Supplementary Material: Boosting the Generalization Capability in Cross-Domain Few-Shot Learning via Noise-enhanced Supervised Autoencoder

1. Discriminability Analysis of Deep Features

Below we give the details of the definition of the Inter-class correlation (ICC) [1, 2]. Let f be an feature extractor and $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_K$ where $\mathcal{D}_j = \{(x_i, y_i) : y_i = j\}$ be a dataset with K classes. Let $\tilde{f}(x_i) := \frac{f(x_i)}{\|f(x_i)\|_2}$ be the normalized feature extracted by the feature extractor f. Then the center of the images features in jth class is defined as

$$\mu(f|\mathcal{D}_j) = |\mathcal{D}_j|^{-1} \sum_{x_i \in \mathcal{D}_j} \tilde{f}(x_i). \tag{1}$$

Then the classical intra-class and inter-class variation on the full dataset $\mathcal{D}$ are defined respectively as

$$\begin{aligned} D_{\text{intra}}(f|\mathcal{D}) &= \frac{1}{K} \sum_{k=1}^{K} \left\{ |\mathcal{D}_k|^{-1} \sum_{x_i \in \mathcal{D}_k} \|\tilde{f}(x_i) - \mu(\mathcal{D}_k)\|^2 \right\}, \\ D_{\text{inter}}(f|\mathcal{D}) &= \frac{1}{K(K-1)} \sum_{k=1}^{K} \sum_{j \neq k} \|\mu(\mathcal{D}_j) - \mu(\mathcal{D}_k)\|^2. \end{aligned} \tag{2}$$

The inter-class variation measures the average pairwise distances of class centers and the intra-class variation measures the within class variation of the image features. Following [2], the intra-class correlation (ICC) is defined as

$$\text{ICC}(f|\mathcal{D}) = D_{\text{inter}}(f|\mathcal{D}) / D_{\text{intra}}(f|\mathcal{D}). \tag{3}$$

Therefore, the ICC of a feature extractor f on dataset $\mathcal{D}$ is larger when the inter-class is larger and the intra-class is smaller. The ICC can therefore measures the discriminability of a feature extractor since a good feature embedding has smaller within-class variation and larger margin across classes.

In our experiment to study the discriminability of the feature extractors, we randomly sample 5 classes and compute the ICC based on the images from these 5 classes. We repeat these procedure for 600 times and use the average ICC as a measure for the discriminability of a feature extractor. The average ICC is computed using the same feature extractor on both the source and the target domains.

2. Model Architecture

In our proposed noise-enhanced supervised autoencoder, we use Conv4 and ResNet10 as the encoder structure and design the corresponding decoders. The decoder can be seen as a mirror mapping of the encoder which consist of deconvolutional blocks, with each block containing 2D transposed convolution operator and ReLU activation, which expand the dimension of the feature map. Before deconvolutional layers, we also add several fully connected layers that transform feature representations from encoder. Detailed architecture and layer specifications of the decoders are shown in Table 1.

Table 1. The architecture and layer specifications of the decoder modules of the Conv4 and ResNet10 based NSAE. Linear represents fully connected layer followed by ReLU activation. Deconv-ReLU represents a ConvTranspose2d-BatchNormalization-ReLU layer. Conv-Sigmoid represents a Conv2d-BatchNormalization-Sigmoid layer.

Module	Specifications
Conv4	Linear, 1600×512
	Linear, 512×1600
	Reshape to 64×5×5
	2×2 Deconv-ReLU, 64 filters, stride 2, padding 0
	2×2 Deconv-ReLU, 64 filters, stride 2, padding 0
	2×2 Deconv-ReLU, 64 filters, stride 2, padding 0
	2×2 Deconv-ReLU, 3 filters, stride 2, padding 0
	3×3 Conv-Sigmoid, 3 filters, stride 1, padding 1
ResNet10	Linear, 512×512
	Linear, 512×6272
	Reshape to 32×14×14
	2× 2 Deconv-ReLU, 32 filters, stride 2, padding 0
	2× 2 Deconv-ReLU, 32 filters, stride 2, padding 0
	2× 2 Deconv-ReLU, 64 filters, stride 2, padding 0
	2× 2 Deconv-ReLU, 64 filters, stride 2, padding 0
	3×3 Conv-Sigmoid, 3 filters, stride 1, padding 1

3. Ablation Study Results

Table 2 gives the detailed experiment result for the 5-way 5-shot ablation study on 8 datasets with various model architectures and loss functions. We use four kinds of combinations of the classification loss functions for pre-training and fine-tuning, i.e. CE+CE, BSR+CE, CE+D, and BSR+D. Meanwhile, we respectively test with Conv4

Table 2. **Ablation study**. The ablation study on the 5-way 5-shot support set on 8 datasets with various model architectures and loss functions.

Encoder	Method	ISIC	EuroSAT	CropDisease	ChestX	Car	CUB	Plantae	Places
Conv4	CE+CE	46.04±0.62	68.88±0.68	83.47±0.67	24.81±0.43	38.36±0.58	52.94±0.70	45.55±0.71	58.74±0.74
	NSAE(-)	46.64±0.62	70.19±0.64	84.89±0.60	24.86±0.41	39.88±0.64	55.08±0.70	46.98±0.69	59.54±0.69
	NSAE	**46.65±0.63**	**70.34±0.65**	**85.22±0.61**	**25.02±0.42**	**39.90±0.62**	**55.35±0.67**	**47.18±0.75**	**59.59±0.69**
	SAE	46.36±0.61	68.88±0.65	83.57±0.62	24.88±0.41	37.94±0.59	52.74±0.65	44.79±0.67	58.34±0.71
	SAE(*)	44.88±0.60	70.10±0.68	83.38±0.64	25.00±0.41	37.29±0.60	53.64±0.70	44.60±0.65	59.12±0.72
	BSR+CE	48.78±0.64	69.34±0.68	85.88±0.61	25.41±0.42	42.54±0.70	60.16±0.73	**50.85±0.78**	62.38±0.77
	NSAE(-)	49.32±0.59	71.84±0.66	86.86±0.59	25.59±0.42	42.54±0.64	60.00±0.71	50.00±0.73	63.36±0.72
	NSAE	**49.34±0.59**	72.00±0.65	**86.87±0.59**	**25.62±0.42**	**42.56±0.65**	**60.18±0.72**	49.48±0.73	**63.40±0.72**
	SAE	48.68±0.63	**72.17±0.69**	86.23±0.60	25.31±0.41	42.38±0.68	60.10±0.76	48.29±0.74	62.12±0.73
	SAE(*)	47.25±0.58	70.52±0.67	85.14±0.62	25.45±0.40	40.81±0.63	59.54±0.74	47.70±0.70	62.46±0.75
	CE+D	50.54±0.66	76.16±0.64	89.65±0.55	24.07±0.41	44.26±0.70	58.61±0.82	52.47±0.74	61.81±0.74
	NSAE(-)	50.94±0.63	77.70±0.70	90.06±0.52	**24.46±0.40**	43.96±0.68	**60.00±0.82**	53.26±0.80	62.26±0.73
	NSAE	**50.95±0.63**	**77.77±0.64**	**90.11±0.52**	24.29±0.40	**44.28±0.72**	59.90±0.78	**53.36±0.78**	**62.42±0.72**
	SAE	50.62±0.65	74.92±0.64	88.11±0.59	23.52±0.41	42.45±0.70	57.04±0.81	51.51±0.80	61.08±0.74
	SAE(*)	49.94±0.66	77.24±0.70	88.31±0.56	24.01±0.40	42.12±0.68	57.15±0.82	51.77±0.82	61.40±0.78
	BSR+D	**50.06±0.65**	75.74±0.67	87.71±0.56	**23.66±0.40**	41.11±0.77	58.81±0.81	51.35±0.81	60.51±0.80
	NSAE(-)	49.95±0.67	76.96±0.70	87.70±0.57	23.60±0.41	41.05±0.72	58.32±0.81	51.74±0.84	60.34±0.81
	NSAE	49.98±0.67	**77.00±0.69**	**87.71±0.58**	23.61±0.41	**41.80±0.72**	**59.42±0.82**	**51.80±0.84**	**60.92±0.85**
	SAE	49.77±0.68	75.58±0.69	87.67±0.57	23.35±0.41	41.75±0.75	58.34±0.81	50.92±0.81	60.25±0.81
	SAE(*)	49.46±0.68	76.17±0.70	86.50±0.60	23.23±0.39	40.16±0.70	58.26±0.79	50.70±0.83	60.86±0.84
ResNet10	CE+CE	51.28±0.62	82.51±0.58	92.45±0.45	26.50±0.43	52.08±0.72	64.14±0.77	59.27±0.70	70.06±0.74
	NSAE(-)	53.52±0.62	83.83±0.56	93.14±0.47	26.69±0.44	53.49±0.72	67.60±0.73	59.70±0.74	70.74±0.71
	NSAE	**54.05±0.63**	**83.96±0.57**	**93.14±0.47**	**27.10±0.44**	**54.91±0.74**	**68.51±0.76**	**59.80±0.74**	**71.84±0.72**
	SAE	52.28±0.63	83.78±0.55	93.01±0.42	26.05±0.45	53.54±0.71	64.27±0.75	59.87±0.73	70.82±0.72
	SAE(*)	52.11±0.65	83.50±0.55	93.05±0.47	26.37±0.45	54.26±0.70	66.62±0.75	59.62±0.75	71.40±0.67
	BSR+CE	54.42±0.66	80.89±0.61	92.17±0.45	26.84±0.44	57.49±0.72	69.38±0.76	61.07±0.76	71.09±0.68
	NSAE(-)	55.27±0.62	84.19±0.54	92.92±0.47	27.23±0.45	**58.35±0.76**	71.30±0.75	61.92±0.76	71.76±0.74
	NSAE	**55.88±0.64**	**84.33±0.55**	**93.31±0.42**	**27.30±0.42**	58.30±0.75	**71.92±0.77**	**62.18±0.77**	**73.17±0.72**
	SAE	54.48±0.65	84.10±0.54	92.92±0.47	27.20±0.45	58.30±0.76	71.30±0.75	61.92±0.76	71.76±0.74
	SAE(*)	54.73±0.68	83.90±0.55	93.02±0.46	26.74±0.43	57.60±0.71	71.50±0.75	62.20±0.78	72.99±0.67
	CE+D	51.62±0.66	83.72±0.59	93.22±0.41	26.23±0.44	55.12±0.76	66.56±0.78	59.09±0.76	72.81±0.73
	NSAE(-)	54.31±0.68	83.77±0.62	93.54±0.40	26.98±0.44	55.67±0.78	67.17±0.76	59.46±0.75	72.90±0.72
	NSAE	**54.41±0.63**	**83.78±0.56**	**93.65±0.40**	**27.25±0.44**	**55.78±0.73**	**67.64±0.76**	**59.74±0.75**	**73.25±0.73**
	SAE	52.64±0.67	83.13±0.63	93.44±0.41	26.34±0.44	55.44±0.74	65.08±0.76	59.70±0.78	73.13±0.71
	SAE(*)	51.37±0.66	83.04±0.63	92.53±0.42	26.44±0.42	55.00±0.73	65.13±0.81	59.46±0.79	73.20±0.67
	BSR+D	52.85±0.65	80.13±0.65	91.20±0.48	**26.80±0.45**	54.99±0.74	68.15±0.84	58.26±0.77	71.97±0.72
	NSAE(-)	53.74±0.67	82.19±0.64	92.22±0.47	26.79±0.45	55.90±0.77	68.32±0.81	60.25±0.77	73.28±0.72
	NSAE	**54.42±0.64**	**82.79±0.62**	**92.45±0.45**	26.69±0.45	**55.92±0.72**	**68.46±0.82**	**60.40±0.78**	**73.33±0.71**
	SAE	51.84±0.65	80.02±0.69	91.95±0.45	26.52±0.42	55.90±0.77	66.64±0.79	59.20±0.80	72.48±0.76
	SAE(*)	53.08±0.67	81.77±0.64	91.63±0.46	26.58±0.45	54.87±0.78	67.97±0.83	58.61±0.79	73.20±0.67

and ResNet10 as backbone of feature encoder. In the table, CE+CE, BSR+CE, CE+D, and BSR+D denote using single feature extractor with different loss functions combinations. SAE denotes that we use auto-encoder but do not further feed in the reconstructed images for classification during the pre-training. SAE(*) denotes that we double the weight on the classification loss of original images as if the auto-encoder works perfectly that the reconstructed images are identical to original images. NSAE(-) denotes using our proposed pre-training strategy but using one-step fine-tuning.

4. Comparison with Handcrafted Noise

The reconstructed images during the pre-training stage can be viewed as noisy inputs to improve the model generalization capability. Can the model generalization capability be improved if we use images with handcrafted noise instead of reconstructed images? To answer this question, we compare the performance of our proposed method with that when images with handcrafted noise are used as data augmentation during pre-training. In our experiment, we consider the following four kinds of handcrafted noise: Gaussian, salt-pepper, Poisson, and speckle. We use the *skimage* package [3] in python to add handcrafted noise to source images. The parameter values for the noise generation are given in Table 3. We use BSR+CE loss combination and consider the following two settings during pre-training: (a) only use the encoder and feed in both source and hand-

Table 3. **Handcrafted Noise Configuration**. The parameters for adding noise to the images.

Noise type	Parameter values
Gaussian	mode='gaussian', mean=0, var=0.1
salt-pepper	mode='s&p', salt_vs_pepper=0.5
Poisson	mode='poisson'
speckle	mode='speckle', mean=0, var=0.05

crafted noisy images for classification; (b) add a decoder to (a) with reconstruction loss, though the reconstructed images are not used for classification. The rest of the hyper-parameter values are the same as that given in Section 4.1 in the main paper. The results averaged over 8 datasets are shown in Fig. 1. It can be seen from the figures that

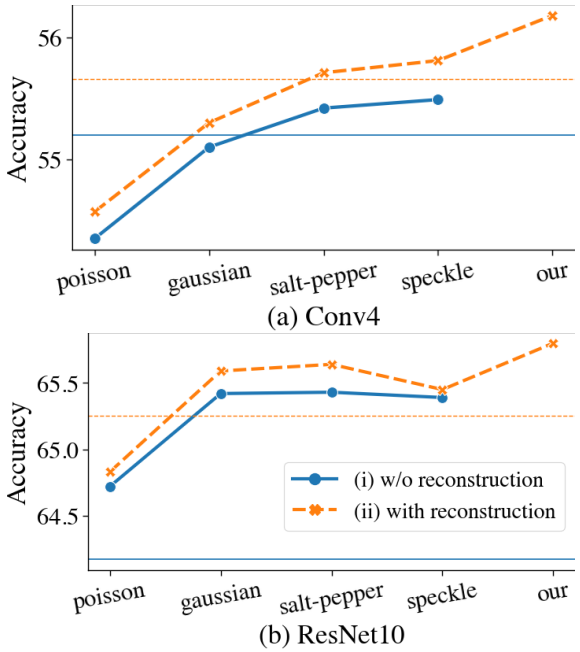

Figure 1. Ablation study with handcrafted noisy images. The two horizontal lines are baselines without noisy images.

1. regardless of the noise type, using auto-encoder scheme with reconstruction loss helps improve the generalization capability owing to regularization effect from decoder, which shows the advantage of our model on top of simple data augmentation;

2. adding handcrafted noise may not improve the accuracy, but our design consistently improves the accuracy and surpasses all results with handcrafted noise.

References

[1] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *European Conference on Computer Vision*, pages 438–455. Springer, 2020. 1

[2] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48. Ieee, 1999. 1

[3] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. 2