Style-Content Metric Learning for Multidomain Remote Sensing Object Recognition

Wenda Zhao[1], Ruikai Yang[1], Yu Liu[2*], You He[2]

[1]Dalian University of Technology, Dalian, China
[2]Tsinghua University, Beijing, China
zhaowenda@dlut.edu.cn; yangruikai@mail.dlut.edu.cn; liuyu77360132@126.com; youhe_nau@163.com

Abstract

Previous remote sensing recognition approaches predominantly perform well on the training-testing dataset. However, due to large style discrepancies not only among multidomain datasets but also within a single domain, they suffer from obvious performance degradation when applied to unseen domains. In this paper, we propose a style-content metric learning framework to address the generalizable remote sensing object recognition issue. Specifically, we firstly design an inter-class dispersion metric to encourage the model to make decision based on content rather than the style, which is achieved by dispersing predictions generated from the contents of both positive sample and negative sample and the style of input image. Secondly, we propose an intra-class compactness metric to force the model to be less style-biased by compacting classifier's predictions from the content of input image and the styles of positive sample and negative sample. Lastly, we design an intra-class interaction metric to improve model's recognition accuracy by pulling in classifier's predictions obtained from the input image and positive sample. Extensive experiments on four datasets show that our style-content metric learning achieves superior generalization performance against the state-of-the-art competitors. Code and model are available at: https://github.com/wdzhao123/TSCM.

Introduction

Remote sensing object recognition (RSOR) attracts more and more attention. On one hand, RSOR possesses many essential applications, *e.g.*, urban planning, disaster assessment and maritime military security. On the other hand, there is an increase in the type and number of remote sensing data (Cheng, Zhou, and Han 2016; Xia et al. 2018; Zhang et al. 2019; Li et al. 2020c), which is collected from various satellites, such as Google Earth, Gaofen series, and Jilin series. Current works (Zhao et al. 2022c,a; Huiming and Fuxin 2021; Cui et al. 2020; Sumbul, Cinbis, and Aksoy 2019; Fang et al. 2021; Li et al. 2020d) adopt convolutional neural networks (CNNs) to improve the performance of RSOR, where the training and testing data manifest similar distributions. However, RSOR in real applications are varied from that in training data, *e.g.*, imaging condition, satellite sensor and time period. These domain gaps bring severe performance degradation, thereby limiting broader applications, as shown in Figure 1.

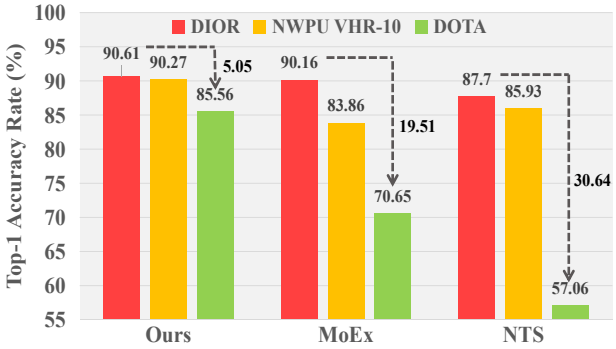
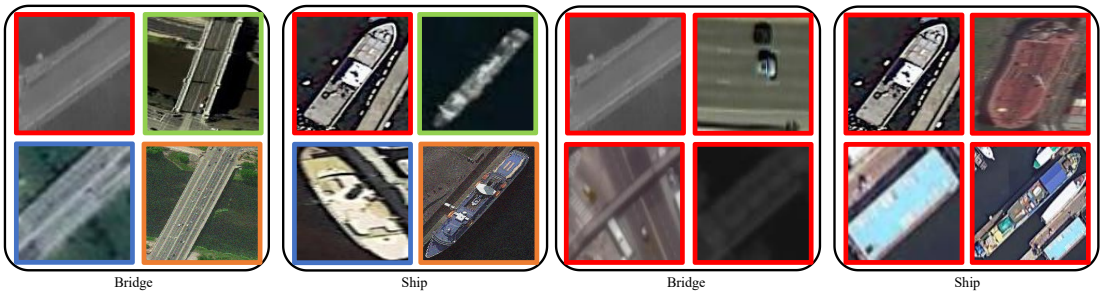


Figure 1: Generalization comparison among ours, MoEx (Li et al. 2021a) and NTS (Yang et al. 2018). The models are trained on DIOR (Li et al. 2020c) and tested on DIOR, NWPU (Cheng, Zhou, and Han 2016) and DOTA (Xia et al. 2018). MoEx and NTS perform well on the training-testing dataset, but suffer from obvious performance degradation on the other datasets. In contrast, ours achieves a better generalization.

Several works (Li et al. 2018b,c; Zhou et al. 2021) are developed to relieve the generalizing problem. A common idea is to learn a domain-invariant representation, *e.g.*, learning generalizing initial parameters via meta-learning (Li et al. 2018a), and training model with augmented data (Zhou et al. 2020b). More recently, single domain generalization methods (Qiao, Zhao, and Peng 2020; Li et al. 2021a; Nam et al. 2021) are developed. For example, works (Zhao et al. 2020; Li et al. 2021b; Wang et al. 2021b) make deep model well against domain discrepancy by training on images transformed with subtly selected augmentation types. Methods (Wang et al. 2019; Huang et al. 2020) use elaborately designed regularization strategies to train models.

Object recognition in remote sensing scenario is often performed on images with multiplex imaging conditions (*e.g.*, sensor parameters, time period and weather), which presents specific challenges for improving RSOR' general-

* Corresponding author.



(a) Style discrepancies across datasets (b) Divergent styles within one dataset

Figure 2: Visual examples of style diversity. Images with red, green, blue and orange rectangular boxes are from DOTA (Xia et al. 2018), NWPU (Cheng, Zhou, and Han 2016), DIOR (Li et al. 2020c) and HRRSD (Zhang et al. 2019), respectively. The same class objects have style discrepancies (*e.g.*, hue, contrast and blur), no matter they are from different datasets or one dataset.

ization ability. From the perspective of image style, we have the following observations. On one hand, apparent style discrepancies across remote sensing datasets exist due to different imaging conditions (see Figure 2(a)). On the other hand, even the same category within one dataset may present divergent styles because of different imaging time periods and weather (see Figure 2(b)). Unfortunately, existing works (Nam et al. 2021; Geirhos et al. 2019) have proved that CNN tends to learn more superficial features (*i.e.*, style) instead of penetrating features (*i.e.*, content), which makes CNN be sensitive to domain shift as image styles change across domains, consequently degrades deep models' generalization ability.

Based on the above analysis, we propose an end-to-end triplet style-content metric learning framework (TSCM) that reduces adverse impact of diverse styles, thus improving RSOR model's generalization ability. Specifically, we force RSOR model to learn discriminative content-biased and style-agnostic features through building style-content metric constraints, as shown in Figure 3. Given input image x, positive sample x^+ (the same class with x) and negative sample x^- (different class from x), we firstly design an inter-class dispersion metric to encourage the model to make decision based on content rather than the style, which is achieved by dispersing predictions based on the contents of x^+, x^- and the style of x. Secondly, we propose an intra-class compactness metric to force the model to be less style-biased by compacting classifier's predictions based on the content of x and the styles of x^+, x^-. However, due to subtle yet detrimental variance within categories, naively reducing style prejudice across different categories may not make the model be sophisticated adequately. Thus, we design an intra-class interaction metric to improve model's recognition accuracy, which is implemented by pulling in classifier's predictions obtained from x and x^+.

Main contributions of this paper are as follows.

- We explore the generalization of RSOR, and successfully implement an end-to-end triplet style-content metric learning framework by reducing adverse impact of diverse styles. TSCM is trained on one remote sensing dataset but performs well on the other unseen datasets.
- Inter-class dispersion metric, intra-class compactness metric and intra-class interaction metric are proposed to encourage the model to capture style-agnostic and content-biased features, thus improving generalization of RSOR.
- Extensive experiments are conducted on four widely-used remote sensing datasets, which demonstrate the superior performance of our model compared with the state-of-the-art methods.

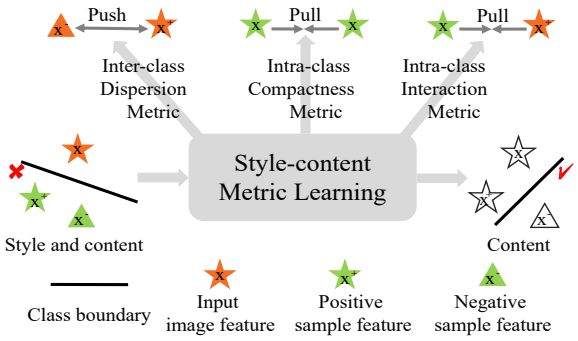


Figure 3: Motivation illustration of our style-content metric learning. Shape and color represent content and style features, respectively. Deep model tends to learn style features that makes inaccurate class boundary (✗), since the image styles change across domains. Thus, style-content metric learning is proposed to encourage the model to learn content features that makes accurate class boundary (✓).

Related Work

Remote Sensing Object Recognition. Early works generally extract features from off-the-peg CNNs, such as VGG (Simonyan and Zisserman 2014) and ResNet (He et al. 2016), for remote sensing object recognition. Recent works have improved CNNs' ability to learn discriminative features by extracting and fusing features of different layers

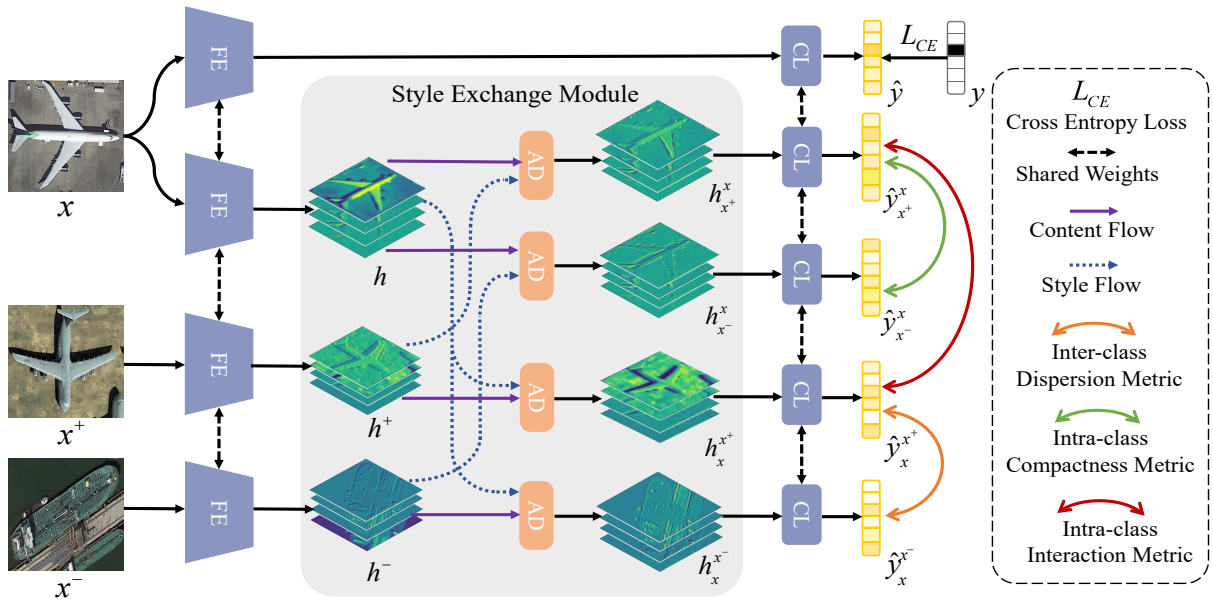


Figure 4: Illustration of the proposed style-content metric learning framework. Input image x, positive sample x^+ and negative sample x^- are fed into feature extractor FE to generate latent features h, h^+ and h^-, respectively. Then, we perform a style exchange module to obtain reorganized features $h^x_{x^+}$, $h^x_{x^-}$, $h^{x^+}_x$ and $h^{x^-}_x$ by exchanging different contents and styles, where superscript indicates the content source and subscript indicates the style source. Finally, cross entropy loss L_{CE}, inter-class dispersion metric, inter-class compactness metric and intra-class interaction metric constraints are imposed to the classifier's outputs $\hat{y}$, $\hat{y}^x_{x^+}$, $\hat{y}^x_{x^-}$, $\hat{y}^{x^+}_x$ and $\hat{y}^{x^-}_x$, leading to a precise prediction and good generalization of RSOR.

(Cui et al. 2020; Fang et al. 2021; Li et al. 2020d; Zhu et al. 2021; Han et al. 2021; Li et al. 2020b). For example, Zhao *et al.* (Zhao et al. 2022c) adopt image super-resolution to help improve weak features. Besides, work (Zhao et al. 2022a) proposes hierarchical distillation to help learn features of tailed data. Moreover, Zhao *et al.* (Zhao et al. 2022b) implement diversity consistency learning to extract features from data with limited labels. Ma *et al.* (Ma et al. 2021) propose adapative weighted intensity-hue-saturation and correlation-based attention to fuse features. Sumbul *et al.* (Sumbul, Cinbis, and Aksoy 2019) adopt attention-driven multisource deep feature representations to improve remote sensing object recognition.

Generally, existing RSOR methods perform well where training data and testing data comes from the same dataset, and suffer from performance degradation when applied to unseen dataset. Here, we explore the generalization of RSOR and propose an end-to-end triplet style-content metric learning framework, which is trained on one dataset but performs well on the other unseen datasets.

Domain Generalization Method. Domain genaralization aims to learn discriminative and general representations from source domains that support the model to perform well on unseen domains (Blanchard, Lee, and Scott 2011). Some domain genaralization methods have been proposed, such as domain distribution alignment (Li et al. 2018d; Ghifary et al. 2015), meta learning (Li et al. 2018a; Du et al. 2020b), and data generation strategies (Zhou et al. 2020a). For example, CIDDG (Li et al. 2018d) and MTAE (Ghifary et al. 2015) use domain alignment theory to minimize feature variance between source domains. The former utilize a conditional invariant adversarial network to learn domain-invariant representations, and the latter extends standard denoising model with the help of naturally occurring inter-domain variability to achieve a generalizable model. Recently, single domain generalization (Qiao, Zhao, and Peng 2020; Zhao et al. 2020; Li et al. 2021b; Wang et al. 2021b; Li et al. 2021a; Nam et al. 2021) are developed, which expects to learn generalized features from one domain. For example, MADA (Qiao, Zhao, and Peng 2020) leverages adversarial training to synthesize new domains to promote the model's generalization. PDEN (Li et al. 2021b) extends image synthesis by progressive domain expansion.

Different from the above methods, we propose novel solution of triplet style-content metric learning that encourages the model to capture style-agnostic and content-biased features, thus improving generalization of RSOR.

Deep Metric Learning. Deep metric learning (Yang, Nan, and Song 2020; Li et al. 2020a; Davis et al. 2007; Koestinger et al. 2012; Kulis et al. 2012) focuses on implementing a distance function to explicitly make the samples from the same category (positive samples) as close as possible in the latent space and disperse samples from different categories (negative samples). For example, TripletNet (Hoffer and Ailon 2015) encodes the pair of distances between positive and negative samples against the input im-

age, and imposes metric loss on the embedding. Cheng *et al.* (Cheng et al. 2018) propose discriminative objective function and metric learning regularization to effectively learn discriminative CNNs. Deng *et al.* (Deng, Jia, and Shi 2019) combine nearest neighbor and metric learning methods, and perform domain alignment through adversarial learning to encourage the target scene's embedding. Li *et al.* (Li et al. 2020a) propose a balance loss and organize training in a meta manner to learn generalized task-level distributions, which consequently learns a more discriminating metric space.

In this paper, we adopt metric learning from a new perspective that forces the model to capture style-agnostic and content-biased features. Specifically, inter-class dispersion metric, intra-class compactness metric and intra-class interaction metric are proposed to reduce adverse impact of diverse styles, thereby improving generalization of RSOR.

Style-Content Metric

Overview

Overall flowchart of our framework is illustrated in Figure 4. Given the input image x, positive sample x^+ and negative sample x^-, we firstly feed them into feature extractor FE to generate latent features. Sequentially, we perform style exchange module to obtain recombined features of different contents and styles. Then, inter-class dispersion metric, intra-class compactness metric and intra-class interaction metric are implemented after classifiers CL to force FE to capture style-agnostic and content-biased features, thus improving generalization of RSOR. Besides, the cross entropy loss is used to ensure the recognition accuracy for x. In the following sections, we present detailed descriptions of our framework.

Style Exchange Module

A remote sensing image can be interpreted as a combination of style (color and texture, *etc*.) and content (semantic feature and behavior pattern, *etc*.), and the content is the determinant for object recognition by deep model. Therefore, we decouple styles and contents within the input images, and then recombine them through the style exchange module (see Figure 4), which aims to obtain a content-biased feature extractor that generalizes well to unseen data.

Given the input image x, positive sample x^+ and negative sample x^-, we feed them into the feature extractor FE to produce latent features $h,\ h^+ \ and\ h^- \in \mathbb{R}^{C\times H\times W}$, where C indicates the number of channels, and H and W represents the height and width of a channel. For a feature $f \in \{h,\ h^+,\ h^-\}$, the channel-wise mean and standard deviation $\mu(h)$, $\sigma(h) \in \mathbb{R}^C$ are computed as style representation:

$$\mu(f) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} f_{i,j} \tag{1}$$

$$\sigma(f) = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(f_{i,j}-\mu(f)\right)^2 + \varepsilon} \tag{2}$$

where i, j indicates the spatial location, and ε is a tiny bias that avoids $\sigma(f) = 0$.

Then, we exchange different styles of x, x^+ and x^- by adaptive instance normalization (AD) (Huang and Belongie 2017), and thus obtaining recombined features as follows.

$$h^x_{x^+} = \sigma(h^+)\left(\frac{h-\mu(h)}{\sigma(h)}\right) + \mu(h^+) \tag{3}$$

$$h^x_{x^-} = \sigma(h^-)\left(\frac{h-\mu(h)}{\sigma(h)}\right) + \mu(h^-) \tag{4}$$

$$h^{x^+}_x = \sigma(h)\left(\frac{h^+-\mu(h^+)}{\sigma(h^+)}\right) + \mu(h) \tag{5}$$

$$h^{x^-}_x = \sigma(h)\left(\frac{h^--\mu(h^-)}{\sigma(h^-)}\right) + \mu(h) \tag{6}$$

where superscript in h indicates the content source and subscript in h indicates the style source.

Style-Content Metric Learning

Now, we obtain recombined features $h^x_{x^+}$, $h^x_{x^-}$, $h^{x^+}_x$ and $h^{x^-}_x$ through interchanging styles of samples x, x^+ and x^-. Sequentially, we feed recombined features to the classifier (CL) to produce object predictions $\hat{y}^x_{x^+}$, $\hat{y}^x_{x^-}$, $\hat{y}^{x^+}_x$ and $\hat{y}^{x^-}_x \in \mathbb{R}^P$, respectively, where P is the number of object categories. After that, inter-class dispersion metric, intra-class compactness metric and intra-class interaction metric are imposed to encourage FE to capture style-agnostic and content-biased features that boots generalization of RSOR.

Inter-class Dispersion Metric. Different object categories have different contents, which can be used to build the inter-class dispersion metric if they are constrained to have the same style, *i.e.*, recombined features $h^{x^+}_x$ and $h^{x^-}_x$. Specifically, we impose the inter-class dispersion metric on their corresponding $\hat{y}^{x^+}_x$ and $\hat{y}^{x^-}_x$:

$$L_{ID} = -\log\left(1-\max\left(0, \frac{1}{N}\sum_{i=0}^{N}\frac{(\hat{y}^{x^+}_x)^\top \hat{y}^{x^-}_x}{\left\|\hat{y}^{x^+}_x\right\|\left\|\hat{y}^{x^-}_x\right\|+\varepsilon}\right)\right) \tag{7}$$

where N stands for batch size, and $\top$ is a transposition operation.

By utilizing L_{ID}, we disperse features with same style from input image and positive sample's and negative sample's contents from each other as much as possible, which encourages the network to make decision based on the content rather than style. Please see Sec. 4.2 for quantitative analysis.

Intra-class Compactness Metric. Simply dispersing samples from different categories cannot contributes sufficiently to reduce adverse impact of diverse styles. Thus, we use recombined features $h^x_{x^+}$ and $h^x_{x^-}$, where the content is from x and the styles are from x^+ and x^-, and then propose the intra-class compactness metric. Specifically, we compact their output spaces $\hat{y}^x_{x^+}$ and $\hat{y}^x_{x^-}$:

$$L_{IC} = -\log\left(\frac{1}{2N}\sum_{i=0}^{N}\frac{(\hat{y}^x_{x^+})^\top \hat{y}^x_{x^-}}{\left\|\hat{y}^x_{x^+}\right\|\left\|\hat{y}^x_{x^-}\right\|+\varepsilon} + \frac{1}{2}\right) \tag{8}$$

Through compacting classifier's predictions based on the same content of x and different styles of x^+, x^-, L_{IC} can force the model to be less style-biased. Quantitative analysis is provided in Sec. 4.2.

Intra-class Interaction Metric. L_{ID} and L_{IC} use content and style variance within different categories to reduce style prejudice, respectively, which ignores model's recognition performance. Addressing this issue, we propose the intra-class interaction metric. Firstly, we calculate cross entropy loss L_{CE} to guarantee the model's recognition accuracy.

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i log(\hat{y}_i) \tag{9}$$

where y_i is the label of input image x_i, and $\hat{y}_i$ is the corresponding recognition prediction.

Then, we implement intra-class interaction metric to compact the predictions of the same category but different individuals to improve recognition accuracy. Moreover, we exchange the styles of these individuals to mitigate adverse effect of style discrepancy within the same category that further improves generalization. Thus, intra-class interaction metric L_{II} can be written as follows.

$$L_{II} = -\log\left(\frac{1}{2N}\sum_{i=0}^{N}\frac{(\hat{y}^x_{x^+})^\top \hat{y}^{x^+}_x}{\left\|\hat{y}^x_{x^+}\right\|\left\|\hat{y}^{x^+}_x\right\| + \varepsilon} + \frac{1}{2}\right) \tag{10}$$

Optimization

We employ ResNet-50 (He et al. 2016) as the model's baseline, and divide it into two parts where the former four convolution blocks serve as the feature extractor and the latter convolution blocks and fully-connected layers serve as the classifier. Especially, three fully-connected layers are added before classification and the style exchange module is appended after the feature extractor.

The proposed model is trained in an end-to-end manner, where the cross entropy loss L_{CE} jointly works with the inter-class dispersion metric L_{ID}, intra-class compactness metric L_{IC} and intra-class interaction metric L_{II} as the overall loss,

$$L_{all} = L_{CE} + \alpha L_{ID} + \beta L_{II} + \gamma L_{IC} \tag{11}$$

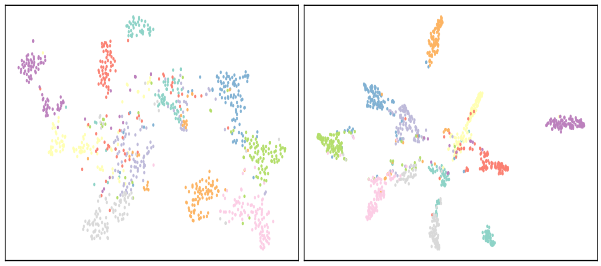
where α, β, and γ are hyper-parameters.

Experiment

Experimental Setting

Datasets. We conduct experiments using four remote sensing datasets: NWPU (Cheng, Zhou, and Han 2016), DOTA (Xia et al. 2018), HRRSD (Zhang et al. 2019) and DIOR (Li et al. 2020c). DOTA includes 26278 objects for training and 23621 instances for testing, which are collected from the Google Earth, GF-2 and JL-1 satellite. DIOR consists of 19730 objects for training and 113899 instances for testing that are collected from Google Earth. NWPU contains around 3896 instances for testing, which are from Google Earth and Vaihingen data. HRRSD contains 10647 objects for testing, which are collected from Google Earth and Baidu Map. The training datasets of DOTA and DIOR are used to train networks and the other testing datasets are adopted for evaluating models' generalization, respectively. Ten common object categories among four datasets are reserved for experiments, $i.e.$, *Airship, Ship, Storage Tank, Baseball Diamond, Tennis Court, Basketball Court, Ground Track Field, Harbor, Bridge* and *Vehicle*.

Implementation Detail. Our model is implemented by Pytorch on a PC with a NVIDIA RTX 2080 Ti GPU. We resize the image size to 256×256 pixels and set the batch size to 36. Adam (Kingma and Ba 2014) is used as the optimizer, and learning rate is 1.25e-4. We exponentially decay the learning rate of each parameter group by gamma set as 0.99 every epoch. Bias ε is set to 1e-6 in case that divisor and square root turn zero. The hyper-parameters are set as $\alpha = 0.1, \beta = 0.5$ and $\gamma = 0.5$. We initialize the backbone with the parameters of ResNet-50 pretrained on ImageNet (Russakovsky et al. 2015). The model is firstly trained for 48 epochs and then the last four fully-connected layers are further finetuned for 10 epochs to improve performance.



(a) RN50 (b) TSCM

Figure 5: Visualization of feature embedding using t-distributed stochastic neighbor embedding (t-SNE). We randomly select 100 test samples from each class from DOTA, and then map the samples to latent feature spaces through the feature extractor ($i.e.$, the first four groups of convolution layers in ResNet-50). The points with different colors denote features from different classes.

Ablation Study

The main innovation of the proposed method is the style-content metric learning, including inter-class dispersion metric, intra-class compactness metric and intra-class interaction metric, that emphasizes content while reducing impact of style. That can consequently lead to a more content-biased and style-agnostic deep model, thus improving the generalization ability of our TSCM. To verify the effectiveness of different metrics in our method, *we conduct ablation studies with model trained on DIOR and tested on NWPU, DOTA and HRRSD*. Details are as follows.

Importance of inter-class dispersion metric. Inter-class dispersion metric disperses classifier's predictions of negative sample pairs with different contents but same style. We study its effect by setting different α in Equation 11, and results are shown in Table 1. The overall accuracy increases from 81.4% to 83.1% as α increases from 0 to 0.1, and then overall accuracy decreases when α is larger. The rea-

Setting	α			
	0	0.1	0.5	1
NWPU	87.4	90.3	88.7	**92.0**
HRRSD	**75.0**	**75.0**	71.7	72.7
DOTA	83.3	**85.6**	84.3	84.5
Overall	81.4	**83.1**	81.2	82.0

Table 1: Overall top-1 accuracy (%) of our method with different α.

Setting	β			
	0	0.1	0.5	1
NWPU	86.5	88.9	**90.3**	89.0
HRRSD	72.1	70.7	**75.0**	65.5
DOTA	84.8	84.8	**85.6**	85.5
Overall	81.4	81.3	**83.1**	80.3

Table 2: Overall top-1 accuracy (%) of our method with different β.

son may be that inter-class dispersion metric can improve performance since the objects with different contents are dispersed, but higher α will make the overall loss disequilibrate. Therefore, we take $\alpha = 0.1$.

Effect of intra-class interaction metric. Intra-class interaction metric compacts classifier's predictions of positive sample pairs with same content but different styles. We study its effect by taking different hyper-parameter β in Equation 11, and the results are reported in Table 2. We find that the overall accuracy increases to 83.1% when β turns to 0.5, and then decreases to 80.3% when β turns to 1. Therefore, we set $\beta = 0.5$.

Influence of intra-class compactness metric. Intra-class compactness metric compacts classifier's predictions of the input sample with the same content but different styles, which forces the model to be less style-biased. We study its influence through setting different hyper-parameter γ in Equation 11. As shown in Table 3, the overall accuracy increases from 80.4% to 83.1% as γ increases from 0 to 0.5. However, as γ gets larger, the overall accuracy decreases to 82.8%. Accordingly, γ is set to be 0.5.

Effect of style-content metric. Lastly, we comprehensively evaluate the effect of the proposed style-content metric by comparing our model (TSCM) with the baseline ResNet-50 (RN50). As shown in Table 4, TSCM outperfroms RN50 by 11.8% on NWPU, 5.0% on HRRSD, and 3.2% on DOTA, respectively. In summary, our style-content metric can encourage the model to emphasize contents and suppress styles, consequently leads to a better generalization for RSOR.

Visualization. The proposed model can effectively capture content-biased feature embedding, thus improving generalization of RSOR. We exploit the t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton 2008) to visualize embedded feature distributions between the proposed method (TSCM) and baseline (RN50). As illustrated in Figure 5(a), although features of different

Setting	γ			
	0	0.1	0.5	1
NWPU	90.0	90.4	90.3	**91.7**
HRRSD	65.7	70.9	**75.0**	74.7
DOTA	85.4	**85.6**	**85.6**	85.1
Overall	80.4	82.0	**83.1**	82.8

Table 3: Overall top-1 accuracy (%) of our method with different γ.

Setting	NWPU		HRRSD		DOTA	
	TSCM	RN50	TSCM	RN50	TSCM	RN50
Overall	**90.3**	78.5	**75.0**	70.0	**85.6**	82.4

Table 4: Overall top-1 accuracy (%) of our method (TSCM) and baseline ResNet-50 (RN50).

classes are separated by the decision boundary, the decision boundary is quite ambiguous. For example, the feature points of RN50 are mixed and cannot be distinguished. In addition, some feature points are clustered, but the cluster centers are close. This may lead to collapse on performance if categories' number expanses greatly. On the contrary, our method maps the samples to their corresponding feature centers, and disperses them from each other, as shown in Figure 5(b). Specifically, the sample features from the same class are clustered more tightly around their centers, which benefits from the intra-class compactness metric and intra-class interaction metric that constrains features of homogeneous samples to be assembled according to their contents. Moreover, the decision boundaries in the feature space embedded by our method are more evident. The reason may be that inter-class dispersion metric pushes features of heterogeneous samples away from each other.

Comparison with State-of-the-art

We compare the proposed model TSCM with the state-of-the-art recognition methods as follows. MoEx (Li et al. 2021a) blends feature normalization with data augmentation using momentum exchange. PDEN (Li et al. 2021b) expanses domains from a single source domain to enhance model's domain generalization ability. SagNet (Nam et al. 2021) enforces deep model to learn more contents. PMG (Du et al. 2020a) performs progressive multi-granularity training with jigsaw patches. NTS (Yang et al. 2018) implements navigator-teacher-scrutinizer framework for object classification. RIDE (Wang et al. 2021a) proposes routing diverse distribution-aware experts for long-tailed object classification. For fair comparisons, we utilize the authors' original codes, and train on one source training dataset and test on the other three datasets.

Training on DIOR. We train the models using DIOR, and test them on NWPU, HRRSD and DOTA. Results are shown in Table 5. Thanks to exploration of the style-content metric, our model achieves best accuracy on both three test datasets, $e.g.$, our overall overall accuracy yields 83.1%, which brings an improvement of 21.7% to NTS, 12.2% to MoEx, 11.2%

Category	MOEX	PDEN	SagNet	PMG	NTS	RIDE	TSCM
Baseball Dd.	98.5	96.9	97.2	98.5	***99.7***	97.9	**99.5**
Basketball Ct.	91.2	***98.1***	***98.1***	82.4	92.5	79.9	**97.5**
Bridge	75.0	**86.3**	82.3	46.8	83.1	54.8	***89.5***
Gnd. Trk. Fld.	**98.8**	**98.8**	***99.4***	82.2	98.2	98.2	***99.4***
Harbor	97.8	**99.6**	98.2	96.9	***100***	***100***	***100***
Airplane	98.2	**99.3**	***99.6***	***99.6***	89.6	97.6	97.4
Ship	**75.2**	52.6	***75.8***	67.2	41.1	34.1	66.6
Vehicle	55.0	**78.6**	65.1	50.2	74.2	65.9	***88.0***
Storage Tank	77.1	71.5	**90.5**	77.3	***97.4***	85.3	89.2
Tennis Court	87.6	87.8	**88.9**	***95.2***	85.7	68.9	81.9
Overall	83.8	85.6	**88.6**	81.8	86.1	80.0	***90.3***

(a) Tested on NWPU

Category	MOEX	PDEN	SagNet	PMG	NTS	RIDE	TSCM
Baseball Dd.	***63.9***	53.4	41.5	54.4	45.6	37.0	**58.9**
Basketball Ct.	27.9	38.6	**54.8**	20.4	22.7	36.6	***59.2***
Bridge	72.5	***91.3***	83.8	72.1	69.0	61.6	**84.6**
Gnd. Trk. Fld.	81.2	54.4	***88.4***	74.1	51.5	81.1	**82.0**
Harbor	***98.5***	92.2	62.5	66.9	87.6	**95.4**	65.8
Airplane	82.9	80.8	90.3	***94.5***	52.0	**90.5**	64.9
Ship	66.6	***75.5***	**73.8**	70.8	60.6	49.3	60.0
Vehicle	9.6	30.0	54.5	36.7	42.4	**58.1**	***91.8***
Storage Tank	81.1	90.4	***97.2***	93.7	***97.2***	90.0	**96.2**
Tennis Court	85.9	***92.2***	90.2	**91.8**	90.2	69.4	81.0
Overall	66.6	70.1	**74.2**	68.1	62.0	67.4	***75.0***

(b) Tested on HRRSD

Category	MOEX	PDEN	SagNet	PMG	NTS	RIDE	TSCM
Baseball Dd.	43.0	44.9	**45.8**	41.6	42.5	40.2	***78.0***
Basketball Ct.	67.4	84.8	**87.1**	55.3	54.5	65.9	***91.7***
Bridge	25.0	***83.0***	**76.9**	64.4	27.2	59.3	75.6
Gnd. Trk. Fld.	53.5	***72.9***	70.1	67.4	59.0	66.0	**72.2**
Harbor	60.0	74.5	**76.3**	***80.6***	67.0	66.8	71.8
Airplane	77.4	**93.6**	***94.5***	91.0	62.3	81.4	89.6
Ship	**89.1**	84.8	87.5	***89.5***	59.6	70.5	86.5
Vehicle	51.8	78.7	**89.2**	85.5	48.3	78.2	***96.1***
Storage Tank	54.9	***80.2***	**72.5**	69.3	49.9	65.9	69.2
Tennis Court	93.9	96.3	94.6	***96.7***	**94.9**	89.7	95.1
Overall	70.6	82.7	**85.4**	84.5	57.1	72.6	***85.6***

(c) Tested on DOTA

Table 5: Top-1 accuracy (%) of different methods trained on DIOR and tested on NWPU, HRRSD and DOTA. The best and second best accuracy rate are marked in bold italics and bold, respectively.

to RIDE, 3.5% to PMG and 0.5% to SagNet. Detailed and representative results are analyzed as follows. (i) On NWPU, our method achieves the best accuracy on four categories and the second best on two categories. Our method outperforms all the others, especially by 10.3% to RIDE and 8.5% to PMG. This may be explained that the style information confuses the other methods, but our method makes decision based on the content. (ii) On HRRSD, our method predicts

Dataset	MOEX	PDEN	SagNet	PMG	NTS	RIDE	TSCM
NWPU	82.4	**86.0**	84.9	76.2	83.8	82.2	***86.2***
DIOR	83.3	**85.4**	83.9	81.2	83.3	83.5	***86.2***
HRRSD	**71.4**	62.6	68.5	50.0	57.3	69.7	***71.9***
Overall	82.3	**83.6**	82.7	78.5	81.1	82.3	***85.0***

Table 6: Overall top-1 accuracy (%) of different methods trained on DOTA and tested on NWPU, HRRSD and DIOR. The best and second best accuracy rate are marked in bold italics and bold, respectively.

best on two categories and achieves second highest accuracy on four categories, which makes TSCM the best on the whole. It should be noted that PDEN achieves peak on three categories, but it lags behind our method 20.6%, 27.6% and 61.8% on *Basketball Court*, *Ground Track Field* and *Vehicle*, which consequently impairs PDEN's overall performance. (iii) On DOTA, our method achieves the best result on three categories. Besides, our TSCM brings large improvement on *Vehicle*, *e.g.*, 47.8% compared to NTS and 6.9% compared to SagNet.

Training on DOTA. The models are trained on DOTA, and tested on NWPU, DIOR and HRRSD, respectively. Here, we give the overall accuracy duo to limited space, as shown in Table 6. Our model achieves best overall accuracy on both three test datasets. For example, our model obtains overall accuracy of 85.0%, which outperforms PMG by 6.5%, NTS by 3.9%, MoEx and RIDE by 2.7%, SagNet by 2.3% and PDEN by 1.4%, respectively.

Conclusion

This paper proposes a novel end-to-end multidomain remote sensing recognition network, which is achieved by implementing the style-content metric learning. We show that the style exchange module could help disentangle and combine features of different styles and contents. Further, inter-class dispersion metric can encourage the model to make decision based on content rather than the style, intra-class compactness metric can force the model to be less style-biased, and intra-class interaction metric can improve model's recognition accuracy. Therefore, the proposed style-content metric encourages the model to capture style-agnostic and content-biased features, thus improving generalization of RSOR. Extensive experiments on four datasets demonstrate the effectiveness of our approach, thereby achieving the goal of training a model on a source domain but generalizing well on the other domains.

Acknowledgements

This work is supported by National Key R&D Program of China under Grant No. 2021YFA0715202, National Natural Science Foundation of China under Grant Nos. 62176038 and 62022092, and Science and Technology Star of Dalian under Grant No. 2021RQ054.

References

Blanchard, G.; Lee, G.; and Scott, C. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24.

Cheng, G.; Yang, C.; Yao, X.; Guo, L.; and Han, J. 2018. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE transactions on geoscience and remote sensing*, 56(5): 2811–2821.

Cheng, G.; Zhou, P.; and Han, J. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12): 7405–7415.

Cui, Z.; Guo, W.; Zhang, Z.; Chen, H.; and Yu, W. 2020. Ellipse-FCN: Oil Tanks Detection from Remote Sensing Images with Fully Convolution Network. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, 2855–2858. IEEE.

Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, 209–216.

Deng, B.; Jia, S.; and Shi, D. 2019. Deep metric learning-based feature embedding for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2): 1422–1435.

Du, R.; Chang, D.; Bhunia, A. K.; Xie, J.; Ma, Z.; Song, Y.-Z.; and Guo, J. 2020a. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, 153–168. Springer.

Du, Y.; Xu, J.; Xiong, H.; Qiu, Q.; Zhen, X.; Snoek, C. G.; and Shao, L. 2020b. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, 200–216. Springer.

Fang, W.; Sun, Y.; Ji, R.; Wan, W.; and Ma, L. 2021. Recognizing Global Dams From High-Resolution Remotely Sensed Images Using Convolutional Neural Networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 6363–6371.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Ghifary, M.; Kleijn, W. B.; Zhang, M.; and Balduzzi, D. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, 2551–2559.

Han, Y.; Yang, X.; Pu, T.; and Peng, Z. 2021. Fine-Grained Recognition for Oriented Ship Against Complex Scenes in Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hoffer, E.; and Ailon, N. 2015. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, 84–92. Springer.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 124–140. Springer.

Huiming, Y.; and Fuxin, X. 2021. A remote sensing image target recognition method based on improved Mask-RCNN model. In *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 436–439. IEEE.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, 2288–2295. IEEE.

Kulis, B.; et al. 2012. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4): 287–364.

Li, B.; Wu, F.; Lim, S.-N.; Belongie, S.; and Weinberger, K. Q. 2021a. On feature normalization and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12383–12392.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Li, H.; Cui, Z.; Zhu, Z.; Chen, L.; Zhu, J.; Huang, H.; and Tao, C. 2020a. RS-MetaNet: Deep Metametric Learning for Few-Shot Remote Sensing Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8): 6983–6994.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018b. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5400–5409.

Li, J.; Tian, J.; Gao, P.; and Li, L. 2020b. Ship Detection and Fine-Grained Recognition in Large-Format Remote Sensing Images Based on Convolutional Neural Network. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, 2859–2862. IEEE.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020c. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159: 296–307.

Li, L.; Gao, K.; Cao, J.; Huang, Z.; Weng, Y.; Mi, X.; Yu, Z.; Li, X.; and Xia, B. 2021b. Progressive Domain Expansion Network for Single Domain Generalization. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 224–233.

Li, L.; Wang, C.; Zhang, H.; and Zhang, B. 2020d. SAR image ship object generation and classification with improved residual conditional generative adversarial network. *IEEE Geoscience and Remote Sensing Letters*.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018c. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 624–639.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018d. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 624–639.

Ma, W.; Shen, J.; Zhu, H.; Zhang, J.; Zhao, J.; Hou, B.; and Jiao, L. 2021. A Novel Adaptive Hybrid Fusion Network for Multiresolution Remote Sensing Images Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–17.

Nam, H.; Lee, H.; Park, J.; Yoon, W.; and Yoo, D. 2021. Reducing Domain Gap by Reducing Style Bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8690–8699.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12556–12565.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sumbul, G.; Cinbis, R. G.; and Aksoy, S. 2019. Multisource region attention network for fine-grained object recognition in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(7): 4929–4937.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, H.; He, Z.; Lipton, Z. C.; and Xing, E. P. 2019. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. 2021a. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. In *International Conference on Learning Representations*.

Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; and Baktashmotlagh, M. 2021b. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 834–843.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974–3983.

Yang, X.; Nan, X.; and Song, B. 2020. D2N4: A discriminative deep nearest neighbor neural network for few-shot space target recognition. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5): 3667–3676.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 420–435.

Zhang, Y.; Yuan, Y.; Feng, Y.; and Lu, X. 2019. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8): 5535–5548.

Zhao, L.; Liu, T.; Peng, X.; and Metaxas, D. 2020. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *arXiv preprint arXiv:2010.08001*.

Zhao, W.; Liu, J.; Liu, Y.; Zhao, F.; He, Y.; and Lu, H. 2022a. Teaching Teachers First and Then Student: Hierarchical Distillation to Improve Long-Tailed Object Recognition in Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12.

Zhao, W.; Tong, T.; Wang, H.; Zhao, F.; He, Y.; and Lu, H. 2022b. Diversity Consistency Learning for Remote-Sensing Object Recognition With Limited Labels. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–10.

Zhao, W.; Tong, T.; Yao, L.; Liu, Y.; Xu, C.; He, Y.; and Lu, H. 2022c. Feature Balance for Fine-Grained Object Classification in Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2021. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020a. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13025–13032.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020b. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, 561–578. Springer.

Zhu, S.; Luo, F.; Du, B.; and Zhang, L. 2021. Adversarial Fine-Grained Adaptation Network for Cross-Scene Classification. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2369–2372. IEEE.