Supplementary Material of ”Self-Feature Distillation with Uncertainty Modeling for Degraded Image Recognition”

Zhou Yang[1], Weisheng Dong[1(✉)], Xin Li[2], Jinjian Wu[1], Leida Li[1], and Guangming Shi[1]

[1] School of Artificial Intelligence, Xidian University, Xi’an, China
yang_zhou@stu.xidian.edu.cn, {wsdong, jinjian.wu}@mail.xidian.edu.cn
{ldli, gmshi}@xidian.edu.cn
[2] Lane Dep. of CSEE, West Virginia University, Morgantown WV, USA
xin.li@mail.wvu.edu

1 Performance on Individual Corruption Types

In this section, we show the performance of our method on individual corruption types. We compared our method with the vanilla model (trained on clean data only) and QualNet[4] (trained on clean and corrupted images) in two types of backbone networks, i.e., ResNet50[2] and ResNeXt101-32x8d[5]. We trained our model with ImageNet-1K[1] and employed 15 corruption types described in ImageNet-C[3] to generate corrupted images. As the results are shown in Table 1, our method is superior to QualNet in most degradation types. It is worth noting that our method has less performance degradation on clean images and is more robust than QualNet, from which we can verify the superiority of our method.

Table 1. The top-1 accuracy on clean images in ImageNet-1K[1] and individual corruption in ImageNet-C[3]. ”*Corrupt.Average*” indicates the average accuracy over 15 corruption types. The other specific types show the individual corruption average accuracy (%) over the five severity levels. The best results are indicated in bold.

method	Architecture	Top-1 Accuracy (%)																
		clean	Corrupt.	Noise			Blur				Weather				Digital			
			Average	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG
Vanilla	ResNet50	**76.82%**	39.17%	29.29%	27.03%	23.81%	38.75%	26.78%	38.67%	36.24%	32.53%	38.14%	45.84%	68.02%	39.06%	45.25%	44.79%	53.41%
QualNet50		75.43%	60.90%	59.83%	60.28%	59.79%	56.20%	57.38%	60.81%	59.83%	59.29%	59.45%	61.89%	68.22%	58.74%	60.94%	66.55%	64.36%
Ours		76.23%	**63.44%**	**62.36%**	**62.67%**	**61.93%**	**59.44%**	**57.94%**	**62.75%**	**61.37%**	**59.91%**	**60.53%**	**64.67%**	**72.98%**	**61.14%**	**63.95%**	**70.37%**	**69.60%**
Vanilla	ResNeXt101	**79.68%**	44.64%	39.12%	36.61%	34.80%	42.87%	28.29%	44.59%	42.17%	36.49%	41.82%	48.87%	69.66%	41.97%	48.84%	53.98%	59.68%
QualNeXt101		77.81%	66.62%	61.18%	60.06%	60.92%	60.21%	58.61%	**69.65%**	**69.93%**	**68.38%**	66.63%	69.62%	**75.71%**	67.41%	**70.23%**	72.62%	68.18%
Ours		79.04%	**69.16%**	**67.42%**	**67.70%**	**67.29%**	**63.92%**	**63.60%**	69.07%	69.76%	67.40%	**67.69%**	**71.46%**	75.61%	**69.86%**	69.24%	**74.82%**	**72.64%**

2 Visual comparison results on the reconstructed final feature map

In this section, we made a visual observation of the final feature map (size:7×7) extracted both from high-quality (HQ) images and low-quality (LQ) images

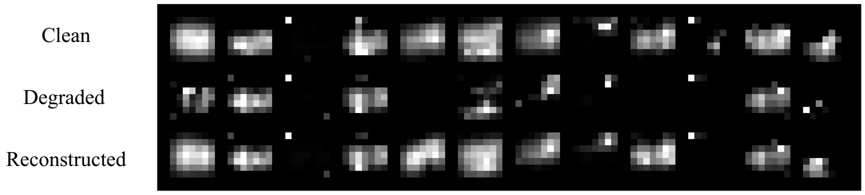

Fig. 1. Visual observation of the final feature map.

by our method, respectively. We selected 12 of the 2048 channels in the final feature map and normalized it to display. As shown in Fig. 1, The first row is the feature map extracted from the clean image, the second row is the degraded feature extracted from the corresponding degraded image, and the third row is the feature reconstructed by our method. We can observe that almost all feature maps extracted from low-quality images by our method are reconstructed well, which is close to those extracted from clean images. This enables our method to improve the recognition accuracy of the model on degraded images

3 Additional ablation studies:

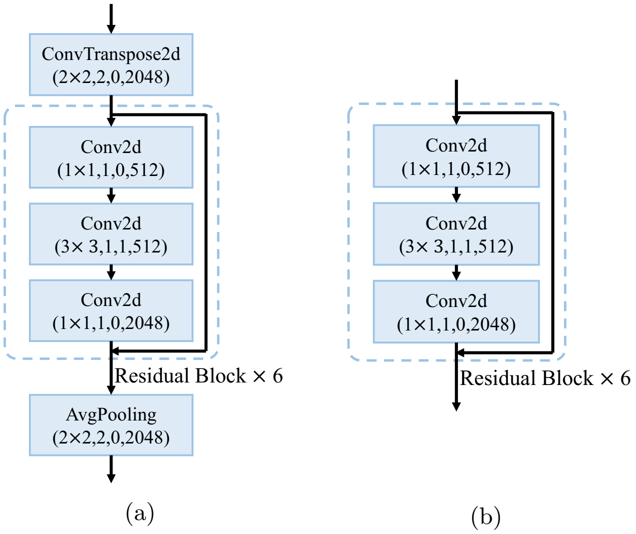

Fig. 2. (a) Uncertainty estimation module with the ConvTranspose2D layer. In order to maintain the consistency of spatial dimensions, we also add an AveragePooling layer with kernel size of 2 and stride 2. (b) Uncertainty estimation module without ConvTranspose2D layer and the corresponding AveragePooling layer.

Table 2. Additional ablation study for the ConvTranspose2d layer in our uncertainty estimation module.

ConvTranspose2D	×	✓
Top-1 Accuracy	62.77%	63.44%

1) ***Effect on the ConvTranspose2d layer in our UEM:***

As the uncertainty estimation module (UEM) described in the paper, we first use a *ConvTranspose2d* layer to expand the spatial dimension of the feature map. We have trained this module with and without this layer in the same number of epochs to investigate the effect of this layer. Fig. 2(a) and Fig. 2(b) show the two different architectures of our uncertainty estimation module (UEM). Table 2 shows the top-1 accuracy of the two structures in the degraded images, from which we can learn that adding this layer can slightly improve the recognition accuracy. The reason may be that by expanding the spatial dimension, it can enrich the information of the feature map and make it easy to estimate uncertainty.

2) ***Training cost of our uncertainty estimation module:***

In our experiments, the batch size was set to 512 and 4 NVIDIA GeForce RTX 3090 GPUs were used. With the proposed uncertainty estimation module (UEM), we got an average training speed of **1.37s/batch** and **2.5 GPU days** for total training. While the training speed was **1.14s/batch** and **2.2 GPU days** for total training without the UEM. We believe that the additional training cost brought by this module is acceptable. During testing, we remove the uncertainty estimation module, so there will be no additional calculation cost.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 1
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
3. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019) 1
4. Kim, I., Han, S., Baek, J.w., Park, S.J., Han, J.J., Shin, J.: Quality-agnostic image recognition via invertible decoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12257–12266 (2021) 1
5. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017) 1