Histopathological Image Analysis with Style-Augmented Feature Domain Mixing for Improved Generalization

Vaibhav Khamankar, Sutanu Bera(✉), Saumik Bhattacharya, Debashis Sen, and Prabir Kumar Biswas

Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India
sutanu.bera@iitkgp.ac.in

Abstract. Histopathological images are essential for medical diagnosis and treatment planning, but interpreting them accurately using machine learning can be challenging due to variations in tissue preparation, staining and imaging protocols. Domain generalization aims to address such limitations by enabling the learning models to generalize to new datasets or populations. Style transfer-based data augmentation is an emerging technique that can be used to improve the generalizability of machine learning models for histopathological images. However, existing style transfer-based methods can be computationally expensive, and they rely on artistic styles, which may negatively impact model accuracy. In this study, we propose a feature domain style mixing technique that uses adaptive instance normalization to estimate style-mixed versions of image features. We compare our proposed method with existing style transfer-based data augmentation methods and found that it performs similarly or better, despite requiring lower computation. Our results demonstrate the potential of feature domain statistics mixing in the generalization of learning models for histopathological image analysis.

Keywords: Domain Shift · Domain Generalization · Mitotic Figure · Style Mixing · Feature Domain Augmentation · Histopathological Image

1 Introduction

Histopathological images play a critical role in medical diagnosis and treatment planning, allowing healthcare providers to visualize the microscopic structures of tissues and organs. However, accurately interpreting these images can be challenging due to variations in tissue preparation, staining and imaging protocols.

V. Khamankar and S. Bera—Contributed equally to this work and share the first authorship

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-47401-9_28.

M. E. Celebi et al. (Eds.): MICCAI 2023 Workshops, LNCS 14393, pp. 285–294, 2023.
https://doi.org/10.1007/978-3-031-47401-9_28

These variations can result in significant differences in image quality, tissue morphology and staining intensity, making it difficult to develop machine learning models for analysis that generalize well to new datasets or populations. Domain generalization is a field of machine learning that seeks to address this limitation by enabling models to generalize to new domains or datasets. In the context of histopathological images, domain generalization methods aim to improve the generalizability of machine learning models by reducing the effects of dataset bias and increasing the robustness of the model to variations in tissue preparation, staining, and imaging protocols. Recently, there has been a growing interest in using style transfer-based data augmentation for learning visual representations that are independent of specific domains for histopathological images viz., [7,13,18]. This technique involves transferring the style or texture of one image to another while maintaining the original content. By generating new images with different styles or textures, this technique can be used to augment the training data and improve the model's generalization performance [18]. Although the style transfer based method achieves good results in domain generalization for histopathological images, it takes a considerable amount of time to generate the augmented data. Further, the collinearity between the various artistic styles used for the style transfer may have a negative impact on the model's accuracy. Unlike the existing methods, in this work, we propose to apply feature domain style mixing for the style transfer. Specifically, we use adaptive instance normalization [6] to mix the feature statistics of the different images to generate a style-augmented version of an image. Feature statistics mixing helps to save a lot of time and computation power as data augmentation is not required, and the dependency on the artistic style is also alleviated. We compare the proposed method with the current state-of-the-art style transfer-based data augmentation methods, on two image classification tasks and one object detection task. We find that the proposed method performs similarly or better than the image domain mixing-based methods, despite having low computation requirements.

2 Related Work

In the field of digital pathology, researchers have developed several deep learning approaches to address challenges related to domain generalization such as normalization and style transfer. One example is StainNet [7], which is designed for stain normalization in digital pathology images. StainNet removes variations in tissue staining across different samples, making it easier to compare and analyze images in a consistent manner. Another approach, STRAP [18], uses a deep neural network to extract features from histopathology images and proposes a style transfer augmentation technique to reduce the domain-specific information in these features. This technique generates a new set of images that have the same content as the original images but in different styles. Domain Adversarial RetinaNet [16], a modified version of the RetinaNet object detection model, has been developed that includes domain adversarial training. The idea is to train in both source and target domain data to address domain generalization challenges.

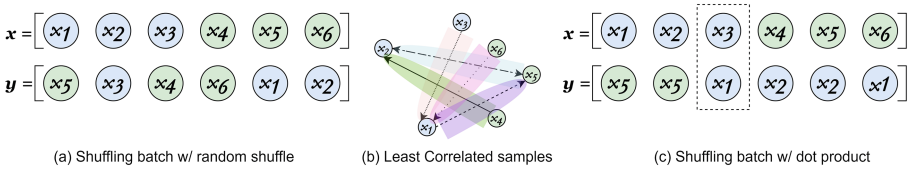


Fig. 1. A graphical illustration of FuseStyle. The shaded areas in (b) are the simulated points for augmentation. The domain label of each sample is colour-coded. There can be cases where the dot product (correlation) is the least within the domain as highlighted in the dotted rectangle in (c).

3 Proposed Method

3.1 Background

Huang et al. [6] introduced Adaptive Instance Normalization (AdaIN) for style transfer based on Instance normalization [15]. AdaIN aims to align the means and variances of instances of the content features (c) with those of the style features (s). It computes the mean ($\mu(s)$) and variance ($\sigma(s)$) parameters from instances of the style input and achieves the style transfer as $AdaIN(c) = \sigma(s)\frac{x-\mu(c)}{\sigma(c)} + \mu(s)$, where $\mu(c)$ and $\sigma(c)$ are respectively the corresponding instance mean and standard deviation of a given content feature tensor. The above parameter adaption allows for arbitrary style transfer, enabling the mixing of the content and style features in a way that produces a new output with the parameters of the style.

3.2 FuseStyle: Proposed Feature Domain Style Mixing

Our feature domain style mixing approach, FuseStyle, is inspired by AdaIN. FuseStyle avoids the use of an image generating network that is usually associated with style transfer based domain generalization. Instead, it regularizes the training of the neural network at hand (for performing a required task) by perturbing the style information of the training instances. It can be easily implemented as a plug-and-play module inserted between the layers of the neural network. So, the need to explicitly create a new style image does not arise.

FuseStyle, depicted in Fig. 1, combines the feature statistics of two instances from the same /different domains as a convex sum using random weights to simulate new styles. As shown in Fig. 1, for an input training batch, x, a reference batch y is generated by shuffling x across the batch dimension. We then compute the means (μ) and variances (σ) of the corresponding instances in x and y, and use them to compute the combined feature statistics as:

$$\gamma_i = \lambda_i \sigma(x_i) + (1 - \lambda_i)\sigma(y_i), \qquad \beta_i = \lambda_i \mu(x_i) + (1 - \lambda_i)\mu(y_i) \tag{1}$$

where i denotes the i^{th} instance and $\lambda_i \sim \text{Beta}(\alpha, \alpha)$ is computed from a Beta distribution having both its shape parameters as α. A style-modified training instance $\tilde{x_i}$ is then computed as:

$$\tilde{x_i} = \gamma_i \frac{x_i - \mu(x_i)}{\sigma(x_i)} + \beta_i \tag{2}$$

where the batch size of $\tilde{x}$ is the same as that of x and y. x is then randomly (binomial-$B(0, .5)$) replaced by $\tilde{x}$ as the training batch for domain generalization.

Generating the reference batch y is crucial for achieving better generalization to unseen domains. While previous studies [20] have used a random sample selection method for creating the reference batch, a recent study [18] in histopathological image domain generalization has shown that mixing medically irrelevant images, such as artistic paintings, with whole slide images (WSI) results in improved performance. This suggests that using the least correlated image in the reference batch could result in a better generalization than using a meaningful stylized image. With this motivation, we propose a new method of generating the reference batch that allows the mixing of the features of a sample with the features of another sample in the batch that is least correlated to the former. This method has inherent advantages over existing methods. For example, when we combine the parameters of two furthest samples linearly, the interpolated parameter values are more likely to represent a simulated sample that is far from the both the original samples than when we combine two close samples (which may happen during random reference batch generation). This allows us to explore more regions in the feature space and simulate a wider variety of augmented domains, as illustrated in Fig. 1b. Consider that FuseStyle is applied between a layer, f_l, and f_{l+1}, and the output feature of the layer f_l is $z_l \in \mathbb{R}^{B \times C \times W \times H}$ (B - batch dimension). Then, the correlation ($\rho \in \mathbb{R}^{B \times B}$) between different samples of the current batch can be computed by:

$$\rho = \hat{z}_l \odot \hat{z}_l^T \tag{3}$$

where $\odot$ represents the matrix multiplication, $\hat{z}_l \in \mathbb{R}^{B \times CWH}$ is the vectorized version of the z_l and T represents the transpose operation. Next, we set i^{th} sample of the reference batch, that is, y_i to be x_j, where $j = \arg\min_j \rho_i$, and $\rho_i \in \mathbb{R}^B$ is the i^{th} row of the matrix ρ. Then, the i^{th} sample of the batch x is mixed with i^{th} sample of the batch y as mentioned in Eq.(2) to get $\tilde{x_i}$. We set α of the Beta distribution to 0.3 to generate all the results reported in this paper. During the learning phase of the neural network model, the probability of using the FuseStyle method is set at 0.5, but it is not applied during the test phase.

Table 1. Comparison of FuseStyle with SoTA methods on Camelyon17-WILDS.

Methods	STRAP	FuseStyle	LISA	Fish	ERM	V-REx	DomainMix	IB-IRM	GroupDRO
Test Accuracy	93.7%	90.49%	77.1%	74.7%	70.3%	71.5%	69.7%	68.9%	68.4%

Table 2. Classification using FuseStyle and STRAP on $MIDOG'21$ Dataset.

	STRAP Test Accuracy(%)			FuseStyle Test Accuracy(%)		
Networks{Train,Test}	XR	S360	CS	XR	S360	CS
{S360+CS,XR}	67.33	87.99	84.67	**77.56**	**91.07**	**90.46**
{XR+CS, S360}	88.35	**76.78**	91.14	**90.06**	75.16	**92.16**
{XR+S360,CS}	**88.92**	**92.70**	**74.28**	86.65	88.96	74.10

4 Experimental Details

4.1 Datasets and Task

In our study, we compared our proposed method with the recent state-of-the-art histopathological domain generalization using two datasets. 1. The MIDOG'21 Challenge dataset [3] consisted of 200 samples of human breast cancer tissue stained with Haematoxylin and Eosin (H&E). Four scanning systems were used to digitize the samples: Leica GT450, Aperio CS2 (CS), Hamamatsu XR (XR), and Hamamatsu S360 (S360), resulting in 50 WSIs from each system. 2. The Camelyon17-WILDS [9] dataset comprised 1,000 histopathology images distributed across six domains, representing different combinations of medical centres and scanners. In our study, we focused on three tasks: classification between mitotic figures and non-mitotic figures using the MIDOG'21 dataset, tumour classification using the Camelyon17 dataset, and detection of mitotic and non-mitotic figures using the MIDOG'21 dataset. TFor the mitotic figure detection task, the details regarding dataset preparation can be found in the supplementary material of our study. For the Camelyon17 WILDS dataset, we used the default settings and train test split as given on the challenge website. For the classification task on MIDOG'21, we cropped patches of size 64×64 around the mitotic and non-mitotic figure, and we then performed an 80–20 train-test split on the cropped patches keeping the patches from each domain separate.

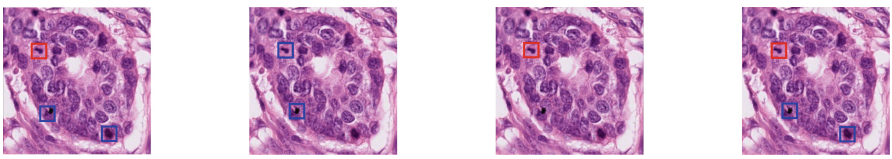
4.2 Model Architecture, Training and Methods

Classification: Here, we employ ResNet50 [5] CNN architecture and integrate FuseStyle after layers 1 and 4 of the network for 15 epochs. We use Binary Cross Entropy (BCE) Loss for training, while Adam Optimizer [8] with a learning rate of 1e-4 is utilized. To facilitate smooth training, a scheduler is used, that is, when no improvement is seen during model training after 2 epochs, the learning rate is reduced by a factor of 0.01. The batch size is set to 256 for both Camelyon17-WILDS [9] and MIDOG'21 Challenge datasets [3]. Recent studies on style transfer indicate that style information can be modified by altering the

instance-level feature statistics in the lower layers of a Convolutional Neural Network (CNN) while preserving the image's semantic content representation [4,6], and hence, we consider layers 1 and 4 of the ResNet to use FuseStyle.

Mitotic Figure Detection: For mitotic figure detection, we utilize RetinaNet [11] with ResNet50 as the backbone architecture and incorporate FuseStyle on layers 1 and 4 of the backbone. We use Focal Loss and train the network for 100 epochs on the MIDOG'21 Challenge dataset with a batch size of 6. Adam Optimizer [8] is used with a learning rate of 1e-4. We use the adaptive learning rate decay scheduler, that is, when no improvement is seen in model training after two epochs, the learning rate is reduced by factor of 0.1 for stable training.

Methods: To assess the effectiveness of our proposed approach, we compare it to eight state-of-the-art domain generalization methods, namely STRAP [18], LISA [19], Fish [14], ERM [9], V-REx [10], DomainMix [17], IB-IRM [1], and GroupDRO [12] for classification task on the Camelyon17-WILDS dataset, where we evaluate the classification accuracy. The best existing approach STRAP [18] based on the performance data on Camelyon17-WILDS dataset is used further for comparison with the proposed approach on the MIDOG'21 Challenge dataset, where both classification and mitotic figure detection are considered. We implemented the networks using the PyTorch library in Python and utilized a GeForce GTX 2080Ti GPU for efficient processing.



(a) GT (b) RetinaNet (c) STRAP (d) FuseStyle

Fig. 2. Mitotic figure detection by different methods in S360 image with model trained on XR & CS, where Red box→Mitotic and Blue box→Non-Mitotic. (Color figure online)

5 Results and Discussion

Classification Task Results: We evaluate the state-of-the-art (SOTA) methods along with ours based on their classification performance in out-of-distribution domains, and we use accuracy as the performance metric. Our approach is first compared to the other methods in Table 1, where the Camelyon17 dataset is used for both training and testing (out-of-distribution). The results

presented in Table 1 demonstrate that our approach outperforms all the methods except STRAP [18].

One should note regarding STRAP that its performance heavily relies on the generated stylized dataset used for training. The time required to generate the stylized data for the Camelyon17- WILDS is around 300 h in our set up and for the MIDOG'21 Challenge dataset, it is around 75 h. On the other hand, there is no data generation involved with our FuseStyle. Further, the main operation in FuseStyle is a dot product, which is computationally cheap, and the complexity of our feature mixing strategy is negligible compared to existing augmentation techniques.

Due to the substantial dependence of STRAP on the generated stylized augmentation, careful selection of style images for every dataset becomes fundamental to reproduce its similar performance on different datasets. Therefore, to further investigate the performance of FuseStyle and STRAP, we conduct a classification experiment on the MIDOG'21 Challenge Dataset, the results of which are presented in Table 2. As seen, if the network is trained on S360 and CS, and tested on XR, there is a 10.23% advantage in test accuracy for FuseStyle over STRAP. Furthermore, the accuracy improves by 5.79% and 3.08% for S360 and CS, which are the seen domains, respectively. In the other cases of Table 2, the approaches outperform each other almost equal number of times, but most importantly, the differences in their accuracies are relatively low. This shows that FuseStyle is at par with STRAP in these cases in spite of it being significantly less complex. We also infer from the table that FuseStyle produces consistent performance irrespective of the training and testing domains.

Mitotic Figure Detection Task Results: We conduct an experiment on this task using three different models: Our FuseStyle, STRAP and RetinaNet [11]. All these models use ResNet50 as their backbone architecture, but Retinanet does not involve any domain generalization. We provide Precision, Recall and F1 score as the performance metrics of detection in Table 3. Here the models are trained using training data from XR and CS scanners. As a result, the images from S360 represent an out-of-distribution scenario. As can be seen, FuseStyle outperforms both STRAP [18] and RetinaNet in most cases in terms of F1 score that incorporates both precision and recall. FuseStyle's superiority over RetinaNet demonstrates the usefulness of our way of domain generalization.

Table 3. Mitotic Figure Detection Analysis on $MIDOG'21$ Challenge Dataset.

Network	Precision			Recall			F1 Score		
	XR	S360	CS	XR	S360	CS	XR	S360	CS
RetinaNet	0.91	0.93	0.93	0.76	0.3	0.76	0.83	0.45	0.84
STRAP	0.85	0.91	0.88	0.88	0.70	0.95	0.87	0.79	0.92
FuseStyle	0.82	0.92	0.90	0.92	0.76	0.90	0.87	0.83	0.90

Table 4. Objective evaluation on $MIDOG'21$ Challenge Dataset.

Network	Methods	XR	S360	CS	Time/epoch (sec)
Train: S360 & CS Test: XR	M1	76.42	89.45	88.76	08
	RA	**77.56**	**91.07**	**90.46**	08
	M2	67.05	88.64	90.29	08
	M3	74.43	87.18	88.76	111
Train: XR & CS Test: S360	M1	84.09	71.59	85.18	08
	RA	90.06	75.16	**92.16**	08
	M2	**90.34**	75.65	91.65	08
	M3	83.07	**78.90**	89.95	111
Train: XR & S360 Test: CS	M1	**88.64**	89.45	72.40	08
	RA	86.65	88.96	74.10	08
	M2	87.78	**90.91**	76.32	08
	M3	86.34	87.34	**78.02**	111

Table 5. Objective evaluation on $MIDOG'22$ Challenge Dataset.

Network	Methods	XR	S360	CS	Time/epoch (sec)
Train: S360 & CS Test: XR	M1	75.10	81.16	80.58	28
	RA	74.27	**83.59**	79.74	28
	M2	76.78	75.68	68.80	28
	M3	**78.74**	83.28	**81.24**	331
Train: XR & CS Test: S360	M1	77.06	80.24	78.89	28
	RA	80.84	80.24	**81.62**	28
	M2	75.94	**81.76**	81.53	28
	M3	**81.26**	**81.76**	79.26	331
Train: XR & S360 Test: CS	M1	81.12	81.46	82.28	28
	RA	**81.34**	**84.19**	**82.75**	28
	M2	73.71	76.60	73.42	28
	M3	80.14	83.89	74.08	331

A visual result of mitotic figure detection using FuseStyle, STRAP and RetinaNet is shown in Fig. 2a along with the ground truth. As we can see from the figure, the use of FuseStyle, unlike the use of the other two, results in accurate detection and classification of all mitotic and non-mitotic figures present. While the use of RetinaNet results in an unsuccessful classification of a mitotic figure, the use of STRAP results in detection failure.

Design Analysis of Our Approach: Our investigation has revealed that combining distant features can lead to the extraction of domain-invariant features. To achieve this, we had proposed using the dot product method, but other techniques for generating a reference batch exist. To explore this further, we conduct an empirical investigation using four different methods: M1: Mixing with Random Shuffle, Reference Approach (RA): Mixing with Least Dot Product (FuseStyle), M2: Mixing with Maximum Euclidean Distance, and M3: Mixing with Maximum KL Divergence. We study the Euclidean distance based approach and also experiment with an advanced approach based on KL divergence. To evaluate the robustness of the proposed approach, we train the ResNet50 model on two scanner datasets and tested it on the third scanner. The comparison of the results obtained from the study are presented in Tables 4 & 5. The comparison is based on the test accuracy (in percentage) of different scanners and the time required for training. The obtained results reveal the effectiveness of the proposed approach of sample selection for mixing. The detailed analysis of the findings is provided in the table, demonstrating the superiority of the proposed method over the other methods.

Based on the results presented in Table 4, it can be observed that the Dot Product method is the most consistent in terms of network performance across

different domains. In contrast, the Random Shuffle method (M1) fails to perform well in the second case, and the Euclidean Distance method (M2) fails in the first case for the out-of-distribution domain. The KL Divergence method (M3) does not perform well in the in-distribution domain, as observed in the second case for the XR scanner, and it also requires a significantly longer computational time compared to the other methods. Therefore, the experimental studies suggest that FuseStyle (RA) provides the most consistent results as well as takes less time compared to KL divergence method (M3) on the MIDOG'21 Challenge dataset. For further analysis of our method, we conduct additional experiments on the MIDOG'22 Challenge dataset [2] as shown in Table 5, using the same reference batch generation methods. The results demonstrate that the FuseStyle (RA) performs well for both in-distribution and out-of-distribution domains.

6 Conclusion

We present, FuseStyle, a novel method that computes generalized features by mixing them in the feature space to address domain shift issues related to histopathological images. It uses a new approach of feature mixing based on correlation computation. FuseStyle has lower computational requirements, with dot product being the main operation in it. We have shown that the performance of our method in classification and detection tasks is at par or better than the state-of-the-art on various datasets. We also find from experimental results that the proposed feature-mixing method has strong domain generalization capabilities. In summary, our method is simple, effective and consistent, and it has the potential to enhance the out-of-distribution performance of any existing machine learning method.

References

1. Ahuja, K., et al.: Invariance principle meets information bottleneck for out-of-distribution generalization. Adv. Neural. Inf. Process. Syst. **34**, 3438–3450 (2021)
2. Aubreville, M., Bertram, C., Breininger, K., Jabari, S., Stathonikos, N., Veta, M.: Mitosis domain generalization challenge 2022 (2022). https://doi.org/10.5281/zenodo.6362337
3. Aubreville, M., et al.: Mitosis domain generalization in histopathology images-the midog challenge. Med. Image Anal. **84**, 102699 (2023)
4. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
7. Kang, H., et al.: Stainnet: a fast and robust stain normalization network. Front. Med. **8**, 746307 (2021)

8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Koh, P.W., et al.: WILDS: A benchmark of in-the-wild distribution shifts. In: International Conference on Machine Learning (ICML) (2021)
10. Krueger, D., et al.: Out-of-distribution generalization via risk extrapolation (rex). In: International Conference on Machine Learning, pp. 5815–5826. PMLR (2021)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
12. 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)
13. Shaban, M.T., Baur, C., Navab, N., Albarqouni, S.: Staingan: Stain style transfer for digital histological images. In: 2019 IEEE 16th International Symposium On Biomedical Imaging (ISBI 2019), pp. 953–956. IEEE (2019)
14. Shi, Y., Seely, J., Torr, P.H., Siddharth, N., Hannun, A., Usunier, N., Synnaeve, G.: Gradient matching for domain generalization. arXiv preprint arXiv:2104.09937 (2021)
15. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
16. Wilm, F., Marzahl, C., Breininger, K., Aubreville, M.: Domain adversarial retinanet as a reference algorithm for the mitosis domain generalization challenge. In: Aubreville, M., Zimmerer, D., Heinrich, M. (eds.) MICCAI 2021. LNCS, vol. 13166, pp. 5–13. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97281-3_1
17. Xu, M., et al.: Adversarial domain adaptation with domain mixup. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6502–6509 (2020)
18. Yamashita, R., Long, J., Banda, S., Shen, J., Rubin, D.L.: Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. IEEE Trans. Med. Imaging **40**(12), 3945–3954 (2021). https://doi.org/10.1109/TMI.2021.3101985
19. Yao, H., et al.: Improving out-of-distribution robustness via selective augmentation. In: International Conference on Machine Learning, pp. 25407–25437. PMLR (2022)
20. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008 (2021)