https://doi.org/10.2352/EI.2023.35.9.IPAS-296

ILIAC: Efficient classification of degraded images using knowledge distillation with cutout data augmentation

Dinesh Daultani [1]***, Masayuki Tanaka*** [1]***, Masatoshi Okutomi*** [1]***, Kazuki Endo*** [2]
[1] ***Tokyo Institute of Technology, Tokyo, Japan*** [2] ***Teikyo Heisei University, Tokyo, Japan***

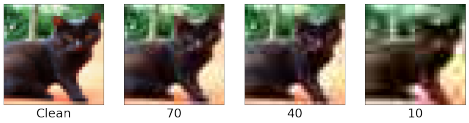
Abstract

Image classification is extensively used in various applications such as satellite imagery, autonomous driving, smartphones, and healthcare. Most of the images used to train classification models can be considered ideal, i.e., without any degradation either due to corruption of pixels in the camera sensors, sudden shake blur, or the compression of images in a specific format. In this paper, we have proposed a novel CNN-based architecture for image classification of degraded images based on intermediate layer knowledge distillation and data augmentation approach cutout named ILIAC. Our approach achieves 1.1% and 0.4% mean accuracy improvements for all the degradation levels of JPEG and AWGN, respectively, compared to the current state-of-the-art approach. Furthermore, ILIAC method is efficient in computational capacity, i.e., about half the size of the previous state-of-the-art approach in terms of model parameters and GFlops count. Additionally, we demonstrate that we do not necessarily need a larger teacher network in knowledge distillation to improve the model performance and generalization of a smaller student network for the classification of degraded images.

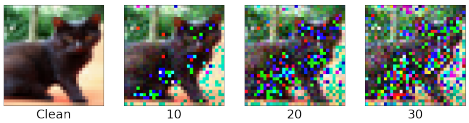
Introduction

Imaging is vital in various areas, such as smartphones, astronomy, medical, robot vision, and self-driving vehicles. A wide variety of research is performed on computer vision tasks where images can be considered ideal; in this paper, we refer to them as clean images. However, different imaging sensors or cameras captured data can be degraded for several reasons, such as motion blur, low-light conditions, and image compression. Hence, the performance decreases when the camera sensor data is non-ideal or degraded. Furthermore, most of the research in the computer vision field is aimed at solving clean images task and can perform poorly when the image quality is degraded [2, 3, 4, 5].

There can be a wide variety of degradations possible in the real world. For example, (1) JPEG compression performed in camera sensors or post-processing pipelines can lead to lossy pixel information. (2) Random noise can occur in an image due to low light or extreme conditions. (3) While taking a photograph, an image can be blurred due to a sudden shake in the camera sensor. (4) Due to defective memory hardware or camera sensor defects, image pixels can be corrupted. This paper mainly deals with the image classification task of computer vision. To systematically simulate and measure the performance in the degradation conditions of image classification, we have discussed two types of degradation: JPEG and additive white Gaussian noise (AWGN), similar to the degradation methods used in the experiments of Endo et al. [2, 3, 4]. JPEG and AWGN compression degradation levels vary between 0 - 100 and 0 - 50, respectively.



(a) JPEG degradation where 70, 40, and 10 represent quality factors levels.



(b) Additive white Gaussian noise (AWGN) degradation where 10, 20, and 30 represent noise levels.

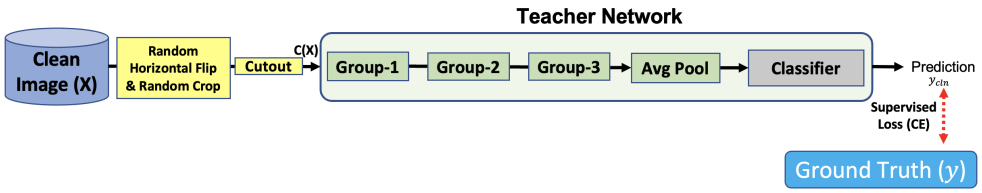
Figure 1: Sample images for different degradation levels on JPEG and AWGN degradation methods.

Figure 1 shows what the input image looks like for two degradation methods and different degradation levels.

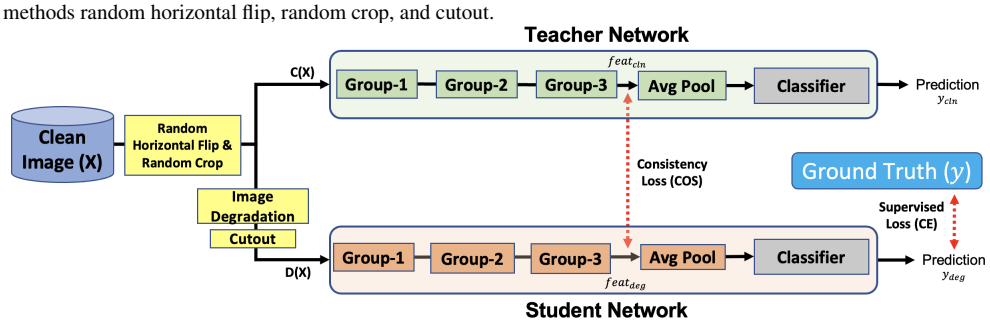
Large models cannot be adequately deployed in limited hardware applications. However, as far as we know, no previous research has investigated the computational efficiency of the methods proposed to solve the different tasks of computer vision on degraded images. In this paper, we measure the efficiency of the proposed model in terms of model parameters and GFlops count. Additionally, rather than increasing the size of our network, we proposed an efficient architecture based on the Intermediate Layer knowledge dIstillAtion and Cutout (ILIAC) to improve the model performance as discussed in section Proposed Method.

The main contributions of our work are summarized as follows:

1. We introduced a novel knowledge distillation-based approach with a cutout method of data augmentation for image classification of degraded images that achieves state-of-the-art performance on the CIFAR-100 dataset for the JPEG and AWGN degradation methods.
2. We empirically demonstrate that the cutout approach of data augmentation can be applied during the teacher and student network training in the distillation process for improving the performance from 0.864 to 0.882, i.e., $\sim$2% mean accuracy improvements for image classification of CIFAR-10 dataset on JPEG degradation.
3. ILIAC is efficient in terms of model parameters and GFlops count since it is about 50% lighter as compared to the previous state-of-the-art method.
4. We have given evidence that in the knowledge distillation setting, a larger teacher network does not necessarily improve the model performance of a smaller student network in the image classification of degraded images.



(a) Step-1: Train teacher network using supervised loss, i.e., cross-entropy loss, while applying data augmentation methods random horizontal flip, random crop, and cutout.



(b) Step-2: Distillation process to train the student network using consistency and supervised loss. The teacher network is non-trainable during the step-2 distillation process. We apply random horizontal flip and random crop data augmentation to teacher and student networks while applying cutout to only student networks.

Figure 2: The above architecture illustrates two steps approach that describes our ILIAC method's training process. Teacher network, student network, and data augmentation methods are represented in green, orange, and yellow colors.

Literature Review

Often researchers have used ensemble networks [3], restoration networks [7], estimation of degradation parameters [2] - [4] for the classification of degraded images. Ensemble networks typically have additional modules for the prediction, increasing the model's computational parameters [3]. Restoration networks typically deal with the restoration of images by applying different types of filters depending upon the type of degradation [14], or a neural network [3] to the original form before performing the end task. In the case of networks with degradation parameters estimator, there is an assumption that degradation levels will be accessible during the training phase [3, 4]. Due to the limitations of the above-proposed approach in terms of either additional model computational parameters or inaccessibility of degradation levels, it makes them non-optimal to deploy in real-world applications with limited hardware, such as self-driving vehicles, drones, and smartphones. On the other hand, several approaches are used to design efficient neural network architectures for limited hardware systems, such as quantization, network pruning, knowledge distillation, weight sharing, and neural architecture search [12].

To deal with the constrained-resource computational requirement for real-world applications, Hinton et al. [1] have introduced the Knowledge Distillation approach. The main idea behind using Knowledge Distillation is to train a teacher network, i.e., a large network or ensemble of networks, to learn the structure from the data and then leverage the information learned to train a much smaller network, i.e., Student Network. A similar approach to knowledge distillation for image classification of degraded images has been proposed in [4] as Direct Extractor and Feature Adjustor method. However, in the case of both networks, there was a degradation levels estimator, which makes it impractical when in real work datasets, we do not have access to the image's degradation levels. Besides, the Feature Adjustor method contains two feature extractors, which substantially increases the computational parameters of the proposed network.

In addition, as an alternative way to improve the performance using knowledge distillation, we used a data augmentation method cutout [13] which works as a regularization to the model and does not lead to an increase in the computational capacity of the network. The method cutout is defined as a way to randomly mask images during the training phase, which increases the model's robustness and performance. There has been similar work done in Stanton et al. [17] related to applying data augmentation methods to Knowledge Distillation. However, the data augmentation approach cutout was not covered in their study. Additionally, our study focuses on at which place we should apply data augmentation methods such as cutout in the knowledge distillation process.

Proposed Method

In this paper, we have proposed an intermediate layer knowledge distillation-based approach along with the data augmentation technique cutout to improve the performance of classification networks for degraded images, as shown in Figure 2. More details about cutout settings are discussed in the Experiments (Cutout) section. The first step is to train a teacher network while applying data augmentation methods: random horizontal flip, random crop, and cutout, which gives us the output $C(X)$ as shown in Figure 2 (a). We pass this clean image as the input to the teacher network, which is trained using a supervised loss function, i.e., cross-entropy applied between the ground truth y and prediction y_{cln}.

The second step is to train a student network through the knowledge distillation process, where we first initialize weights of teacher and student network with the pre-trained model on clean images. We input clean images $C(X)$ and degraded images $D(X)$ after applying data augmentation methods as shown in Figure 2 (b) to the teacher and student networks, respectively. During the

distillation process, we apply consistency loss function g after the feature extractor (group-3 in our case) between the outputs $feat_{cln}$ and $feat_{deg}$ i.e. represented as $L_{con} = g(feat_{cln}, feat_{deg})$ and supervised loss function h between the ground truth y and prediction y_{deg} i.e. represented as $L_{sup} = h(y, y_{deg})$. γ_{con} and γ_{sup} are the weights of the respective loss functions. We formalize the loss function equation for the second step, i.e., the distillation process of our ILIAC method, as follows:

$$L = \gamma_{con} L_{con} + \gamma_{sup} L_{sup}. \tag{1}$$

Specifically, Cosine Embedding (COS) loss is applied as a consistency loss function to transfer the knowledge from the teacher network to the student network. Moreover, a cross-entropy loss is applied as the supervised loss function that calculates the difference between probability distributions of ground truth y and prediction y_{deg}. Although we have primarily used COS as a consistency loss function, we could easily replace it with other loss functions that could be applied in the intermediate layers during the knowledge distillation. For example: Mean Squared Error (MSE), Attention Transfer (AT) [15], and Factor Transfer (FT) [15].

In our primary experiments, we use same backbone as Endo et. al. [4] i.e. PyramidNet [10] with Shake drop regularization [11] for the knowledge distillation. However, ILIAC differs from Endo et al. [4] since it does not require additional estimators such as scale, bias, and degradation level estimators in the architecture; hence there is no additional increase in computational parameters of the network. What's more, our approach can be applied to different types of CNN-based backbones such as ResNet [8] , VGG [9], and PyramidNet [10] irrespective of number of groups shown in the proposed architecture Figure 2. Group numbers in the backbone are shown to represent where exactly consistency loss is applied during the distillation process.

Experiments

In this section, we will discuss experiment settings, datasets and preprocessing, evaluation metrics, preliminary experiments, model comparisons, and results analysis as follows:

Experiment settings

There are two different experimental setups for training teacher networks and student networks. To train teacher networks, we have used an SGD optimizer with an initial learning rate of 0.1, Nesterov momentum of 0.9, and weight decay 5e-4. In addition, we have used a multi-step learning rate scheduler with milestones [60, 120, 160] and gamma 0.2. On the other hand, to train student networks, we have used a RAdam optimizer with an initial learning rate of 0.001 and a weight decay value of 1e-4. In addition, we have used cosine annealing learning rate scheduler with T_{max} = total epochs. As shown in Endo et al. [5] performance of COS loss is better than KLD and MSE loss; hence we have used COS embedding loss for knowledge distillation. Furthermore, we have trained the teacher networks and student networks for 200 and 100 epochs, respectively.

In addition, we have performed our experiments mainly on ResNet [8], and PyarmidNet [10] with Shake drop regularization [11] backbones. Like Endo et al. [4], we name PyramidNet with Shake drop regularization as ShakePyramidNet. We mainly use the ShakePyramidNet backbone with depth=110 and alpha=270, i.e., represented as ShakePyramidNet110. For ResNet, we have used ResNet20, ResNet56, and ResNet110 backbones, where the numbers represent the number of layers in the network. Moreover, Knowledge Distillation backbones are represented in the notation of BackboneA-B, where A and B represent the number of layers in the teacher and student network, respectively. For example: ResNet20-56 represents ResNet20 and ResNet56 backbones for teacher and student network respectively.

Datasets and preprocessing

We have mainly used CIFAR-10 and CIFAR-100 [6] datasets in our paper which are widely used for image classification. Since predominantly CIFAR-10 and CIFAR-100 datasets were used to measure the performance of image classification methods on different degradation levels in previous research [2, 3, 4], we have mainly used the same datasets as well. In addition, we have primarily used the CIFAR-100 dataset for proposed method model comparisons since the CIFAR-10 dataset can be easily overfitted on the backbone of ShakePyramidNet due to the large network size. On the other hand, we have used the CIFAR-10 dataset for preliminary experiments on knowledge distillation backbones.

Several data augmentation methods were used in this study, as shown in the proposed architecture Figure 2. Those methods are as follows: random horizontal flip, random crop, and cutout [13]. Random crop and horizontal flip were also used in [4]. We have also shown different knowledge distillation variations on applying cutout in Table 1.

Evaluation metric

Accuracy is a widely used metric for image classification, representing a ratio of the number of correct predictions by the total number of predictions. Since the model's performance depends on varying degradation levels, we have used the **interval mean accuracy** metric introduced in Endo et al. [2, 4], which can measure the performance for different degradation levels. The definition of the interval mean accuracy metric is as follows:

$$\overline{Acc}(\theta, Q_l, Q_u) \stackrel{def}{=} \frac{\sum_{q=Q_l}^{Q_u} Acc(f(D(\mathbf{X}, q) : \theta), \mathbf{Y})}{Q_u - Q_l + 1}, \tag{2}$$

where $\mathbf{X}$ denotes clean input images without any degradation for respective ground truth labels $\mathbf{Y}$, $\{Q_l, Q_u | Q_l < Q_u\}$ denotes the range of degradation levels, D represents the degradation operator for a degradation level q for clean image X, θ is the model parameter, and Acc represents the accuracy.

Preliminary experiments

In this paper, we explored several ways to improve the performance of a student network from a teacher network. First, we try to transfer features from varying sizes of teacher networks such as ResNet110, ResNet56, and ResNet20 to student network ResNet56. Then, we applied data augmentation techniques such as cutout to improve the generalization of student networks with different variations, as discussed in the below subsection. All preliminary experiments are performed on ResNet backbones.

Size of model backbones

As commonly shown in the research community that knowledge distillation helps to transfer information from a larger

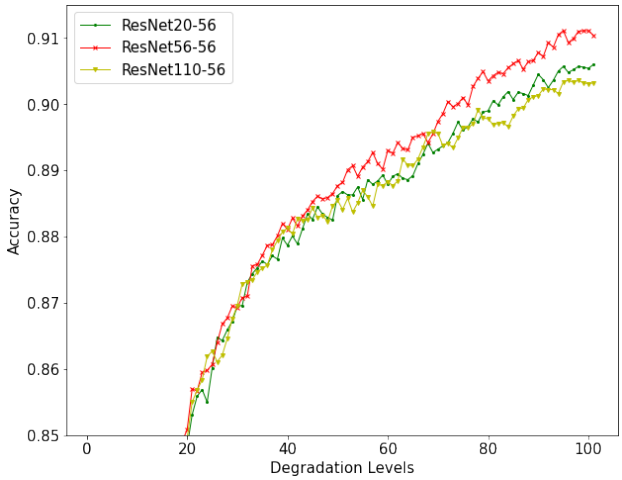


Figure 3: Accuracy for different sizes of teacher network such as ResNet20, ResNet56, and ResNet110 with student network ResNet56 on CIFAR-10 dataset, where degradation levels are JPEG quality factors.

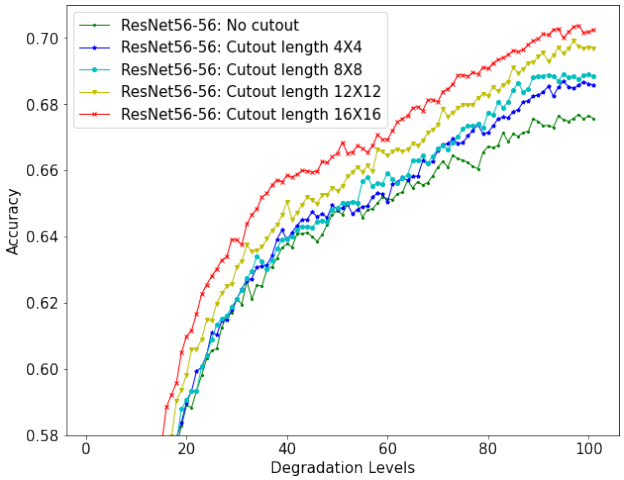


Figure 4: After applying different cutout patch lengths only on the student network, as described in Figure 2 step-2's feature extractor based on the ResNet56-56 backbone on the CIFAR-100 dataset. The X-axis represents accuracy, and Y-axis represents degradation levels for JPEG quality factors.

teacher network to a smaller student network, we inspected several variations of backbones size in our network to see if that holds in case of image classification of degradation images. We observed that a larger teacher network does not help to improve the student network performance, as shown in Figure 3. Similar research has also shown that larger teacher does not consistently improve student networks' performance in knowledge distillation [16, 17].

In our experiments, we observed that the same size network, i.e., teacher network and student network with ResNet56 backbone, performs the best among the three models as compared in Figure 3. A similar phenomenon has been reported in Stanton et al. [17] where the same size network performs the best. On the other hand, we get the worst performance with the larger backbone of the teacher network, i.e., ResNet110. Based on these facts, we have proposed ILIAC with the same backbone for both teacher and student networks, i.e., ShakePyramidNet110.

Cutout

Since cutout patch length can be dependent on the dataset, as shown in DeVries et al. [13] where optimal cutout patch length is 16X16 for CIFAR-10 and 8X8 for CIFAR-100. Similar to the cutout patch lengths examined in DeVries et al. [13], we experimented on five variations of lengths, i.e., no cutout, 4X4, 8X8, 12X12, 16X16 on JPEG degradation for the CIFAR-100 dataset, to determine the optimal cutout patch length for ILIAC. As shown in Figure 4, our method works best with a larger cutout patch length of 16X16.

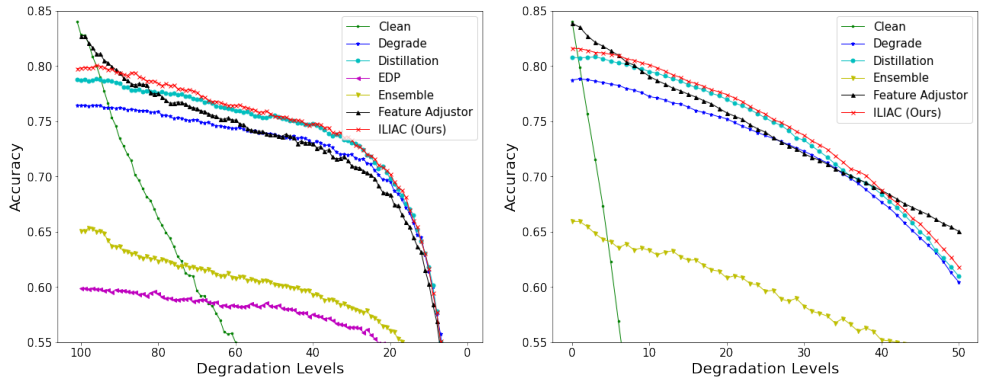
Additionally, we examined eight combinations of cutouts for applying cutout to the teacher and student network's training as shown in Table 1. Column "Pre-trained Teacher" represents whether to apply cutout during step 1 of our proposed method, i.e., training of the teacher network. Columns "Distillation - Teacher" and "Distillation - Student" represent whether to apply cutout during the distillation process on either the teacher network or the student network. For example, case 1 is our baseline when the cutout is not applied to training teacher and student networks while applying distillation. Experiments results for all eight variations of cutout are shown in Table 1. Our baseline case 1 provides the $\overline{Acc}$(All) performance of 0.864 when no cutout is applied while training either teacher or student networks. We get the best $\overline{Acc}$(All) performance of 0.882 for case 6 when we train the teacher network with cutout and apply cutout only to the student network during the distillation process. There is 0.018 mean accuracy or roughly ∼2% accuracy improvements by the usage of the cutout in the model's performance.

Table 1: Model results comparison between different approaches for cutout usage variations on CIFAR-100 dataset for JPEG compression based on the teacher and student backbones of ResNet56.

Case	Pre-trained Teacher	Distillation - Teacher	Distillation - Student	$\overline{Acc}$(All)
1				0.864
2			✓	0.874
3		✓		0.861
4		✓	✓	0.873
5	✓			0.875
6	✓		✓	**0.882**
7	✓	✓		0.873
8	✓	✓	✓	0.880

Table 2: Model comparisons between different approaches

	Existing methods						Proposed
	Clean	Degrade	Distillation	EDP [2]	Ensemble [3]	FA [4]	ILIAC
Loss functions	CE	CE	CE+CL	CE, DL	RL, DL, CE	CL+DL	CE+CL
CL Function	-	-	KLD	-	COS	COS	COS
CL location	-	-	after softmax	-	-	after avg pool	after group3
Cutout	-	-	-	-	-	-	yes
DL Function	-	-	-	MSE	MSE	MSE	-
Classifier training	train	train	train	train	fix	fix	train
Training image	clean	degrade	clean & degrade	clean & degrade	clean & degrade	clean & degrade	clean & degrade



(a) JPEG degradation

(b) AWGN degradations

Figure 5: Accuracy for the feature extractor based of ShakePyramidNet backbone on CIFAR-100 dataset with (a) JPEG and (b) AWGN where degradation levels are JPEG quality factors and standard deviations of the Gaussian distribution for the 8bit images respectively.

Model comparisons

There were 7 model comparisons in our study, as shown in Table 2. The first method, "Clean" and the second method, "Degrade" is focused on directly training the model using either clean or degraded images, respectively, using cross-entropy (CE) loss functions. The third method, "Distillation" is our baseline method where we apply knowledge distillation using the Kullback Leibler divergence (KLD) loss function after the softmax layer of the network. The fourth method, "EDP" Estimation of the Degradation parameters, is proposed by Endo et al. [2] where the Mean Squared Error (MSE) loss function is applied to estimate the degradation levels along with the cross-entropy loss function. The fifth method, "Ensemble" based on an ensemble network, is proposed by Endo et al. [3], where first restoration loss (RL) and degradation level estimator loss (DL) were used to train respective networks. Next, the classification network is trained using the cross-entropy (CE) loss function while fixing the weights of the restoration and degradation level estimator. The sixth method, "FA" Feature Adjustor, is proposed by Endo et al. [4] which is the previous state-of-the-art approach for image classification of degraded images, where degradation level estimator loss (DL) and consistency loss (CL) are applied. Lastly, our proposed method, "ILIAC", is defined in the section "Proposed Method".

Results analysis

We have evaluated ILIAC on two points in the below subsections. First, the performance quality of our model on the interval mean accuracy metric for different intervals of degradation methods such as JPEG and AWGN. Second, we compare our ILIAC model's computation efficiency to the previous state-of-the art method.

Performance quality

We have evaluated our ILIAC method in comparison with six other methods described in subsection Models Comparison. As discussed in the Introduction section, two degradation methods were inspected, JPEG and AWGN, on the CIFAR-100 dataset. As shown in Tables 3 and 4, our method outperforms the previous state-of-the-art method i.e. Feature Adjustor with the same backbone ShakePyramidNet110-110 in $\overline{Acc}$(All) i.e. the mean interval accuracy over all the degradation levels. Specifically, ILIAC's

Table 3: Interval mean accuracy for the feature extractor based on ShakePyramidNet and Endo et al. [2, 3, 4] with JPEG CIFAR-100.

Degradation Interval	Clean	Degrade	Distillation	EDP [2]	Ensemble [3]	FA [4]	ILIAC (Ours)
$\overline{Acc}(1, 20)$	0.144	0.575	0.574	0.454	0.461	0.565	**0.578**
$\overline{Acc}(21, 40)$	0.389	0.718	0.729	0.563	0.581	0.711	**0.731**
$\overline{Acc}(41, 60)$	0.512	0.738	0.754	0.581	0.603	0.739	**0.756**
$\overline{Acc}(61, 80)$	0.605	0.750	0.770	0.588	0.617	0.762	**0.775**
$\overline{Acc}(81, 100)$	0.747	0.762	0.783	0.596	0.638	**0.798**	0.794
Clean Image	**0.841**	0.765	0.788	-	-	0.836	0.798
$\overline{Acc}(1, 100)$	0.479	0.709	0.722	0.557	0.580	0.715	**0.727**
$\overline{Acc}$(All)	0.483	0.709	0.723	-	-	0.716	**0.727**

Table 4: Interval mean accuracy for the feature extractor based on ShakePyramidNet and Endo et al. [3, 4] with AWGN CIFAR-100.

Degradation Interval	Clean	Degrade	Distillation	Ensemble [3]	FA [4]	ILIAC (Ours)
Clean Image	**0.841**	0.787	0.808	0.659	0.839	0.816
$\overline{Acc}(1, 10)$	0.588	0.782	0.803	0.642	**0.812**	0.809
$\overline{Acc}(11, 20)$	0.169	0.762	0.782	0.623	0.773	**0.786**
$\overline{Acc}(21, 30)$	0.040	0.736	0.751	0.597	0.738	**0.754**
$\overline{Acc}(31, 40)$	0.020	0.700	0.705	0.567	0.703	**0.711**
$\overline{Acc}(41, 50)$	0.015	0.640	0.645	0.539	**0.667**	0.652
$\overline{Acc}$(All)	0.180	0.725	0.739	0.596	0.740	**0.744**

Table 5: Computation summary for ILIAC method and previous state-of-the-art method i.e. Feature Adjustor [5].

Model	Params	GFlops	$\overline{Acc}$(All)	
			JPEG	AWGN
ILIAC - SPN110-110	28.51 M	473.34	**0.727**	**0.744**
ILIAC - ResNet56-56	0.88 M	12.74	0.627	0.644
Feature Adjustor [4]	57.02 M	946.68	0.716	0.740

$\overline{Acc}$(All) is 0.727 for JPEG and 0.744 for AWGN. Also, refer to the Figures 5 (a) and 5 (b) for more granular view at each degradation level.

Computation efficiency

To measure our network's computational efficiency, we have used two metrics that are widely used to measure the efficiency of neural networks: model parameters (in millions) and flops. These results are shared in Table 5 along with mean interval accuracy for all degradation levels on JPEG and AWGN. ILIAC method with ShakePyramidNet (SPN) backbone is about half the size of the previous state-of-the-art feature adjustor method [5]. Feature Adjustor method flops, and model parameters are estimated based on the author's comments in the paper that the model is twice the size of the ShakePyramidNet (SPN110-110) backbone since it requires two feature extractors and several estimator modules. Additionally, we have included our ILIAC approach with the ResNet56-56 backbone, which is much smaller than the ShakePyramidNet backbone, along with satisfactory performance results.

Conclusion and Future Work

Overall, our ILIAC method can outperform the previous state-of-the-art methods on degradation methods such as JPEG and AWGN on the CIFAR-100 dataset. Moreover, our ILIAC method is efficient in model parameters and GFlops; specifically, it requires roughly half the computation compared to the previous state-of-the-art method. Additionally, we demonstrate through our experiments that we do not require larger networks to generalize well on image classification of the degraded images. Future studies could examine three relevant research directions as follows: (1) investigation of several data augmentation methods that can work best with image degradations, (2) designing efficient neural network architectures using NAS or similar methods for image classification of degraded images, and (3) affect of image degradation on other computer vision tasks such as object detection and semantic segmentation.

References

[1] G. Hinton, O. Vinyals and J. Dean, Distilling the knowledge in a neural network, Proc. NeurIPS (2015).

[2] K. Endo, M. Tanaka, and M. Okutomi, CNN-based classification of degraded images, Proc. Int. Symp. Electron. Imag. (IS&T), pp. 28-1-28-6 (2020).

[3] K. Endo, M. Tanaka and M. Okutomi, CNN-Based Classification of Degraded Images With Awareness of Degradation Levels, Proc. IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 4046-4057 (2021).

[4] K. Endo, M. Tanaka and M. Okutomi, CNN-Based Classification of Degraded Images Without Sacrificing Clean Images, Proc. IEEE Access, vol. 9, pp. 116094-116104 (2021).

[5] K. Endo, M. Tanaka, and M. Okutomi, Classifying degraded images over various levels of degradation, Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 1691-1695 (2020).

[6] A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report of Univ of Toronto (2009).

[7] M. Suin, K. Purohit and A. N. Rajagopalan, Degradation Aware Approach to Image Restoration Using Knowledge Distillation, IEEE Journal of Selected Topics in Signal Processing, vol. 15, no. 2, pp. 162-173 (2021).

[8] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 770-778 (2016).

[9] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, Proc. Int. Conf. on Learning Representations (ICLR) (2015).

[10] D. Han, J. Kim and J. Kim, Deep Pyramidal Residual Networks, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 6307-6315 (2017).

[11] Y. Yamada, M. Iwamura, T. Akiba and K. Kise, Shakedrop Regularization for Deep Residual Learning, Proc. IEEE Access, vol. 7, pp. 186126-186136, (2019).

[12] W. Roth, G. Schindler, M. Zohrer, L. Pfeifenberger, R. Peharz, S. Tschiatschek, H. Froning, F. Pernkopf, and Z. Ghahramani, Resource-Efficient Neural Networks for Embedded Systems, ArXiv, abs/2001.03048 (2020).

[13] T. DeVries, G. Taylor, Improved Regularization of Convolutional Neural Networks with Cutout, ArXiv, abs/1708.04552 (2017).

[14] B. R. Mohapatra, A. Mishra and S. K. Rout, A Comprehensive Review on Image Restoration Techniques, Int. Journal of Research in Advent Technology, vol. 2, no.3 (2014).

[15] J. Gou, B. Yu, S. J. Maybank and D. Tao, Knowledge Distillation: A Survey, Int. Journal of Computer Vision 129, 1789–1819 (2021).

[16] J. H. Cho and B. Hariharan, On the Efficacy of Knowledge Distillation, Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 4793-4801 (2019).

[17] S. Stanton, P. Izmailov, P. Kirichenko, A. A. Alemi, and A. G. Wilson, Does Knowledge Distillation Really Work?, Proc. NeurIPS (2021).

Author Biography

Dinesh Daultani *received his B.Eng. degree in computer science from RGPV, Bhopal, India, and his MS in Information Systems from Illinois State University, the USA, in 2010 and 2014, respectively. He is currently a Ph.D. student since 2021 in the Department of Systems and Control Engineering, Tokyo Institute of Technology. Additionally, he worked as a Research Scientist at Rakuten from 2018 to 2021 and as a Data Scientist at Woven Planet from 2021 to 2022.*

Masayuki Tanaka *received the bachelor's and master's degrees in control engineering and the Ph.D. degree from Tokyo Institute of Technology, in 1998, 2000, and 2003, respectively. He joined Agilent Technology, in 2003. He was a Research Scientist at Tokyo Institute of Technology, from 2004 to 2008. He was a Visiting Scholar with the Department of Psychology, Stanford University, CA, USA. Since 2008, he has been an Associate Professor at the Graduate School of Science and Engineering, Tokyo Institute of Technology.*

Masatoshi Okutomi *received the B.E. degree from The University of Tokyo in 1981 and the M.E. degree from Tokyo Institute of Technology in 1983. He joined Canon Research Center in 1983. From 1987 to 1990, he was a Visiting Research Scientist with the School of Computer Science, Carnegie Mellon University. He received the PhD degree by dissertation from Tokyo Institute of Technology in 1993. Since 1994, he has been with Tokyo Institute of Technology, where he is currently the Professor with the Department of Systems and Control Engineering.*

Kazuki Endo *received a bachelor's degree in mathematics, a master's degree in industrial engineering and management, and a D.Eng. degree in systems and control engineering from Tokyo Institute of Technology, Tokyo, Japan, in 1997, 1999, and 2022, respectively. He joined Industrial Bank of Japan, Ltd., Tokyo, in 1999. Since 2022, he has been an Associate Professor with the Department of Business, Teikyo Heisei University, Tokyo, Japan.*