Grounding Visual Representations with Texts for Domain Generalization

Seonwoo Min[1], Nokyung Park[2], Siwon Kim[3], Seunghyun Park[4], and Jinkyu Kim[2]

[1] LG AI Research, South Korea
[2] Computer Science and Engineering, Korea University, South Korea
[3] Electrical and Computer Engineering, Seoul National University, South Korea
[4] Clova AI Research, NAVER Corp., South Korea
Correspondence: jinkyukim@korea.ac.kr

Abstract. Reducing the representational discrepancy between source and target domains is a key component to maximize the model generalization. In this work, we advocate for leveraging natural language supervision for the domain generalization task. We introduce two modules to ground visual representations with texts containing typical reasoning of humans: (1) *Visual and Textual Joint Embedder* and (2) *Textual Explanation Generator*. The former learns the image-text joint embedding space where we can ground high-level class-discriminative information into the model. The latter leverages an explainable model and generates explanations justifying the rationale behind its decision. To the best of our knowledge, this is the first work to leverage the vision-and-language cross-modality approach for the domain generalization task. Our experiments with a newly created CUB-DG benchmark dataset demonstrate that cross-modality supervision can be successfully used to ground domain-invariant visual representations and improve the model generalization. Furthermore, in the large-scale DomainBed benchmark, our proposed method achieves state-of-the-art results and ranks 1st in average performance for five multi-domain datasets. The dataset and codes are available at https://github.com/mswzeus/GVRT.

Keywords: Domain generalization, Image Classification, Textual Explanation, Visual-Textual Joint Embedding

1 Introduction

Machine learning systems assume that in-samples (training) and out-of-samples (test) are independent and identically distributed – this assumption, however, rarely holds in real-world scenarios where domain shift often occurs. Various domain generalization (DG) approaches have been introduced to make models generalize well to unseen novel domains. They mainly focus on learning domain-invariant representations so that the model can leverage such invariances during deployment in unseen test domains. In the DG task, samples from target domains are not available during training, thus these approaches are different from domain

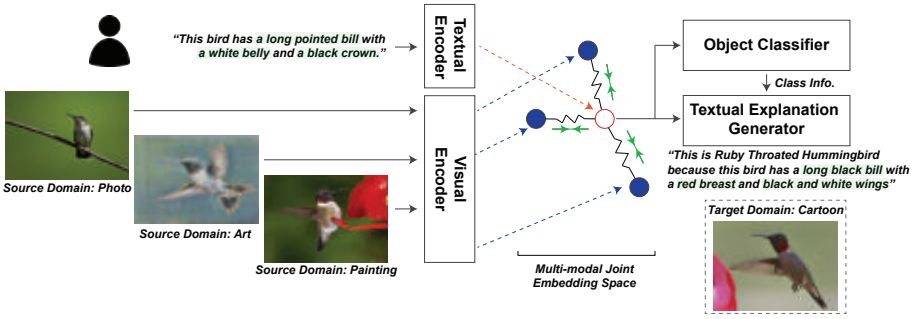

Fig. 1. Our model leverages the text modality by (1) *Visual and Textual Joint Embedder* and (2) *Textual Explanation Generator*. Our model takes advantage of a *pivot* embedding (red circle) from a sentence that describes the class discriminative evidence in a natural language, e.g. a white belly or a pointy beak. Our visual encoder is optimized to produce an embedding (blue-filled circles) that aligns well with the corresponding pivot embedding. The latter further trains the model to justify why it made a certain prediction in a natural language.

adaptation (DA), semi-supervised domain adaptation (SSDA), and unsupervised domain generalization (UDA) [10].

Reducing the discrepancy between source and target domains is a key component to maximize the model generalization. This is often achieved by (i) explicitly matching the feature distribution across domains using a similarity metric to measure the distance between each domain or by (ii) using contrastive loss to map the latent representations of positive pairs close together and those of negative pairs further away in the feature space. Such distance-based approaches need to optimize all pairwise sample distances, thus potentially resulting in models that are susceptible to outliers and unfair to subgroups in the imbalanced data – where the classes are not represented equally.

Distribution of training data also limits the networks' understanding of the data. In computer vision models, their opaque reasoning can be simplified to a situation-specific dependence on visible objects in the image. However, instead of learning the true semantics, they often attend to background objects that are salient to the class labels. (e.g. attending to sea for classifying boats). these models will likely behave well in environments similar to those for which it was trained but typically will not generalize well beyond them [9]

To address this issue, we propose a novel approach that grounds visual representations with explicit (verbalized) knowledge from humans about typical reasoning on visual cues (Figure 1). For example, our model learns to understand the user's utterance ("an elephant is a heavy plant-eating mammal with a prehensile trunk, long curved ivory tusks, and large ears.") and ground it in the trained perceptual primitives. To ground (or internalize) explicit knowledge, we use the following two modules: (1) *Visual and Textual Joint Embedder* and

(2) *Textual Explanation Generator*. The former aligns the perceptual primitives with the (verbalized) thought process of humans by minimizing the distance between the textual and the visual latent representations. The latter leverages the representational power of explainable models. Regardless of image domains, we train the model to consistently verbalize why it made a certain prediction with natural language, e.g. "This is Ruby Throated Hummingbird because this bird has a long pointed bill with a white belly and a black crown."

To the best of our knowledge, this is the first work to leverage the vision-and-language cross-modality approach for the DG task. For the empirical evaluations under natural language supervision, we created a new benchmark built upon the Caltech UCSD Birds 200-2011 (CUB) dataset [45]. Our quantitative and qualitative experiment results demonstrate that cross-modality supervision can be successfully used to improve the model representational generalization power as well as to justify its visual predictions. Furthermore, we conducted large-scale experiments on the DomainBed benchmark [10], a popular testbed for DG algorithms. The proposed method achieved state-of-the-art results and ranked ranks 1st in average performance for five multi-domain datasets.

2 Related Work

Domain Generalization. Generating domain-invariant representations is the key component in the DG task. Such learned invariances can be leveraged to improve the model generalization to unseen test domains. Of a landmark work, Empirical Risk Minimization (ERM) minimizes the sum of errors across domains, thus matching distributions across different domains [37]. Along this line of work, notable variants have been introduced. DANN [8] and CDANN [22] utilized an adversarial network to minimize unconditional and class-conditional distributional differences across domains, respectively. Such a shared feature space is also optimized by different distance metrics: i.e. maximum mean discrepancy [21], transformed feature distribution distance [25], and covariances (CORAL) [36]. Inter-domain mixup techniques [49,48,44] were introduced to perform ERM on linearly interpolated examples from random pairs across domains. SelfReg [15] leveraged the self-supervised learning approaches to address the unstable training, which is often caused by the usage of negative pairs.

In this work, we explore the benefit of grounding visual representations by using cross-modality supervision. We introduce two modules for leveraging texts containing the thought process of humans. First, we train a model in the image-text joint embedding space where we can ground high-level class-discriminative information into the model. Second, we adopt an explainable model that can generate explanations justifying the rationale behind its decision.

Visual and Textual Explanations. Explainability and interpretation of deep neural networks have become increasingly important in various machine learning communities [11,16]. In computer vision, numerous works have explored explaining a target model through visualizations. Early works obtain visual explanations

through deconvolutions of layer activations [50] or synthesizing those that maximize the network output [52]. Attention-based approaches try to measure how spatial features formally affect the network output [42,47]. They directly extract salient areas of a given image that the network pays the most attention to produce its output. On the other hand, some works emphasize the importance of justifying the model decision in a human-understandable manner, i.e. in natural language. They adopt an encoder-decoder framework which is usually composed of a convolutional neural network (CNN) as the encoder and a long short-term memory (LSTM) caption generator as the decoder [12,13]. The latter generates textual explanations from the representations produced from the former.

Following this stream of work, we advocate for leveraging the representational power of explainable models for the DG task. Especially, generating textual justifications requires capturing class-discriminative and high-level semantic information. Therefore, we argue that it can help ground domain-invariant visual representations and improve the model generalization.

3 Method

In this paper, we aim to solve the DG problem: i.e. we train a model on a single or multiple source domains $\{\mathcal{S}_1, \mathcal{S}_2, \dots\} \in \mathcal{S}$ and evaluate it on unseen target domains, $\{\mathcal{T}_1, \mathcal{T}_2, \dots\} \in \mathcal{T}$. Formally, we train a model by minimizing the following data-dependent upper bound on the expected worst-case loss [34]:

$$\underset{\theta}{\text{minimize}} \sup_{\mathcal{T}:\mathcal{D}(\mathcal{S},\mathcal{T})\leq\rho} \mathbb{E}\big[\mathcal{L}_{\text{task}}(\mathcal{S};\theta)\big] \tag{1}$$

where a dissimilarity $D(\mathcal{S},\mathcal{T})$ is used to measure the discrepancy between $\mathcal{S}$ and $\mathcal{T}$ with an arbitrary upper bound ρ. $\mathcal{L}_{\text{task}}$ is a task-specific loss function over a model parameter θ where we use the following cross-entropy loss as we focus on the classification problem:

$$\mathcal{L}_{\text{task}}(\mathcal{S};\theta) = \mathcal{L}_{\text{task}}(\mathbf{y},\hat{\mathbf{y}}) = -\sum_i y_i \log(\hat{y}_i) \tag{2}$$

where $\mathbf{y}$ is the one-hot vector representing each label's class and $\hat{\mathbf{y}}$ is the softmax distribution produced from the visual feature $\mathbf{x}$.

A key component of the DG task is to address the problem is learning domain-invariant representations that help improve the model generalization. In this work, we advocate for leveraging cross-modality supervision with semantic cues. Specifically, as shown in Figure 2, we use the following two main modules to ground visual representations with texts containing class-discriminative and high-level semantic information. First, *Visual and Textual Joint Embedder* encourages our visual encoder to produce a latent representation that is aligned with textual semantics in the joint embedding space. Second, *Textual Explanation Generator* produces a class-discriminative sentence detailing how visual evidence is compatible with a class prediction. Note that both modules are only

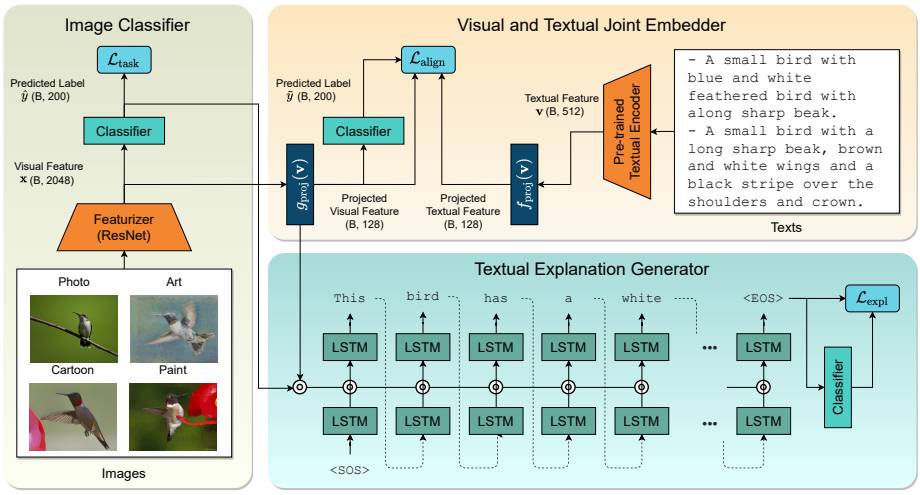

Fig. 2. An overview of our proposed model. An image classifier is trained by minimizing the cross-entropy loss $\mathcal{L}_{\text{task}}$ given images from source domains. Built upon it, our model incorporates two modules for improving DG performance: (i) *Visual and Textual Joint Embedder*, which produces a joint latent representation that is aligned with textual semantics, and (ii) *Textual Explanation Generator*, which generates a class-specific sentence detailing how visual evidence is compatible with a system prediction.

required during the training phase for grounding the visual encoder. Nevertheless, the latter can be also optionally used during the inference to obtain textual explanation along with a class prediction.

3.1 Visual and Textual Joint Embedder

Our Visual and Textual Joint Embedder aims to learn domain-invariant visual representations from textual explanations. We argue that learning from natural language supervision has potential strength over image-only training approaches, especially for the DG task. In contrast to the vulnerability of CNNs against domain shift, the human visual recognition system generalizes well across domains, e.g. even very young children can easily transfer object concepts from picture books to the real world [7]. Thus, we advocate for using explicit knowledge from humans and we train a model to better align with the thought process of humans via their textual explanations. It ultimately provides more semantically-rich information compared to standard crowd-sourced labeling for image classification.

We use a sentence-level textual encoder, which takes a variable-length sentence and yields a fixed-size latent vector $\mathbf{v}$. Given the *pivot* textual latent representations $\mathbf{v}$, we optimize the visual encoder to produce representations $\mathbf{x}$ that align well with the corresponding pivot. Thus, our model needs to understand the textual justification from human annotators and to map it into the image-

text joint embedding space. We assume that such textual justification will often contain class-discriminative evidence reflecting visual semantic cues, thus our visual encoder can internalize knowledge from natural language supervision.

Specifically, we minimize the following loss function $\mathcal{L}_{\text{align}}$ based on l_2 distance between the projected visual and texture features:

$$\mathcal{L}_{\text{align}} = ||f_{\text{proj}}(\mathbf{v}) - g_{\text{proj}}(\mathbf{x})||_2 - \sum_i y_i \log(\tilde{y}_i) \tag{3}$$

where f_{proj} and g_{proj} are the projection layers for text and visual feature, respectively. $\mathbf{y}$ is the one-hot vector representing each label's class and $\tilde{y}_i$ is the softmax distribution produced from the projected visual feature $g_{\text{proj}}(\mathbf{x})_i$. Note that we use the second cross-entropy term to make the projected visual features more class-discriminative. It prevents collapsing into collapsing solutions, e.g., always projecting them to the same point.

Pre-trained (Supervised) Textual Encoder (PTE). One way to obtain the pivot textual latent representation is via pre-trained language models. These pre-trained encoders can be adopted from off-the-shelf sentence-level textual encoders that are often pre-trained with a large-scale dataset. In this work, we adopt the widely used CLIP (i.e. Contrastive Language-Image Pre-Training) model, which can embed texts and images into the joint representation space [30]. The text encoder of CLIP is a $63M$-parameter Transformer architecture with 12-layer, 512-wide, and 8 attention heads [39]. It was jointly trained with a Vision Transformer (ViT)-based image encoder [5] to predict the pairing of texts and images. In this work, we only used the text encoder of the CLIP-ViT-B-32 model, but other pre-trained language models are also applicable.

Self-supervised Textual Encoder (STE). Another way to obtain the pivot textual latent representation is via self-supervision. Since our textual explanation generator justifies the rationale behind the model in the natural language, we can use it as a self-supervised textual encoder. As we will explain in the next subsection, during the training, we iteratively sample a sentence from an LSTM-based explanation generator to compute its training loss. Therefore, it is an intuitive choice to use its last hidden states as a fixed-size latent vector $\mathbf{v}$.

3.2 Textual Explanation Generator

Our textual explanation generator is similar to image captioning models based on an encoder-decoder framework. It contains a two-layer LSTM network that takes high-level features from the visual encoder as input and generates variable-length per-word softmax probabilities. The difference is that it is trained to explain the rationale behind the classifier, reflecting typical visual semantic cues. Since it needs the prediction outputs from the classifier as an input as well, we concatenate the category information with a projected visual feature $g_{\texttt{proj}}(\mathbf{x})$. The concatenated vector is then used to update the LSTM network for a textual explanation generation.

Specifically, the first LSTM layer takes the previously generated output token o_{t-1} as input and updates its hidden state, producing an output $\mathbf{z_t}$. This output is then fed into the second LSTM layer along with the concatenated vector of projected visual features and prediction outputs. The second LSTM layer yields the per-word softmax probabilities $p(o_t)$. Further, following [12], we use the discriminative sentence generation loss function based on reinforcement learning so that a model learns to generate sentences that are more likely to be class-discriminative. Specifically, we first sample a sentence from the textual explanation generator and we minimize the expectation of the negative reward $-R(\tilde{o})$ over the sampled sentences $\tilde{o} \sim p(o|\mathcal{I}, \mathcal{C})$. The probability distribution $p(o|\mathcal{I}, \mathcal{C})$ is the model's estimated conditional distribution over descriptions o conditioned on the input image $\mathcal{I}$ and the category $\mathcal{C}$. Concretely, for training our textual explanation generator, we minimize the following loss function $\mathcal{L}_{\text{expl}}$:

$$\mathcal{L}_{\text{expl}} = -\sum_{t} \log p(o_{t+1}|o_{0:t}, \mathcal{I}, \mathcal{C}) - \mathbb{E}_{\tilde{o}\sim p(o|\mathcal{I},\mathcal{C})}\big[R(\tilde{o})\big] \tag{4}$$

We use the reward function as $R(\tilde{o}) = p(\mathcal{C}|\tilde{o})$, which is the per-class softmax probabilities over the category $\mathcal{C}$ conditioned on the generated sentence $\tilde{o}$. A more class-discriminative sentence receives a higher reward. Using REINFORCE [46] algorithm, we compute the following expected reward gradient as:

$$\nabla_\theta \mathbb{E}_{\tilde{o}\sim p(o|\mathcal{I},\mathcal{C})}\big[R(\tilde{o})\big] = \mathbb{E}_{\tilde{o}\sim p(o|\mathcal{I},\mathcal{C})}\big[R(\tilde{o})\nabla_\theta \log p(\tilde{o})\big] \tag{5}$$

Loss function. To summarize, we train our entire model end-to-end by minimizing the following loss function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{expl}}\mathcal{L}_{\text{expl}} \tag{6}$$

where we use hyperparameters λ_{align} and λ_{expl} to control the strengths of each training objective term.

4 Caltech UCSD Birds - Domain Generalization Extension (CUB-DG) Dataset

No previous DG benchmarks provide viable natural language supervision. Thus, in order to thoroughly investigate the effectiveness of the cross-modality supervision in the DG task, we have created a new benchmark built upon the CUB dataset [45]. This dataset contains overall 11,788 images for 200 classes of North American bird species. Ten sentences for each of the images have been previously collected [31], which provides a detailed description of the content of the image, e.g., "this bird has a long pointed bill with a white belly and a black crown." This dataset has been an ideal benchmark for the visual explanation task as sentences are class-specific and class-discriminative.

CUB Dataset for Domain Generalization Since the CUB dataset is only composed of the Photo domain, we used pre-trained style transfer models to

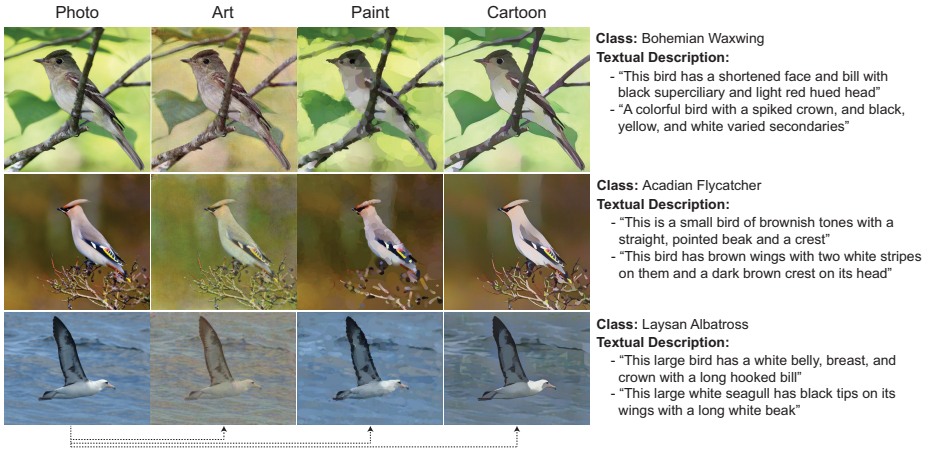

Fig. 3. We create an extended dataset for the DG task based on Caltech UCSD Birds 200-2011 (CUB) dataset. This dataset provides a pair of images and detailed descriptions of the content of the image, e.g. a description "this is a small bird of brownish tones with a straight, pointed beak and a crest" for Acadian Flycatcher. On top of this dataset, we applied off-the-shelf style transfer techniques to obtain images from three other domains: Art, Paint, and Cartoon.

obtain images from three other domains, i.e. Art, Paint, and Cartoon. For the Photo-to-Art translation, we used the CycleGAN [54] *Monet* model which was trained in the absence of paired examples based on adversarial and cycle-consistency losses. For the Photo-to-Paint translation, we used the *Watercolor* neural render model [55]. It imitates painting creation processes by producing a sequence of strokes. For the Photo-to-Cartoon translation, we used a generative adversarial network model which separately identifies surface, structure, and texture representations of cartoons [43].

The generated CUB-DG dataset contains 11,768 sets of images and corresponding text descriptions. Each set illustrates the same content of a bird in four different domains. To evaluate DG algorithms in common experimental protocols, we used the following data split procedures (Figure S1). We start from the official split of the CUB dataset where the train-validation and test sets consist of 5,994 and 5,794 samples, respectively. We divide the train-validation set into three groups. For the multi-source DG task, we select a different group from each source domain, so that the different domains do not share the *siblings* of the same image. For the single-source DG task, we use all three groups from a source domain. For both tasks, we evaluate DG algorithms on the test set from unseen target domains. Note that the CUB-DG dataset holds evident domain shifts such that an ERM model trained only on the Photo domain performs well on the same domain (71.2 accuracy; results not shown) but significantly deteriorates on the other domains (Table 2).

Table 1. Out-of-distribution test accuracies on the CUB-DG benchmark dataset. We compare with 12 DG algorithms in the multi-source DG setting. Note that we use the validation set (from source domains) for the model selection. *Abbr.* D: learning domain-invariant features by matching distributions across different domains, A: adversarial learning strategy, M: inter-domain mix-up, T: learning textual representations. PTE: pre-trained (supervised) textual encoder, STE: self-supervised textual encoder.

Model	D	A	M	T	Target Domain				Avg
					Photo	Cartoon	Art	Paint	
Ours w/ PTE	✓			✓	**74.6**	**64.2**	**52.2**	37.0	**57.0**
Ours w/ STE	✓			✓	74.3	63.9	50.0	**38.1**	56.6
CORAL [36]	✓				72.2	63.5	50.3	35.8	55.4
SD [29]					71.3	62.2	50.8	34.8	54.7
SagNet [26]	✓	✓			67.4	60.7	44.0	34.2	51.6
MixStyle [53]			✓		59.0	56.7	50.3	35.8	50.4
Mixup [49]			✓		67.1	55.9	51.1	27.2	50.3
DANN [8]	✓	✓			67.5	57.0	42.8	30.6	49.5
CDANN [22]	✓	✓			65.3	55.2	43.2	30.5	48.6
VREx [17]	✓				63.9	54.9	38.6	30.1	46.9
ERM [38]					62.5	53.2	37.4	29.0	45.5
ARM [51]					62.3	51.2	38.2	28.4	45.0
GroupDRO [32]	✓				60.9	54.8	36.5	27.0	44.8
IRM [1]					60.6	51.6	36.5	30.3	44.8

5 Experiments

Multi-Source Domain Generalization Performance. We first look into the multi-source DG task, where a single domain is used as a test domain and the others as training domains in rotation. We compare our model with 11 DG algorithms from DomainBed on our newly created CUB-DG dataset. Compared methods include ERM [38], IRM [1], GroupDRO [32], Mixup [49], CORAL [36], DANN [8], CDANN [22], SagNet [26], MixStyle [53], ARM [51], VREx [17], and SD [29]. We report averaged results across three independent runs. Please refer to the supplementary materials for complete implementation details.

We observe in Table 1 that our proposed models outperform the other recent approaches in all test domains (compare the top two rows vs. others), and the average image recognition accuracy is 1.6-12.2% better than alternatives. While the performance difference between our model variants is marginal, we also observe that our model with the PTE generally shows better performance than a model with the STE. Therefore, in the following, we focus on analyzing our model with the PTE. Our model can be used together with other approaches (e.g. SD [29] and SWA [4]) that are not based on matching distributions across domains, which would be worth exploring as future work.

Table 2. Out-of-distribution test accuracies in the single-source DG setting where we train our model with a single source domain (rows) and evaluate with other remaining target domains (columns). We compare with SD [29] and report differences between ours in the last row (+ indicates that ours performs better).

SD [29]	Target Domain					**Ours**	Target Domain				
	Photo	Cartoon	Art	Paint	Avg.		Photo	Cartoon	Art	Paint	Avg.
Photo	-	42.4	51.3	20.4	38.0	Photo	-	49.1	54.2	19.5	40.9
Cartoon	66.9	-	29.3	34.6	43.6	Cartoon	69.5	-	33.6	36.3	46.5
Art	69.0	33.4	-	15.7	39.4	Art	75.6	37.9	-	16.3	43.2
Paint	58.0	49.9	30.0	-	46.0	Paint	63.7	57.3	35.6	-	52.2
Avg	64.6	41.9	36.9	23.6	41.7	Avg	69.6	48.1	41.1	24.0	45.7
							(+5.0%)	(+6.2%)	(+4.3%)	(+0.5%)	(+4.0%)

Single-Source Domain Generalization Performance. We also evaluate our model in an extreme case for the DG task, i.e. single-source DG. In this setting, we assume that only a single domain is available during the training. We then evaluate with examples from all the other remaining target domains. In Table 2, we compare ours with those of SD [29]. We show differences between ours and SD in the last row (+ indicates that ours performs better) and present them as a heatmap in the Figure S2. Additionally, the full results for comparing our model with six DG algorithms are also available in the Table S1. We excluded algorithms that are inapplicable for the single-source DG setting. We report scores for all source-target combinations, i.e. rows and columns for source and target domains, respectively. The scores are averaged across three independent runs. We observe in Table 2 that ours outperforms alternatives, where the average accuracy is improved by 4.0% than SD [29].

Ablation Studies. To better understand different aspects of our proposed models, we present results from ablation studies. We vary our base model in several directions and measured the performance on the multi-source DG task. We report averaged results across three independent runs.

First, we vary the amount of natural language supervision. While we have assumed *Per-Image* texts are available, obtaining them across different domains may not be easy in real life. Therefore, we introduce a more practical scenario where we only use the same single sentence for all the images within each class. Intuitively, it can be understood as *Per-Class* textual definitions. In Table 3 row (A), we observe that even with the *Per-Class* texts, the cross-modality supervision still enables outperforming all the compared DG algorithms in Table 1.

In Table 3 rows (B), we investigate the importance of each module. We observe that removing the Joint Embedder significantly hurts the DG performance. The Explanation Generator plays a complementary role in grounding visual representations by justifying model predictions in natural language. In rows (C), we look into the sensitivity to the hyperparameters λ_{align} and λ_{expl}. We can

Table 3. Results from ablation studies. We vary our base model in several directions and measured the performance on the multi-source DG task.

	Available Texts	Joint Embedder	Explanation Generator	λ_{align}	λ_{expl}	Target Domain				Avg
						Photo	Cartoon	Art	Paint	
Base	Per-Image	Yes	Yes	1.0	1.0	74.6	**64.2**	**52.2**	**37.0**	**57.0**
(A)	Per-Class					**74.8**	63.2	51.9	36.1	56.5
(B)		No				68.5	57.2	42.3	29.1	49.3
			No			73.7	63.8	50.2	36.5	56.1
(C)				0.1	1.0	73.0	63.1	50.1	33.0	54.8
				1.0	0.1	73.1	63.8	50.3	36.4	55.9
				0.1	0.1	72.0	61.3	46.7	33.9	53.5

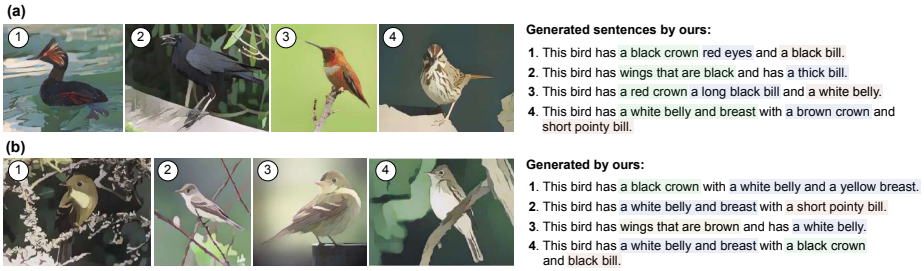

Fig. 4. (a) Textual explanations generated by our model. Our model generates plausible sentences that describe fine details about the class-discriminative attributes. We highlight such attributes with colors. (b) We further compare the generated explanations between different same-class images (i.e. Acadian Flycatcher).

see that the former is more crucial in the training of our proposed model. We provide more extensive results as a heatmap in the Figure S3. Furthermore, in Table S2, we compare the impact of embeddings from various PTEs. While we use CLIP as default, different PTEs also successfully produce domain-invariant representations.

Generated Textual Justification Quality. Next, we evaluate the quality of our generated textual justification. In Figure 4 (a), we provide sample explanations generated by our model. Note that the images shown in the figure are from unseen target domains. The model was trained in the photo, art, and paint domains and tested in the cartoon domain. Qualitatively, our textual explanation generation module accurately describes fine class-discriminative details such as "red eyes" or "white belly and breast." These are important and domain-invariant visual cues to determine their classes. For a better understanding, we highlight class discriminative attributes in the generated sentences.

Table 4. We report the quality of the generated textual explanations. We rely on standard metrics: BLEU [27], METEOR [18], CIDEr-D [40], and ROUGE_L [23].

Model	BLEU-4	METEOR	CIDEr-D	ROUGE_L
Ours w/ Joint Embedder	**48.0**	**31.7**	**40.7**	**61.8**
Ours w/o Joint Embedder	42.9	28.0	28.4	58.1

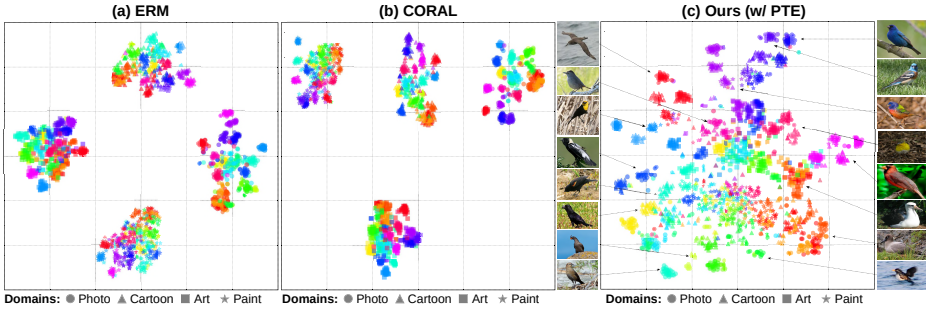

Fig. 5. Visualizations by t-SNE [24] for (a) ERM [38], (b) CORAL [36], and (c) ours. We extract latent representations from each model in the multi-source DG setting. We also provide sample images across different classes. Note that we differently color-coded each point according to its class and differently shaped to its domain.

As shown in Figure 4 (b), we further provide generated explanations for different images of Acadian Flycatcher (in the Cartoon domain). As we expected, our model describes fine details of the diverse class-discriminative attributes, which are consistent over different same-class images. This may imply that our network's visual representations are grounded by such consistent cues, which helps in providing the model generalization. For a better understanding, we highlight the same attributes with the same color.

We further quantitatively evaluate the quality of generated sentences. We use popular metrics: BLEU [27], METEOR [18], CIDEr-D [40], and ROUGE_L [23]. These metrics are widely used for the automatic evaluation of image captioning models against ground truth. The scores are averaged across three independent runs. We observe in Table 4 that our model with the Visual and Textual Joint Embedder as well as the Textual Explanation Generator obtains higher scores in all metrics than its counterpart.

Qualitative Analysis on the Latent Space We use t-SNE [24] to compute pairwise similarities of embeddings in the latent space and visualize them in a low dimensional space. In Figure 5, we provide a comparison of t-SNE visualizations of ERM [38], CORAL [36], and ours. Marker styles and colors indicate the target domain and the ground truth classes, respectively. The more generalizable model should map images belonging to the same class closely even if they

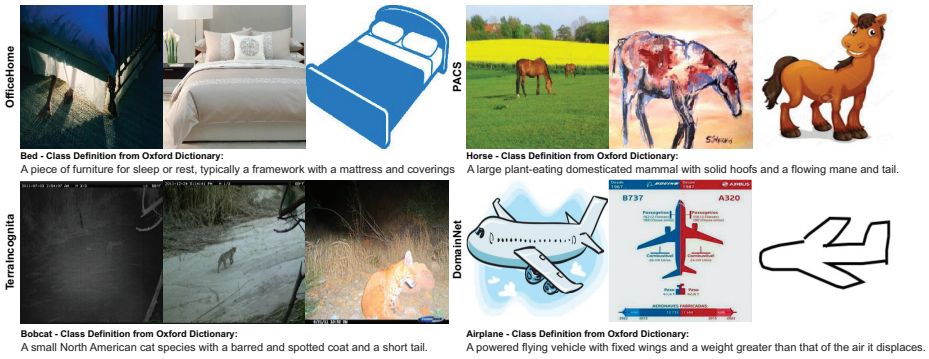

Fig. 6. We use the definition of each class as textual supervision from Oxford English Dictionary [35]. Examples from OfficeHome, PACS, TerraIncognita, and DomainNet on different domains are visualized.

are from different domains. We can observe that the baseline models produce scattered multiple clusters for each domain, which confirms that discrepancy between domains is not successfully reduced (see embeddings of the same domain are clustered closely). Ours is not the case for this. Objects from the same class (or similar attributes) but different domains tend to form a merged cluster, making latent representations close to each other in the high-dimensional space. Additionally, in Figure S4, we provide Grad-CAM [33] visualizations which highlight image regions where the model attends to classify the given object.

Large-Scale Experiments on DomainBed. To further verify the effectiveness of the proposed algotithm, we conduct large-scale experiments on DomainBed [10], which is a unified testbed useful for evaluating DG algorithms. We evaluate our algorithm on the following five multi-domain datasets (i.e. VLCS [6], PACS [19], OfficeHome [41], TerraIncognita [2], and DomainNet [28]) and compare with 14 DG algorithms (i.e. CORAL [36], SagNet [26], MLDG [20], Mixup [49], ERM [38], MTL [3], RSC [14], DANN [8], CDANN [22], VREx [17], ARM [51], IRM [1], GroupDRO [32], and MMD [21]).

The results of compared DG algorithms are excerpted from DomainBed [10]. For each algorithm, they have conducted a random search of 20 hyperparameter choices. Thus, we also conduct a random search of 20 hyperparameter choices from the following: learning rate from $5 \cdot 10^{\text{Uniform}(-5,-4)}$, weight decay from $10^{\text{Uniform}(-4,-3)}$, dropout probability from RandomChoice($[0, 0.1, 0.5]$), and a batch size from $2^{\text{Uniform}(5,5.5)}$. Other hyperparameters are fixed to the default values. We report averaged results across three independent runs.

Here, we leverage cross-modality supervision from *Per-Class* texts. Specifically, we use definitions from Oxford English Dictionary [35] to ground visual representations. In Table 5, we observe that the proposed algorithm shows state-of-the-art performance, where it ranks 1st in average performance for five multi-domain datasets. Additionally, we provide per-domain results on each dataset in

Table 5. Average out-of-distribution test accuracies on the DomainBed setting. Here we compare with 14 DG algorithms on the following five multi-domain datastes: VLCS [6], PACS [19], OfficeHome [41], TerraIncognita [2], and DomainNet [28]. The results of compared DG algorithms are excerpted from DomainBed [10]. Note that we use the validation set (from source domains) for the model selection.

Algorithm	**VLCS [6]**	**PACS [19]**	**OfficeHome [41]**	**TerraIncognita [2]**	**DomainNet [28]**	**Avg**
Ours w/ PTE	**79.0 ± 0.2**	85.1 ± 0.3	**70.1 ± 0.1**	48.0 ± 0.2	**44.1 ± 0.1**	**65.2**
CORAL [36]	78.8 ± 0.6	86.2 ± 0.3	68.7 ± 0.3	47.6 ± 1.0	41.5 ± 0.1	64.6
SagNet [26]	77.8 ± 0.5	**86.3 ± 0.2**	68.1 ± 0.1	**48.6 ± 1.0**	40.3 ± 0.1	64.2
MLDG [20]	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	47.7 ± 0.9	41.2 ± 0.1	63.6
Mixup [49]	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	39.2 ± 0.1	63.4
ERM [38]	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.1	63.3
MTL [3]	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	45.6 ± 1.2	40.6 ± 0.1	62.9
RSC [14]	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	46.6 ± 1.0	38.9 ± 0.5	62.7
DANN [8]	78.6 ± 0.4	83.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	38.3 ± 0.1	62.6
CDANN [22]	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	45.8 ± 1.6	38.3 ± 0.3	62.0
VREx [17]	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	46.4 ± 0.6	33.6 ± 2.9	61.9
ARM [51]	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	45.5 ± 0.3	35.5 ± 0.2	61.7
IRM [1]	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 2.8	61.6
GroupDRO [32]	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	43.2 ± 1.1	33.3 ± 0.2	60.7
MMD [21]	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	42.2 ± 1.6	23.4 ± 9.5	58.8

Table S3-S7. We suppose the inferior performance on some datasets is because they often do not contain enough semantic ques that can be aligned with the textual definitions. For example, it is difficult to recognize "a barred and spotted coat" from the images of the TerraIncognita dataset in Figure 6.

6 Conclusion

Towards learning more domain-invariant representations, we advocate for leveraging the cross-modality supervision. Specifically, we propose a new approach where class-discriminative natural language sentence is used during training. *Visual and Textual Joint Embedder* encourages learning visual representations that align with the pivot sentence embedding. *Textual Explanation Generator* encourages to consistently verbalize why it made a certain prediction with natural language. The experiments with the newly created CUB-DG dataset and the DomainBed benchmarks show that our model outperforms prior work under the standard DG evaluation setting. Our analysis further shows that the text modality can be successfully used to justify visual predictions as well as improve the model's representational generalization power.

Acknowledgements. This work was supported by supported by the National Research Foundation of Korea grant (NRF-2021R1C1C1009608), Basic Science Research Program (NRF-2021R1A6A1A13044830), and ICT Creative Consilience program (IITP-2022-2022-0-01819).

References

1. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
2. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
3. Blanchard, G., Deshmukh, A.A., Dogan, U., Lee, G., Scott, C.: Domain generalization by marginal transfer learning. arXiv preprint arXiv:1711.07910 (2017)
4. Cha, J., Cho, H., Lee, K., Park, S., Lee, Y., Park, S.: Domain generalization needs stochastic weight averaging for robustness on domain shifts. arXiv preprint arXiv:2102.08604 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. 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 (ICCV). pp. 1657–1664 (2013)
7. Ganea, P.A., Pickard, M.B., DeLoache, J.S.: Transfer between picture books and the real world by very young children. Journal of cognition and development **9**(1), 46–66 (2008)
8. 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)
9. Geirhos, R., et al.: ImageNet-trained CNNs are biased towards texture. In: ICLR (2019), `https://openreview.net/forum?id=Bygh9j09KX`
10. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. arXiv preprint arXiv:2007.01434 (2020)
11. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA) (2017)
12. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European conference on computer vision. pp. 3–19. Springer (2016)
13. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Grounding visual explanations. In: ECCV (2018)
14. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
15. Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.: Selfreg: Self-supervised contrastive regularization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9619–9628 (2021)
16. Kim, S., Yi, J., Kim, E., Yoon, S.: Interpretation of nlp models through input marginalization. arXiv preprint arXiv:2010.13984 (2020)
17. Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Priol, R.L., Courville, A.: Out-of-distribution generalization via risk extrapolation (rex). arXiv preprint arXiv:2003.00688 (2020)
18. Lavie, A., Agarwal, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: EMNLP (2005)
19. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)

20. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
21. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5400–5409 (2018)
22. 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)
23. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
24. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
25. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 10–18. PMLR (2013)
26. Nam, H., et al.: Reducing domain gap by reducing style bias. In: CVPR (2021)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
28. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1406–1415 (2019)
29. Pezeshki, M., Kaba, S.O., Bengio, Y., Courville, A., Precup, D., Lajoie, G.: Gradient starvation: A learning proclivity in neural networks. arXiv preprint arXiv:2011.09468 (2020)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
31. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 49–58 (2016)
32. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
34. Sinha, A., Namkoong, H., Volpi, R., Duchi, J.: Certifying some distributional robustness with principled adversarial training. ICLR (2017)
35. Stevenson, A.: Oxford dictionary of English. Oxford University Press, USA (2010)
36. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 443–450. Springer (2016)
37. Vapnik, V.: Statistical learning theory new york. NY: Wiley (1998)
38. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
40. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: ICCV (2015)

41. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5018–5027 (2017)
42. Wang, D., Devin, C., Cai, Q.Z., Yu, F., Darrell, T.: Deep object centric policies for autonomous driving. ICRA (2019)
43. Wang, X., Yu, J.: Learning to cartoonize using white-box cartoon representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8090–8099 (2020)
44. Wang, Y., Li, H., Kot, A.C.: Heterogeneous domain generalization via domain mixup. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3622–3626. IEEE (2020)
45. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010)
46. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning **8**(3), 229–256 (1992)
47. Wu, J., Mooney, R.J.: Faithful multimodal explanation for visual question answering. arXiv preprint arXiv:1809.02805 (2018)
48. Xu, M., Zhang, J., Ni, B., Li, T., Wang, C., Tian, Q., Zhang, W.: Adversarial domain adaptation with domain mixup. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6502–6509 (2020)
49. Yan, S., Song, H., Li, N., Zou, L., Ren, L.: Improve unsupervised domain adaptation with mixup training. arXiv preprint arXiv:2001.00677 (2020)
50. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. pp. 818–833. Springer (2014)
51. Zhang, M., Marklund, H., Gupta, A., Levine, S., Finn, C.: Adaptive risk minimization: A meta-learning approach for tackling group shift. arXiv preprint arXiv:2007.02931 (2020)
52. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)
53. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: International Conference on Learning Representations (2020)
54. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
55. Zou, Z., Shi, T., Qiu, S., Yuan, Y., Shi, Z.: Stylized neural painting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15689–15698 (2021)