Learning from Extrinsic and Intrinsic Supervisions for Domain Generalization

Shujun Wang[1][0000−0003−1495−3278], Lequan Yu[2⋆][0000−0002−9315−6527],
Caizi Li[3], Chi-Wing Fu[1,3][0000−0002−5238−593X], and
Pheng-Ann Heng[1,3][0000−0003−3055−5034]

[1] The Chinese University of Hong Kong
{sjwang,cwfu,pheng}@cse.cuhk.edu.hk
[2] Stanford University
lequany@stanford.edu
[3] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology,
Chinese Academy of Sciences, Shenzhen, China
cz.li@siat.ac.cn

Abstract. The generalization capability of neural networks across domains is crucial for real-world applications. We argue that a generalized object recognition system should well understand the relationships among different images and also the images themselves at the same time. To this end, we present a new domain generalization framework (called EISNet) that learns how to generalize across domains *simultaneously* from *extrinsic relationship supervision* and *intrinsic self-supervision* for images from multi-source domains. To be specific, we formulate our framework with feature embedding using a multi-task learning paradigm. Besides conducting the common supervised recognition task, we seamlessly integrate a momentum metric learning task and a self-supervised auxiliary task to collectively integrate the extrinsic and intrinsic supervisions. Also, we develop an effective momentum metric learning scheme with the K-hard negative mining to boost the network generalization ability. We demonstrate the effectiveness of our approach on two standard object recognition benchmarks VLCS and PACS, and show that our EISNet achieves state-of-the-art performance.

Keywords: Domain generalization, unsupervised learning, metric learning, self-supervision

1 Introduction

The rise of deep neural networks has achieved promising results in various computer vision tasks. Most of these achievements are based on supervised learning, which assumes that the models are trained and tested on the samples drawn from the same distribution or domain. However, in many real-world scenarios,

[⋆] Corresponding Author

the training and test samples are often acquired under different criteria. Therefore, the trained network may perform poorly on "unseen" test data with domain discrepancy from the training data. To address this limitation, researchers have studied how to alleviate the performance degradation of a trained network among different domains. For instance, by utilizing labeled (or unlabeled) target domain samples, various domain adaptation methods have been proposed to minimize the domain discrepancy by aligning the source and target domain distributions [11, 18, 23, 35, 41, 42].

Although these domain adaptation methods can achieve better performance on the target domain, there exists an indispensable demand to pre-collect and access target domain data during the network training. Moreover, it needs to re-train the network to adapt to every new target domain. However, in real-world applications, it is often the case that adequate target domain data is not available during the training process [28, 49]. For example, it is difficult for an automated driving system to know which domain (*e.g.*, city, weather) the self-driving car will be used. Therefore, it has a broad interest in studying how to learn a generalizable network that can be directly applied to new "unseen" target domains. Recently, the community develops *domain generalization* methods to improve the model generalization ability on unseen target domains by utilizing the multiple source domains.

Most existing domain generalization methods attempt to extract the shared domain-invariant semantic features among multiple source domains [8,26–28,31]. For example, Li *et al.* [28] extend an adversarial auto-encoder by imposing the Maximum Mean Discrepancy (MMD) measure to align the distributions among different domains. Since there is no specific prior information from target domains during the training, some works have investigated the effectiveness of increasing the diversity of the inputs by creating synthetic samples to improve the generalization ability of networks [49, 50]. For instance, Yue *et al.* [49] propose a domain randomization method with Generative Adversarial Networks (GANs) to learn a model with high generalizability. Meta-learning has also been introduced to address the domain generalization problem via an episodic training [8, 27]. Very recently, Carlucci *et al.* [2] introduce a self-supervision task by predicting relative positions of image patches to constrain the semantic feature learning for domain generalization. This shows that the self-supervised task can discover invariance in images with different patch orders and thus improve the network generalization. Such self-supervision task only considers the regularization within images but does not explore the valuable relationship among images across different domains to further enhance the discriminability and transferability of semantic features.

The generalization of deep neural networks relies crucially on the ability to learn and adapt knowledge across various domains. We argue that a generalized object recognition system should well understand the relationships among different objects and the objects themselves at the same time. Particularly, on the one hand, exploring the relationship among different objects (*i.e.*, *extrinsic supervision*) guides the network to extract domain-independent yet category-specific

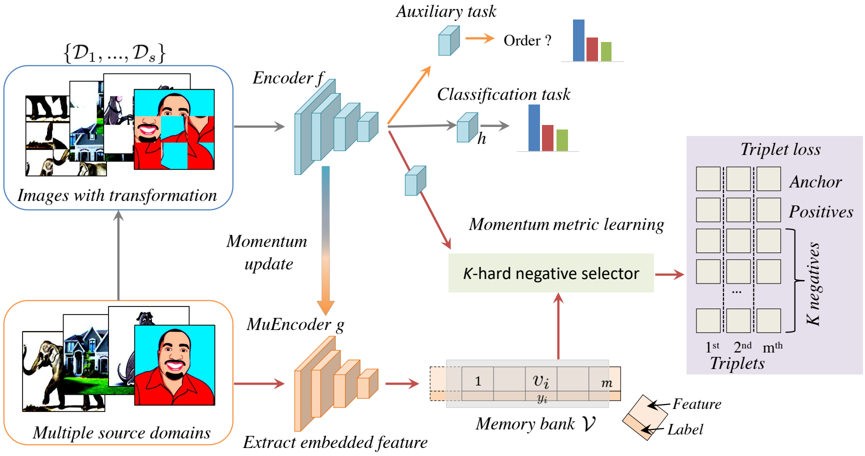


Fig. 1. The framework of the proposed EISNet for domain generalization. We train a feature Encoder f for discriminative and transferable feature extraction and a classifier for object recognition. Two complementary tasks, a momentum metric learning task and a self-supervised auxiliary task, are introduced to prompt general feature learning. We maintain a momentum updated Encoder (MuEncoder) to generate momentum updated embeddings stored in a large memory bank. Also, we design a K-hard negative selector to locate the informative hard triplets from the memory bank to calculate the triplet loss. The auxiliary self-supervised task predicts the order of patches within an image.

representation, facilitating decision-boundary learning. On the other hand, exploring context or shape constraint within a single image (*i.e.*, *intrinsic supervision*) introduces necessary regularization for network training, broadening the network understanding of the object.

To this end, we present a new framework called EISNet that learns how to generalize across domains by simultaneously incorporating *extrinsic supervision* and *intrinsic supervision* for images from multi-source domains. We formulate our framework as a multi-task learning paradigm for general feature learning, as shown in Fig. 1. Besides conducting the common supervised recognition task, we seamlessly integrate a momentum metric learning task and a self-supervised auxiliary task into our framework to utilize the *extrinsic* and *intrinsic* supervisions, respectively. Specifically, we develop an effective momentum metric learning scheme with the K-hard negative selector to encourage the network to explore the image relationship and enhance the discriminability learning. The K-hard negative selector is able to filter the informative hard triplets, while the momentum updated encoder guarantees the consistency of embedded features stored in the memory bank, which stabilizes the training process. We then introduce a jigsaw puzzle solving task to learn the spatial relationship of images parts. The three kinds of tasks share the same feature encoder and are optimized in an end-to-end manner. We demonstrate the effectiveness of our approach on two object recognition benchmarks. Our EISNet achieves the state-of-the-art performance.

2 Related Work

Domain adaptation and generalization The goal of unsupervised domain adaptation is to learn a general model with source domain images and unlabeled target domain images, so that the model could perform well on the target domain. Under such a problem setting, images from the target domain can be utilized to guide the optimization procedure. The general idea of domain adaption is to align the source domain and target domain distributions in the input level [3, 19], semantic feature level [9, 34], or output space level [5, 39, 40, 43, 45]. Most methods adopt Generative Adversarial Networks and achieve better performance on the target domain data. However, training domain adaptation models need to access unlabeled target domain data, making it impractical for some real-world applications.

Domain generalization is an active research area in recent years. Its goal is to train a neural network on multiple source domains and produce a trained model that can be applied directly to unseen target domain. Since there is no specific prior guidance from the target domain during the training procedure, some domain generalization methods proposed to generate synthetic images derived from the given multiple source domains to increase the diversity of the input images, so that the network could learn from a larger data space [49, 50]. Another promising direction is to extract domain-invariant features over multiple source domains [12, 26–28, 31]. For example, Li *et al.* [25] developed a low-rank parameterized CNN model for domain generalization and proposed the domain generalization benchmark dataset PACS. Motiian *et al.* [31] presented a unified framework by exploiting the Siamese architecture to learn a discriminative space. A novel framework based on adversarial autoencoders was presented by Li *et al.* [28] to learn a generalized latent feature representation across domains. Recently, meta-learning-based episodic training was designed to tackle domain generalization problems [8, 27]. Li *et al.* [27] developed an episodic training procedure to expose the network to domain shift that characterizes a novel domain at runtime to improve the robustness of the network. Our work is most related to [2], which introduced self-supervision signals to regularize the semantic feature learning. However, besides the self-supervision signals within a single image, we further exploit the extrinsic relationship among image samples across different domains to improve the feature compactness.

Metric learning Our work is also related to metric learning, which aims to learn a metric to minimize the intra-class distances and maximize the inter-class variations [14, 46]. With the development of deep learning, distance metric also benefits the feature embedding learning for better discrimination [17, 44]. Recently, the metric learning strategies have attracted a lot of attention on face verification and recognition [36], fine-grained object recognition [44], image retrieval [48], and so on. Different from previous applications, in this work, we adopt the conventional triplet loss with more informative negative selection and momentum feature extraction for domain generalization.

Self-supervision Self-supervision is a recent paradigm for unsupervised learning. The idea is to design annotation-free (*i.e.*, self-supervised) tasks for feature learning to facilitate the main task learning. Annotation-free tasks can be predictions of the image colors [24], relative locations of patches from the same image [2, 32], image inpainting [33], and image rotation [13]. Typically, self-supervised tasks are used as network pre-train to learn general image features. Recently, it is trained as an auxiliary task to promote the mainstream task by sharing semantic features [4]. In this paper, we inherit the advantage of self-supervision to boost the network generalization ability.

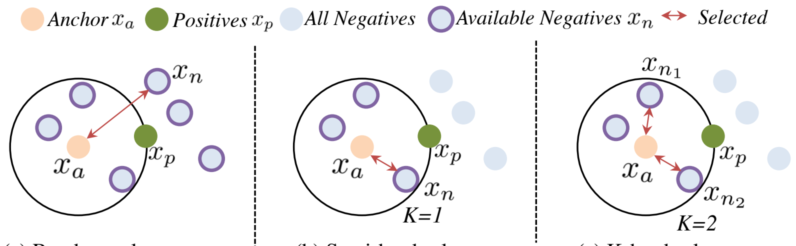
3 Method

We aim to learn a model that can perform well on "unseen" target domain by utilizing multiple source domains. Formally, we consider a set of S source domains $\{\mathcal{D}_1, ..., \mathcal{D}_s\}$, with the j-th domain $\mathcal{D}_j$ having N_j sample-label pairs $\{(x_i^j, y_i^j)\}_{i=1}^{N_j}$, where x_i^j is the i-th sample in $\mathcal{D}_j$ and $y_i^j \in \{1, 2, ..., C\}$ is the corresponding label. In this work, we consider the object recognition task and aim to learn an Encoder $f_\theta : \mathcal{X} \to \mathcal{Z}$ mapping an input sample x_i into the feature embedding space $f_\theta(x_i) \in \mathcal{Z}$, where θ denotes the parameters of Encoder f_θ. We assume that Encoder f_θ could extract discriminative and transferable features, so that the task network (*e.g.*, classifier) $h_\psi : \mathcal{Z} \to \mathbb{R}^C$ can be prompted on the unseen target domain.

The overall framework of the proposed EISNet is illustrated in Fig. 1. We adopt the classical classification loss, *i.e.*, Cross-Entropy, to minimize the objective $\mathcal{L}_c(h_\psi(f_\theta(x)), y)$ that measures the difference between the ground truth y and the network prediction $\hat{y} = h_\psi(f_\theta(x))$. To avoid performance degradation on unseen target domain, we introduce two additional complementary supervisions to our framework. One is an extrinsic supervision with momentum metric learning, and the other is an intrinsic supervision with a self-supervised auxiliary task. The momentum metric learning is employed by a triplet loss with a K-hard negative selector on the momentum updated embeddings stored in a large memory bank. We implement a self-supervised auxiliary task by predicting the order of patches within an image. All these tasks adopt a shared encoder f and are seamlessly integrated into an end-to-end learning framework. Below, we introduce the extrinsic supervision and intrinsic self-supervision in detail.

3.1 Extrinsic Supervision with Momentum Metric Learning

For the domain generalization problem, it is necessary to ensure the features of samples with the same label close to each other, while the features of different class samples being far apart. Otherwise, the predictions on the unseen



(a) Random selector (b) Semi-hard selector (c) K-hard selector

Fig. 2. The schematic diagram of triplet negative sample selectors. We draw a circle with the anchor (x_a) as the center and the distance between the anchor (x_a) and positive (x_p) as the radius. We ignore the relaxation margin here and set K as 2 for illustration. The selected negatives (x_n) are shown with arrows.

target domain may suffer from ambiguous decision boundaries and performance degradation [8, 20]. This is well aligned with the philosophy of metric learning. Therefore, we design a momentum metric-learning scheme to encourage the network to learn such domain-independent yet class-specific features by considering the mutual relation among samples across domains. Specifically, we propose a novel K-hard negative selector for triple loss to improve the training effectiveness by selecting informative triplets in the memory bank, and a momentum updated encoder to guarantee the representation consistency among the embeddings stored in the memory bank.

***K*-hard negative selector for triplet loss** The triplet loss is widely used to learn feature embedding based on the relative similarity of the sampled pairs. The goal of the original triplet loss is to assign close distance to pairs of similar samples (*i.e.*, positive pair) and long distance to pairs of dissimilar samples (*i.e.*, negative pair). For example, we can extract the feature representation v_i of each image x_i from multi-source domains with the feature Encoder f_θ. Then by fixing an anchor sample x_a, we choose a corresponding positive sample x_p with the same class label as x_a, and a random negative sample x_n with different class label from x_a to form a triplet $\mathcal{T} = \{(x_a, x_p, x_n) | y_a = y_p,\ y_a \neq y_n\}$. Accordingly, the objective of the original triplet loss is formulated as

$$\mathcal{L}_\mathcal{T} = [d(x_a, x_p)^2 - d(x_a, x_n)^2 + \text{margin}]_+, \tag{1}$$

where $[\cdot]_+ = max(0, \cdot)$, $d(x_i, x_j)$ represents the distance between the samples, and the margin is a standard relaxation coefficient. In general, we use the Euclidean distance to measure distances between the embedded features. Then, the distance between samples x_i and x_j is defined as

$$d(x_i, x_j) = \sqrt{\| \ v_i - v_j \ \|^2} = \sqrt{\| \ f_\theta(x_i) - f_\theta(x_j) \ \|^2}. \tag{2}$$

The negative sample selection process in the original triplet loss is shown in Fig. 2(a). Since the selected negative sample may already obey the triplet constraint, the training with the original triplet loss selector may not be efficient. To avoid useless training, inspired by [36,37], we propose a novel K-hard negative online selector, which extends the triplet with K negatives that violate the triplet constraint within a certain of margin. Specifically, given a sampled anchor, we randomly choose one positive sample with the same class label as the anchor, and select K hard negative samples $x_{n_i}, i = \{1, 2, ..., K\}$ following

$$\mathcal{T}' = \{(x_a, x_p, x_{n_i}) | y_a = y_p, y_a \neq y_{n_i}, d(x_a, x_{n_i})^2 < d(x_a, x_p)^2 + \text{margin}\}. \quad (3)$$

In an extreme case, the number of hard negative samples may be zero, then we random select negative samples without the distance constraint. Therefore, the objective of the proposed triple loss with K-hard negative selector can be represented as

$$\mathcal{L}_{\mathcal{T}'} = \frac{1}{K} \sum_{i=1}^{K} [d(x_a, x_p)^2 - d(x_a, x_{n_i})^2 + \text{margin}]_+. \quad (4)$$

We illustrate the triplet selection process of semi-hard selector ($K = 1$) and K-hard selector ($K = 2$) in Fig. 2 (b) (c) for a better understanding. Compared with the original triplet loss, our proposed triple loss equipped with the K-hard negative selector considers more informative hard negatives for each anchor, thus facilitating the feature encoder to learn more discriminative features.

Efficient learning with memory bank The way to select informative triplet pairs has a large influence on the feature embedding. Good features can be learned from a large sample pool that includes a rich set of negative samples [15]. However, selecting K-hard triplets from the whole sample pool is not efficient. To increase the diversity of selected triplet pairs while reducing the computation burden, we maintain memory bank $\mathcal{V}$ to store the feature representation v_i of historical samples [47] with a size of m. Instead of calculating the embedded features of all the images at each iteration, we utilize the stored features to select the K-hard triplet samples. Note that we also keep the class label y_i along with representation v_i in the memory bank to filter the negatives, as shown in Fig. 1. During the network training, we dynamically update the memory bank by discarding the oldest items and feeding the new batch of embedded features, where the memory bank acts as a queue.

Momentum updated encoder With the memory bank, we can improve the efficiency of triplet sample selection. However, the representation consistency between the current samples and historical samples in the memory bank is reduced due to the rapidly-changed encoder [15]. Therefore, instead of utilizing the same feature encoder to extract the representation of current samples and historical samples, we adopt a new **M**omentum **u**pdated **E**ncoder (MuEncoder) to generate feature representation for the samples in the memory bank. Formally, we denote the parameters of Encoder and MuEncoder as θ_f and θ_g, respectively. The Encoder parameter θ_f is optimized by a back-propagation of the loss function, while the MuEncoder parameter θ_g is updated as a moving average of Encoder parameters θ_f following

$$\theta_g = \delta * \theta_g + (1 - \delta) * \theta_f, \quad \text{where } \delta \in [0, 1), \tag{5}$$

where δ is a momentum coefficient to control the update degree of MuEncoder. Since the MuEncoder evolves more smoothly than Encoder, the update of different features in the memory bank is not rapid, thereby easing the triplet loss update. This is confirmed by the experimental results. In our preliminary experiments, we found that a large momentum coefficient δ by slowly updating θ_g could generate better results than rapid updating, which indicates that a slow update of MuEncoder is able to guarantee the representation consistency.

3.2 Intrinsic Supervision with Self-supervised Auxiliary Task

To broaden the network understanding of the data, we propose to utilize the intrinsic supervision within a single image to impose a regularization into the feature embedding by adding auxiliary self-supervised tasks on all the source domain images. A similar idea has been adopted in domain adaptation and Generative Adversarial Networks training [4, 38]. The auxiliary self-supervised task is able to exploit the intrinsic semantic information within a single image to provide informative feature representations for the main task.

There are plenty of works focusing on designing auxiliary self-supervised tasks, such as rotation degree prediction and relative location prediction of two patches in one image [7, 13, 21]. Here, we employ the recently-proposed *solving jigsaw puzzles* [2, 32] as our auxiliary task. However, most of the self-supervised tasks focusing on high-level semantic feature learning can be incorporated into our framework. Specifically, we first divide an image into nine (3×3) patches, and shuffle these patches within the 30 different combinations following [2]. As pointed by [2], the model achieves the highest performance when the class number is set as 30 and the order prediction performance decreases when the task becomes more difficult with more orders. A new auxiliary task branch h_a follows the extracted feature representation f_θ to predict the ordering of the patches. A Cross-Entropy loss is applied to tackle this order classification task:

$$\mathcal{L}_a = -\frac{1}{N * 31} \sum_{i=1}^{N} \sum_{c_a=0}^{30} y^a_{i,c_a} * \log(p^a_{i,c_a}), \tag{6}$$

where y^a and p^a are the ground-truth order and predicted order from the auxiliary task branch, respectively. We use $c_a = 0$ to represent the original images without patch shuffle, leading to a total of 31 classes.

Overall, we formulate the whole framework as a multi-task learning paradigm. The total objective function to train the network is represented as

$$\mathcal{L} = \alpha * \mathcal{L}_c + \beta * \mathcal{L}_{\mathcal{T}'} + \gamma * \mathcal{L}_a, \tag{7}$$

where α, β, and γ are hyper-parameters to balance the weights of the basic classification supervision, extrinsic relationship supervision, and intrinsic self-supervision, respectively.

4 Experiments

4.1 Datasets

We evaluate our method on two public domain generalization benchmark datasets: **VLCS** and **PACS**. **VLCS** [10] is a classic domain generalization benchmark for image classification, which includes five object categories from four domains (PASCAL VOC 2007, LabelMe, Caltech, and Sun datasets). **PACS** [25] is a recent domain generalization benchmark for object recognition with larger domain discrepancy. It consists of seven object categories from four domains (Photo, Art Paintings, Cartoon, and Sketches datasets) and the domain discrepancy among different datasets is more severe than VLCS, making it more challenging.

4.2 Network Architecture and Implementation Details

Our framework is flexible and one can use different network backbones as the feature Encoder. We utilized a fully-connected layer with 31-dimensional output as the self-supervised auxiliary classification layer following the setting in [2] for a fair comparison. To enable the momentum metric learning, we further employed a fully-connected layer with 128 output channels following the Encoder part and added an L2 normalization layer to normalize the feature representation v of each sample. The MuEncoder has the same network architecture as the Encoder, and the weight of MuEncoder was initialized with the same weight as Encoder. We followed the previous works in the literature [1, 2, 8, 26] and employed the leave-one-domain-out cross-validation strategy to produce the experiment results, i.e., we take turns to choose each domain for testing, and train a network model with the remaining three domains.

We implemented our framework with the PyTorch library on one NVIDIA TITAN Xp GPU. Our framework was optimized with the SGD optimizer. We totally trained 100 epochs, and the batch size was 128. The learning rate was set as 0.001 and decreased to 0.0001 after 80 epochs. We empirically set the margin of the triplet loss as 2. We also adopted the same on-the-fly data augmentation as JiGen [2], which includes random cropping, horizontal flipping, and jitter.

4.3 Results on VLCS Dataset

We followed the same experiment setting in previous work [2] to train and evaluate our method. The extrinsic metric learning and intrinsic self-supervised learning was developed upon the "FC7" features of AlexNet [22] pretrained on ImageNet [6]. We set the size of the memory bank as 1024 and the number of negatives K in the triplet loss Eq. (4) as 256. The hyper-parameters α, β, and

Table 1. Domain generalization results on **VLCS** dataset with object recognition accuracy (%) using **AlexNet** backbone. The top results are highlighted in **bold**.

Target	**Within**	**D-MTAE**	**CIDDG**	**CCSA**	**DBADG**	**MMD-**	**MLDG**	**Epi-**	**JiGen**	**MASF**	**EISNet**
	domain	[12]	[29]	[31]	[25]	**AAE** [28]	[26]	**FCR** [27]	[2]	[8]	(Ours)
PASCAL	82.07	63.90	64.38	67.10	69.99	67.70	67.7	67.1	**70.62**	69.14	69.83±0.48
LabelMe	74.32	60.13	63.06	62.10	63.49	62.60	61.3	64.3	60.90	**64.90**	63.49±0.82
Caltech	100.0	89.05	88.83	92.30	93.63	94.40	94.4	94.1	96.93	94.78	**97.33**±0.36
Sun	77.33	61.33	62.10	59.10	61.32	64.40	65.9	65.9	64.30	67.64	**68.02**±0.81
Average	83.43	68.60	69.59	70.15	72.11	72.28	72.3	72.9	73.19	74.11	**74.67**

Table 2. Domain generalization results on **PACS** dataset with object recognition accuracy (%) using **AlexNet** backbone. The top results are highlighted in **bold**.

Target	**Within**	**D-MTAE**	**CIDDG**	**DBADG**	**MLDG**	**Epi-**	**MetaReg**	**JiGen**	**MASF**	**EISNet**
	domain	[12]	[29]	[25]	[26]	**FCR** [27]	[1]	[2]	[8]	(Ours)
Photo	97.80	91.12	78.65	89.50	88.00	86.1	91.07	89.00	90.68	**91.20**±0.00
Art painting	90.36	60.27	62.70	62.86	66.23	64.7	69.82	67.63	70.35	**70.38**±0.37
Cartoon	93.31	58.65	69.73	66.97	66.88	72.3	70.35	71.71	**72.46**	71.59±1.32
Sketch	93.88	47.68	64.45	57.51	58.96	65.0	59.26	65.18	67.33	**70.25**±1.36
Average	93.84	64.48	68.88	69.21	70.01	72.0	72.62	73.38	75.21	**75.86**

γ in total objective function Eq. (7) were set as 1, 0.1, and 0.05, respectively. For our results, we report the average performance and standard deviation over three independent runs.

We compare our method with other nine previous state-of-the-art methods. **D-MTAE** [12] utilized the multi-task auto-encoders to learn robust features across domains. **CIDDG** [29] was a conditional invariant adversarial network that learns the domain-invariant representations under distribution constraints. **CCSA** [31] exploited a Siamese network to learn a discriminative embedding subspace with distribution distances and similarities. **DBADG** [25] developed a low-rank parametrized CNN model for domain generalization. **MMD-AAE** [28] aligned the distribution through an adversarial auto-encoder by Maximum Mean Discrepancy. **MLDG** [26] was a meta-learning method by simulating train/test domain shift during training. **Epi-FCR** [27] was an episodic training method. **JiGen** [2] solved a jigsaw puzzle auxiliary task based on self-supervision. **MASF** [8] employed a meta-learning based strategy with two complementary losses for encoder regularization. Moreover, we include the **Within domain** performance of all the datasets as a comparison to reveal the performance drop due to domain discrepancy. We trained **Within domain** using a supervised way with training and test images from the same domain.

The comparison results with the above methods are shown in Table 1. It is observed that our EISNet achieves the best performance on both Caltech and Sun datasets and comparable results on PASCAL VOC and LabelMe datasets. Overall, EISNet achieves an average accuracy of 74.67% over four domains, outperforming the previous state-of-the-art method **MASF** [8]. Our method also outperforms **JiGen** [2] on three domains and achieves comparable results on the remaining PASCAL VOC domain, demonstrating that utilizing extrinsic relationship supervision can further improve the network generalization ability.

Table 3. Domain generalization results on **PACS** dataset with object recognition accuracy (%) using **ResNet** backbones. The top results are highlighted in **bold**.

Target	ResNet-18			ResNet-50		
	DeepAll	**MASF** [8]	**EISNet** (Ours)	**DeepAll**	**MASF** [8]	**EISNet** (Ours)
Photo	94.25	94.99	**95.93**±0.06	94.83	95.01	**97.11**±0.40
Art painting	77.38	80.29	**81.89**±0.88	81.47	82.89	**86.64**±1.41
Cartoon	75.65	**77.17**	76.44±0.31	78.61	80.49	**81.53**±0.64
Sketch	69.64	71.69	**74.33**±1.37	69.69	72.29	**78.07**±1.43
Average	79.23	81.04	**82.15**	81.15	82.67	**85.84**

4.4 Results on PACS Dataset

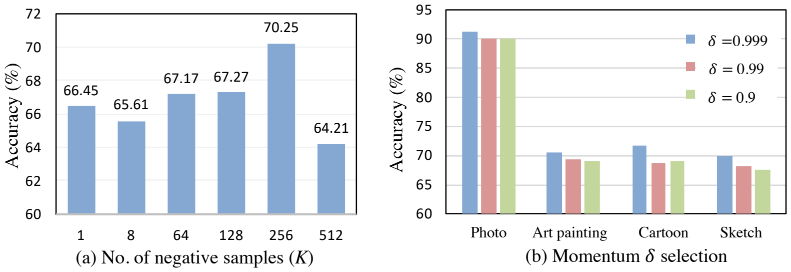
To show the effectiveness of our framework under different network backbones on PACS dataset, we evaluate our method with three different backbones: AlexNet, ResNet-18, and ResNet-50 [16]. The size of memory bank was set as 1024 and K in the triplet loss Eq. (4) was set as 256. The hyper-parameters in total objective function Eq. (7) were set as 1, 0.5, and 0.7 for α, β, and γ, respectively. For our results, we also report the average performance and standard deviation over three independent runs.

Table 2 summarizes the experimental results developed with AlexNet backbone. We compare our methods with eight other methods that achieved previous best results on this benchmark dataset. **MetaReg** [1] utilized a novel classifier regularization in the meta-learning framework. As we can observe from Table 2, by simultaneously utilizing momentum metric learning and intrinsic self-supervision for images across different source domains, our method achieves the best performance on three datasets. Across all domains, our method achieves an average accuracy of 75.86%, setting a new state-of-the-art performance.

We also compare our method with baseline method (DeepAll) and the state-of-the-art method **MASF** [8] using ResNet-18 and ResNet-50 backbones in Table 3. In the ResNet-50 experiment, we reduce the batch size to 64 to fit the limited GPU memory. The DeepAll method is trained with all the source domains without any specific network design. As shown in Table 3, our method consistently outperforms **MASF** about 1.11% and 3.17% on average accuracy with ResNet-18 and ResNet-50 backbone, respectively. This indicates that our designed framework is very general and can be migrated to different network backbones. Note that the improvement over **MASF** is more obvious with a deeper network backbone, showing that our proposed algorithm is more beneficial for domain generalization with deeper feature extractors.

4.5 Analysis of Our Method

We conduct extensive analysis of our method. Firstly, we investigate the effectiveness of extrinsic and intrinsic supervision using ResNet-50 backbone on **PACS** dataset, and the experimental results are illustrated in Table 4. The **Extrinsic**



(a) No. of negative samples (K)

(b) Momentum δ selection

Fig. 3. The performance of our method under different number of negative samples K and momentum update coefficient δ.

Table 4. Ablation study on key components of our method with the **PACS** dataset (%). The top results are highlighted in **bold**.

Extrinsic	**Intrinsic**	**P**hoto	**A**rt painting	**C**artoon	**S**ketch	**Average**
-	-	94.85	81.47	78.61	69.69	81.15
✓	-	97.06	81.97	80.70	76.81	84.14
-	✓	97.02	85.17	76.35	76.97	83.88
✓	✓	**97.11**	**86.64**	**81.53**	**78.07**	**85.84**

Table 5. Comparison of our proposed K-hard negative selector with original random selector and semi-hard negative selector.

Selector	Random	Semi-hard	K-hard
Accuracy (%)	65.38	68.08	70.25

supervision indicates that the momentum metric learning is used, while **Intrinsic** supervision denotes that the auxiliary self-supervision loss is optimized. The method without these two supervisions is the baseline model, which is the same with DeepAll results in Table 3. From the results in Table 4, we observe that each supervision plays an important role in our framework. Specifically, equipping the extrinsic supervision into the baseline model yields about 2.99% average accuracy improvement. Meanwhile, we also achieve 2.73% average accuracy improvement over the baseline model by incorporating intrinsic self-supervision of the images. By combing extrinsic and intrinsic supervision, performance is further improved across all settings, indicating these two supervisions are complementary.

We then analyze five key components in our framework, that is a) the number of different negative samples K in momentum metric learning, b) the effectiveness of momentum update coefficient δ, c) the effectiveness of hard negative selector, d) the size of memory bank m, and e) time cost. All below comparison experiments are implemented with AlexNet backbone on the PACS benchmark.

Table 6. Comparison among different memory bank size.

Memory bank size m	1024	512	256	128
No. of negatives K	256	128	64	32
Accuracy (%)	70.25	68.80	68.24	67.93

a. The number of negative samples K is a key parameter of our designed K-hard negative selector in momentum metric learning. We investigate the network performance under different options. We select six K values at different magnitudes, which are 1, 8, 64, 128, 256, and 512. The Sketch dataset results are shown in Fig. 3 (a), We can observe that a large number of negative samples would lead to better results in general and the network generates the best result with $K = 256$. However, the performance drops drastically if we set $K = 512$, demonstrating that too large K will produce a burden on the metric distance calculation and make the network difficult to learn.

b. The momentum update coefficient δ is important to control the feature consistency among different batches of embedded features in the memory bank. We show the accuracy with different momentum coefficient δ in Fig. 3 (b). It is observed that the network performs well when δ is relatively large, *i.e.*, 0.999. A small coefficient would degrade the network performance, suggesting that a slow updating MuEncoder is beneficial to the feature consistency.

c. To validate the effectiveness of K-hard negative selector in our proposed metric learning, we compare our proposed K-hard negative selector with original random triplet selector and semi-hard negative selector. The Sketch dataset results are shown in Table 5. Equipped with semi-hard negative selector, the accuracy improves 2.70%. By selecting more negative pairs from the memory bank, we obtain the accuracy of 70.25%, demonstrating the effectiveness of the proposed K-hard negative selector.

d. The size of memory bank m can be adjusted according to different tasks. Here, we show the results of four different settings with the number of negatives changing as well in Table 6. In general, our method is able to generate better results with a large memory bank size and negative samples. However, a too large memory bank will increase the burden to calculate the pair-wise distance in triplet loss. Therefore, we need to balance the accuracy and computation burden.

e. Apart from the performance improvement over other methods, our method has much lower computation cost. Under the same server setting (one TITAN XP GPU) and AlexNet backbone, our method only takes 1.5 hours to train the network on PACS dataset, while the total training time of the state-of-the-art MASF is about 17 hours. Therefore, our method could save more than 91% time cost on training phase.

We also employ t-SNE [30] to analyze the feature level discrimination of our method and the visualization results are shown in Fig. 4. Compared with the feature extracted from the ImageNet pre-trained network, the distance between

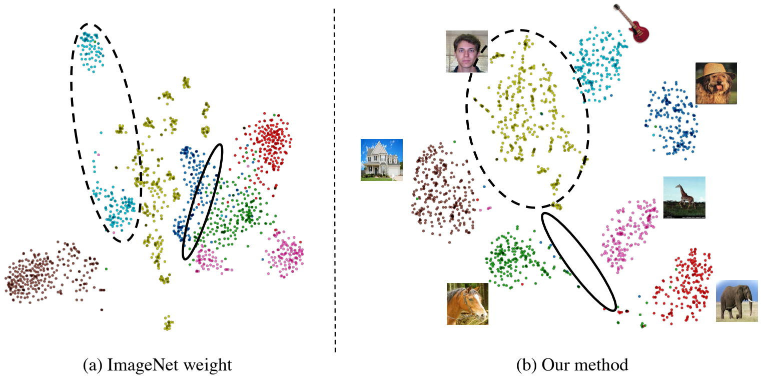


Fig. 4. t-SNE visualization on one target domain to show the discrimination of the network. (a) is the feature embedding extracted from the IMAGENET pre-trained network. (b) shows the feature embedding distributions extracted from our EISNet.

different class clusters in our method becomes evident, indicating that equipped with our proposed extrinsic and intrinsic supervision, the model is able to learn more discriminative features among different object categories regardless domains.

5 Conclusions

We have presented a multi-task learning paradigm to learn how to generalize across domains for domain generalization. The main idea is to learn a feature embedding simultaneously from the extrinsic relationship of different images and the intrinsic self-supervised constraint within the single image. We design an effective and efficient momentum metric learning module to facilitate compact feature learning. Extensive experimental results on two public benchmark datasets demonstrate that our proposed method is able to learn discriminative yet transferable feature, which lead to state-of-the-art performance for domain generalization. Moreover, our proposed framework is flexible and can be migrated to various network backbones.

Acknowledgments. We thank anonymous reviewers for the comments and suggestions. The work described in this paper was supported in parts by the following grants: Key-Area Research and Development Program of Guangdong Province, China (2020B010165004), Hong Kong Innovation and Technology Fund (Project No. ITS/426/17FP and ITS/311/18FP), and National Natural Science Foundation of China with Project No. U1813204.

References

1. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: MetaReg: Towards domain generalization using meta-regularization. In: Advances in Neural Information Processing Systems. pp. 998–1008 (2018)
2. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2229–2238 (2019)
3. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In: Proceedings of The Thirty-Third Conference on Artificial Intelligence (AAAI). pp. 865–872 (2019)
4. Chen, T., Zhai, X., Ritter, M., Lucic, M., Houlsby, N.: Self-supervised gans via auxiliary rotation loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12154–12163 (2019)
5. Chen, Y., Li, W., Van Gool, L.: ROAD: Reality oriented adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7892–7901 (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
7. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1422–1430 (2015)
8. Dou, Q., de Castro, D.C., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. In: Advances in Neural Information Processing Systems. pp. 6447–6458 (2019)
9. Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.A.: Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 691–697 (2018)
10. Fang, C., Xu, Y., Rockmore, D.N.: Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1657–1664 (2013)
11. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
12. Ghifary, M., Bastiaan Kleijn, W., Zhang, M., Balduzzi, D.: Domain generalization for object recognition with multi-task autoencoders. In: Proceedings of the IEEE international conference on computer vision. pp. 2551–2559 (2015)
13. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
14. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1735–1742. IEEE (2006)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
16. 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 (2016)

17. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition. pp. 84–92. Springer (2015)
18. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
19. Hong, W., Wang, Z., Yang, M., Yuan, J.: Conditional generative adversarial network for structured domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1335–1344 (2018)
20. Kamnitsas, K., Castro, D.C., Folgoc, L.L., Walker, I., Tanno, R., Rueckert, D., Glocker, B., Criminisi, A., Nori, A.: Semi-supervised learning via compact latent space clustering. arXiv preprint arXiv:1806.02679 (2018)
21. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1920–1929 (2019)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. pp. 1097–1105 (2012)
23. Kumar, A., Saha, A., Daume, H.: Co-regularization based semi-supervised domain adaptation. In: Advances in Neural Information Processing Systems. pp. 478–486 (2010)
24. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: European Conference on Computer Vision. pp. 577–593. Springer (2016)
25. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5542–5550 (2017)
26. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Learning to generalize: Meta-learning for domain generalization. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
27. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., Hospedales, T.M.: Episodic training for domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1446–1455 (2019)
28. Li, H., Jialin Pan, S., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5400–5409 (2018)
29. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 624–639 (2018)
30. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
31. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5715–5725 (2017)
32. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision. pp. 69–84. Springer (2016)
33. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544 (2016)

34. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 201–209. Springer (2018)
35. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018)
36. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823 (2015)
37. Sohn, K.: Improved deep metric learning with multi-class N-pair loss objective. In: Advances in Neural Information Processing Systems. pp. 1857–1865 (2016)
38. Sun, Y., Tzeng, E., Darrell, T., Efros, A.A.: Unsupervised domain adaptation through self-supervision. arXiv preprint arXiv:1909.11825 (2019)
39. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)
40. Tsai, Y.H., Sohn, K., Schulter, S., Chandraker, M.: Domain adaptation for structured output via discriminative patch representations. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1456–1465 (2019)
41. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4068–4076 (2015)
42. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176 (2017)
43. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
44. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1386–1393 (2014)
45. Wang, S., Yu, L., Yang, X., Fu, C.W., Heng, P.A.: Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Transactions on Medical Imaging **38**(11), 2485–2495 (2019)
46. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**(Feb), 207–244 (2009)
47. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018)
48. Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 814–823 (2017)
49. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., Gong, B.: Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2100–2110 (2019)

50. Zakharov, S., Kehl, W., Ilic, S.: DeceptionNet: Network-driven domain randomization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 532–541 (2019)