visualized for the algorithm analysis. Finally, both the robustness and generalization performances of the proposed method are evaluated based on the comparison with the state-of-the-art methods.

Table 1. Summary of Experiment Configurations and TENET Training Gains.

Task-[protocol]	Dataset	Previous SOTA	Gain
Standard Classification-[4]	PASCAL VOC 2012[9]	Group Orthogonal Training [4]	**2.9%**
Robustness against Adversarial Attack-[8, 24]	CIFAR-10/100 [18]	A. T. [24] Augmix[14]	**5.75%** **15.56%***
Robustness against Common Corruption-[13, 14, 21]	CIFAR-10/100-C [13] ImageNet-C [13]	Augmix[14]	**1.77%** **2.8%**†
Generalization-[2]	CUB-200 [26]	GLICO [2]	**2.75%**

* The gain is obtained in CIFAR-10 against FGSM (8/255).
† The gain is obtained by following 90-epoch Protocol [21].

4.1. Data Sets and Implementation Details

We evaluate the performance of TENET Training from three aspects, i.e. standard classification, robustness and generalization (see Table 1).

Standard Classification. In this case, ResNet-18 [11] is selected as the backbone in our TENET Training. PASCAL VOC 2012 [9] is used for the evaluation, while 5,717 and 5,823 images are used for the training and validation, respectively. The protocol in [4] is adopted. The CNNs for evaluation are pretrained on the ImageNet [6], and fine-tuned on PASCAL VOC 2012 training set. In the training stage, the shorter side of image is resized to a random value within [256,480] for the scale augmentation. The resized image is then randomly cropped to the size of 224×224 for the training based on the batch size of 256. In the testing stage, ten-crop testing is used to evaluate the performance.

Robustness. In this case, the robustness of the proposed algorithm against both adversarial attack and image corruption is evaluated on CIFAR 10/100 [18], CIFAR 10/100-C [12] and ImageNet-C [12]. ResNeXt-29 [31] and ResNet-50 [11] are chosen as the backbones. To test the robustness of the proposed method against adversarial attacks, two popular attacks, FGSM [10] and PGD [1], are employed. The performance is then evaluated according to the protocol in [8]. The perturbation budget (ϵ) is set to $8/255$ or $4/255$ under l_∞ norm distance for the two attacks. PGD-K stands for K-step attack with a step size of $2/255$. Meanwhile, adversarial training is used to defense powerful iterative attacks of PGD. To make the results more convincing, an efficient adversarial training method (free-AT) [24] is adopted, where the *hop step* of free-AT, i.e. m, is set to 4.

Against image corruption, 15 different kinds of corruptions, such as noise, blur, weather and digital corruptions, are performed on CIFAR 10/100-C and ImageNet-C for the

Table 2. The Ablation Study of the Proposed Method on the Validation Dataset of Pascal VOC 2012 in terms of Average Precision (%).

Baseline	Channel-wise Inhibition	Group-wise Inhibition	L_o	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	prsn	plant	sheep	sofa	train	tv	**mean**
√	×	×	×	94.8	83.8	91.5	79.4	56.6	88.2	78.9	90.8	64.8	61.5	57.9	90.9	73.7	83.8	96.0	51.6	77.1	58.2	89.8	77.1	77.1
√	√	×	×	94.2	82.8	**92.9**	83.3	62.2	90.8	81.0	**92.8**	71.1	74.1	63.0	88.2	83.9	88.5	93.5	**58.4**	**85.2**	64.7	93.1	80.6	81.2
√	×	√	×	93.9	81.7	92.5	**83.7**	**63.8**	90.9	82.7	91.5	69.5	76.4	64.6	89.6	**85.9**	**89.3**	**96.5**	58.1	84.6	64.5	93.2	**83.7**	81.8
√	×	√	√	**95.6**	**84.3**	91.1	83.1	61.3	**91.4**	**83.2**	91.6	**72.8**	**77.4**	**65.9**	**91.3**	84.4	89.2	96.3	57.4	83.9	**67.6**	**94.5**	83.1	**82.3**

Table 3. Performance Comparison between the Proposed Method and the State of the Arts on the Validation Dataset of Pascal VOC 2012 in terms of Average Precision (%).

Model	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	prsn	plant	sheep	sofa	train	tv	**mean**
ResNet18[11] reported in [4]	95.2	79.3	90.2	82.8	52.6	90.9	78.5	90.2	62.3	64.9	64.5	84.2	81.1	82.0	91.4	50.0	78.0	61.1	92.7	77.5	77.5
ResNet18 trained in this paper	94.8	83.8	91.5	79.4	56.6	88.2	78.9	90.8	64.8	61.5	57.9	90.9	73.7	83.8	96.0	51.6	77.1	58.2	89.8	77.1	77.1
GoCNN [4]	**96.1**	81.0	90.8	**85.3**	56.0	**92.8**	78.9	91.5	63.6	69.7	65.1	84.8	84.0	83.9	92.3	52.0	83.9	64.2	93.8	78.6	79.4
TENET (Binary Mask)	93.2	83.8	91.3	83.2	59.8	91.6	79.6	90.6	66.3	75.2	62.1	89.7	84.7	88.4	96.3	**58.0**	**87.0**	65.2	93.1	82.1	81.1
TENET (Instance-wise Inhibition)	93.1	82.7	**92.6**	82.9	61.1	90.9	81.8	**91.6**	70.6	73.7	63.3	**91.5**	**85.6**	88.5	**96.4**	56.8	85.1	61.8	93.2	82.3	81.3
TENET	95.6	**84.3**	91.1	83.1	**61.3**	91.4	**83.2**	**91.6**	**72.8**	**77.4**	**65.9**	91.3	84.4	**89.2**	96.3	57.4	83.9	**67.6**	**94.5**	**83.1**	**82.3**

evaluation, and each kind of corrupted data has five different severity levels [12]. We follow the training protocols and evaluation metrics used in Augmix [14] and WResNet50 [21]. The *Clean Error* is the regular classification error on the original (uncorrupted) test or validation dataset, and *mCE (Mean Corruption Error)* for CIFAR-10/100-C is the mean over all 15 corruptions. Meanwhile, the *mCE* for ImageNet-C is normalized by the corruption error of AlexNet [19]. Due to the computational efficiency, Augmix without Jensen-Shannon divergence (JSD) loss is implemented.

Generalization. Since CUB-200 [26] contains only 30 images for each of the 200 species of birds, it is used as a popular benchmark to test the generalization of CNNs. We follow the protocol in [2], and evaluate the generalization with three numbers of samples per class (SPC) for training, i.e. 10, 20 and 30. For a fair comparison, the same ResNet-50 [11] in the protocol [2] is adopted as the backbone. To train the CNNs, the smaller side of the images from CUB-200 is resized to 256, the scaled images are then randomly cropped to the size of 224×224. In the testing stage, the prediction is based on the center cropping with the size of 224×224.

TENET Training. For the hyper parameter setting, the cluster number N_G is set to 6, while α and μ are set as 0.1 and 0.1, respectively.

The public platform pytorch [22] is used for the implementation of all the experiments on a work station with CPU of 2.8GHz, RAM of 512GB and GPU of NVIDIA Tesla V100.

4.2. Effectiveness Analysis of the Proposed Method

Ablation Study. To quantify the contribution of each module in TENET Training, we test the discriminative performance of the variant with or without this module. Table 2 shows the results carried out for standard classification. Since GMW is based on CFG module, these two modules denoted as Channel-wise Inhibition and Group-wise Inhibition are evaluated integratedly. Table 2 shows that the performance of the baseline in the first row can be improved by both channel-wise inhibition and group-wise inhibition. Specifically, an improvement of 4.1% in terms of mAP is achieved by channel-wise inhibition. To study the performance of GMW and CFG modules, Table 2 shows that the group-wise inhibition further improves the performance using $\mathcal{L}_o(A)$. The most significant improvement of TENET Training happens when all the proposed modules are employed, i.e. the proposed method achieves a mAP of 82.3%, which largely outperforms the baseline with a mAP of 77.1%.

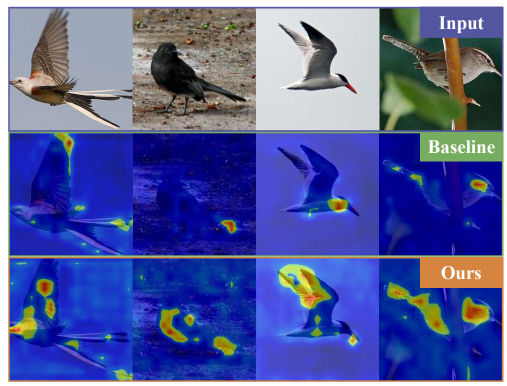


Figure 4. The visualization of the discriminative regions for image classification of CUB-200 using Grad-CAM [23, 34]. The 1st-3rd rows show the input samples, the discriminative regions extracted by ResNet-50 and the results based on TENET Training.

Visualization of TENET Training. To study the diversity of the learned features with the proposed TENET Training, we visualize the discriminative regions of the input samples from CUB-200 using Grad-CAM [23, 34] in Fig. 4. Compared with the baseline, the CNN using TENET Training derives more discriminative regions, such as wings, heads and tails, for classification.

To study the distribution of the extracted features, we

Table 4. Top-1 error rates (%) on ImageNet and Top-1 mCE rates (%) on ImageNet-C with ResNet-50. Aug. stands for Augmix.

	Protocol	Clean	Noise			Blur				Weather				Digital				
		Error	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	**mCE**
Baseline [11] reported in [14]		23.8	79	80	82	82	90	84	80	86	81	75	65	79	91	77	80	80.6
Cutout [7]	90-epoch Protocol[21]	23.2	79	81	80	77	90	80	81	80	78	70	61	74	87	74	75	77.7
WResNet50 (Haar) [21]		23.1	77	79	79	71	86	77	77	80	75	66	57	71	84	75	77	75.3
Augmix [14]		23.0	71	71	71	72	88	72	72	78	78	67	60	72	86	75	76	73.9
TENET		23.1	73	78	75	74	87	76	80	79	78	67	63	73	84	72	71	75.3
TENET (Aug.)		22.8	69	69	69	**69**	87	69	70	76	75	**64**	**56**	69	82	72	73	71.1
Augmix [14]	180-epoch Protocol[14]	22.5	**68**	69	70	73	**81**	69	**67**	75	**73**	67	61	61	**80**	71	72	70.5
TENET (Aug.)		**22.4**	69	**67**	**68**	72	**81**	**66**	69	**74**	74	65	59	**60**	82	**69**	**70**	**69.6**

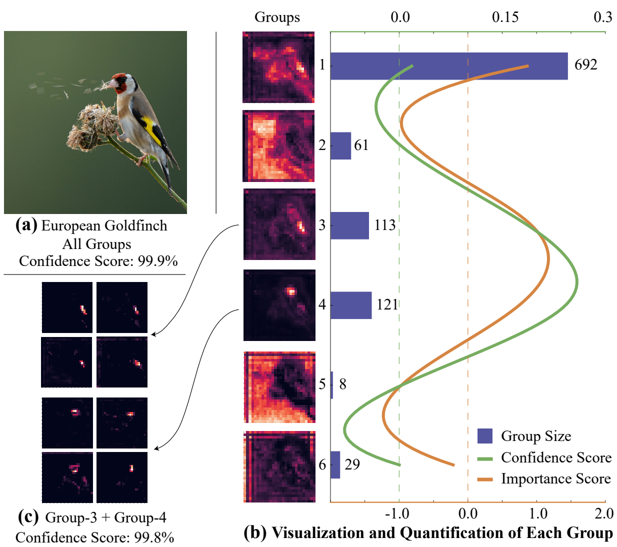


Figure 5. The visualization and quantification of the feature maps extracted by the 3rd residual block of ResNet-50 using TENET Training. (a) An input image with the label of European Goldfinch. (b) The activation distribution, the corresponding importance and confidence scores of each group clustered by CFG module. (c) The example feature maps selected from the 3rd and 4th groups.

further visualize the group-wise maps with different importance scores of the input image in Fig. 5, where feature maps are clustered into six groups. The confidence score of each group corresponds to the variant with or without the selected group. Fig. 5 (b) shows that, the importance score (orange line) calculated by GMW module is similar with confidence score (green line) in tendency, which illustrates the effectiveness of the GMW module. Meanwhile, Fig. 5 (b) also shows the large variations among the activation distributions of the group-wise features, which indicates the reasonability of group-independent processing. As a contrast, the instance-wise operation involved in traditional methods can not regularize the most important features but only the features with the largest group size (i.e. group-1 in Fig. 5), based on the average of activation maps or annotations. Thus, the proposed group-independent processing can facilitate our TENET training to achieve better performance than other regularization methods.

Fig. 5 (c) shows that group-3 and group-4 out of six groups are the most important for CNN, which can improve the confidence score output by the CNN from 0 to 99.8%. Group-1 is relatively less important than group-3 and group-4 but can increase confidence score, while the impacts of group-2, 5 and 6 on the classification performance is very limited. More precisely, when these three groups are not used, the confidence score has dropped by only 0.09%. This observation indicates that the inhibition of the important groups can help improve the efficiency without losing accuracy. Hence, in the proposed method, we only regularize the groups with higher importance scores.

4.3. Comparison with Related Methods

Comparison in Standard Classification. To study the classification performance of the proposed method, we compare it with the group orthogonal training [4] denoted as GoCNN in Table 3. In additional, we include TENET (Binary Mask) and TENET (Instance-wise Inhibition) for the comparison. TENET (Binary Mask) refers to the proposed method that suppresses the activation value using binary masks rather than the smoothed reversed maps. In TENET (Instance-wise Inhibition), CFG module and GMW module are replaced by Grad-CAM [23, 34], which process features by instance-wise operation. Table 3 shows that TENET Training outperforms the competing methods significantly. The proposed method achieves a mAP of 82.3%, exceeding group orthogonal training by 2.9% absolutely. This indicates group-wise inhibition using the smoothed reversed maps is suitable for classification. Meanwhile, the proposed method uses less information than group orthogonal training, i.e. large-scale dense annotations, e.g. segmentation or localization labels, are not demanded. While state-agnostic inhibition used in group orthogonal training regularizes features in a coarse way, it limits both the accuracy and efficiency. However, based on the proposed group-wise inhibition, our method can consistently improve the classification performance, and does not demand any extra annotations.

Comparison in Robustness. We compare the proposed method with two state-of-the-art regularization methods [7, 14], a wavelet integrated method [21] and an adversarial training one [24], for robustness evaluation against image corruption and adversarial attacks in Tables 4, 5 and 6. One can observe that TENET Training outperforms the competing methods in each case. For the recognition

Table 5. Top-1 error rates (%) on CIFAR-10 and Top-1 mCE rates (%) on CIFAR-10-C trained with various methods based on ResNeXt-29. A.T. stands for Adversarial Training. The brackets following the adversarial attack method show the perturbation budget (ϵ).

	Clean	mCE	FGSM (8/255)	PGD-7 (4/255)	PGD-100 (8/255)
Baseline [31]	5.72	29.88	72.81	94.15	-
Cutout [7]	3.97	29.20	71.07	97.19	-
Augmix [14]	3.95	13.32	76.03	93.67	-
TENET	3.89	26.46	61.05	91.28	-
TENET (Aug.)	**3.50**	**12.31**	60.47	90.45	-
A.T. [24]	-	-	36.37	22.61	42.82
TENET (A.T.)	-	-	**31.75**	**20.07**	**37.07**

against image corruption, the best performance is achieved with the combination of TENET Training and Augmix (denoted as TENET(Aug.)), which achieves 69.6%, 12.31% and 35.73% error rates on ImageNet-C, CIFAR-10-C and CIFAR-100-C, respectively. Augmix [14] with JSD loss can achieve a mCE of 68.4% on ImageNet-C, while it requires three times the GPU memory and runtime cost compared with the proposed method.

For robustness against adversarial attacks, two attack paradigms, namely FGSM and PGD, are employed to test the trained CNNs with different regularization methods. Tables 5 and 6 show that the CNNs using the proposed method outperform those with other regularization methods by a large margin. When FGSM is considered, our method can achieve an error rate of 60.47%, exceeding other regularization methods by around 10% absolutely. Meanwhile, our method is complementary to the Adversarial Training (denoted as A.T.). Typically, the proposed method achieves the error rates of 37.07% and 63.13% against PGD-100 on CIFAR-10/100, which outperforms Adversarial Training clearly, i.e. 37.07% vs. 42.82% and 63.13% vs. 65.17%.

Table 6. Top-1 error rates (%) on CIFAR-100 and Top-1 mCE rates (%) on CIFAR-100-C trained with various methods based on ResNeXt-29.

	Clean	mCE	FGSM (8/255)	PGD-7 (4/255)	PGD-100 (8/255)
Baseline [31]	23.33	53.40	85.93	95.96	-
Cutout [7]	20.73	54.60	87.03	98.13	-
Augmix [14]	21.83	37.50	84.65	95.32	-
TENET	20.56	51.21	78.71	94.62	-
TENET (Aug.)	**19.46**	**35.73**	75.28	93.54	-
A.T. [24]	-	-	60.13	47.99	65.17
TENET (A.T.)	-	-	**58.60**	**46.17**	**63.13**

Comparison in Generalization. To further study the generalization performance achieved by TENET Training, we compare the proposed method with regularization methods [3, 7, 33], data augmentation method [2] and adversarial training [24] in Table 7. Table 7 shows the evident improvements of TENET Training over other methods in every case. Typically, when 20 samples per class are used for training, the proposed method can achieve 76.91% in terms of Top-1 accuracy. As a comparison, adversarial training [24] achieves the Top-1 accuracy of only 57.91% in this case. It seems that adversarial training can improve the robustness, while it may also largely impair the generalization performance. Hence, Table 7 illustrates that the proposed method can better maintain the generalization performance compared with other methods.

Table 7. Comparison of TOP-1 Accuracy (%) for CUB-200 based on ResNet-50 with Different Numbers of Training Samples Per Class (SPC).

Methods	SPC = 10	SPC = 20	SPC = 30
MixMatch [3]	36.02	60.57	70.41
Random Erase [33]	63.72	66.14	73.74
Cutout [7]	64.33	68.47	74.97
GLICO [2]	65.13	74.16	77.75
A.T. [24]	44.53	57.91	63.67
TENET	**66.07**	**76.91**	**80.34**

5. Conclusion

In this paper, we proposed a group-wise inhibition based feature regularization method to improve the robustness and generalization of CNNs. In the proposed algorithm, CNN is regularized dynamically when learning, where the most discriminative regions with significant activation values are suppressed to enable the network to explore more diverse features. Richer features then help to better represent images even with malicious variations. The effectiveness of the proposed method was verified in terms of standard classification, adversarial robustness and generalization performance based on small number of training samples.

Acknowledgment

The work is partially supported by the National Natural Science Foundation of China under grants no. 62076163, 91959108, 61602315 and U1713214,the Science and Technology Project of Guangdong Province under grant no. 2020A1515010707, the Shenzhen Fundamental Research fund JCYJ20190808163401646, JCYJ20180305125822769 and JCYJ20190808165203670, and Tencent "Rhinoceros Birds"-Scientific Research Foundation for Young Teachers of Shenzhen University.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circum-

venting defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[2] Idan Azuri and Daphna Weinshall. Learning from small data through sampling an implicit conditional generative latent optimization model. In *ICPR*, 2021.

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NIPS*, pages 5049–5059, 2019.

[4] Yunpeng Chen, Xiaojie Jin, Jiashi Feng, and Shuicheng Yan. Training group orthogonal neural networks with privileged information. In *IJCAI*, pages 1532–1538, 2017.

[5] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, pages 2219–2228, 2019.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[8] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *CVPR*, pages 321–331, 2020.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

[14] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *ICLR*, 2020.

[15] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[16] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NIPS*, pages 549–559, 2018.

[17] Brett Jefferson and Carlos Ortiz Marrero. Robust assessment of real-world adversarial examples. In *CVPRW*, pages 792–793, 2020.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[21] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavelet integrated cnns for noise-robust image classification. In *CVPR*, pages 7245–7254, 2020.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[24] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NIPS*, pages 3358–3369, 2019.

[25] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *CVPR*, June 2015.

[26] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[27] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, pages 12275–12284, 2020.

[28] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, pages 1568–1576, 2017.

[29] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *CVPR*, pages 819–828, 2020.

[30] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, pages 501–509, 2019.

[31] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017.

[32] Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. Efficient adversarial training with transferable adversarial examples. In *CVPR*, pages 1181–1190, 2020.

[33] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020.

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.