ConStyX: Content Style Augmentation for Generalizable Medical Image Segmentation

Xi Chen[1,2,†], Zhiqiang Shen[1,2,†], Peng Cao[1,2,3(✉)], Jinzhu Yang[1,2,3], and Osmar R. Zaiane[4]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2] Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China
[3] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang, China
caopeng@mail.neu.edu.cn
[4] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada

Abstract. Medical images are usually collected from multiple domains, leading to domain shifts that impair the performance of medical image segmentation models. Domain Generalization (DG) aims to address this issue by training a robust model with strong generalizability. Recently, numerous domain randomization-based DG methods have been proposed. However, these methods suffer from the following limitations: 1) constrained efficiency of domain randomization due to their exclusive dependence on image style perturbation, and 2) neglect of the adverse effects of over-augmented images on model training. To address these issues, we propose a novel domain randomization-based DG method, called content style augmentation (ConStyX), for generalizable medical image segmentation. Specifically, ConStyX 1) augments the content and style of training data, allowing the augmented training data to better cover a wider range of data domains, and 2) leverages well-augmented features while mitigating the negative effects of over-augmented features during model training. Extensive experiments across multiple domains demonstrate that our ConStyX achieves superior generalization performance. The code is available at *https://github.com/jwxsp1/ConStyX*.

Keywords: Domain generalization · Domain randomization · Medical image segmentation · Deep features.

1 Introduction

Medical image segmentation is an important task in computer-aided diagnosis and treatment. In recent years, this field has witnessed significant advancements, attributed to the progress of deep learning [1]. However, learned segmentation models encounter significant performance drops when the training and test sets

[1] † Xi Chen and Zhiqiang Shen contributed equally to this work.

are sampled from different distributions, where domain shifts arise due to variations in acquisition processes and patient populations [5,19]. Domain generalization (DG) has been proposed to improve the generalizability of segmentation models, with the setting that a model is trained using data from single or multiple domains and tested on unseen domains [25,17].

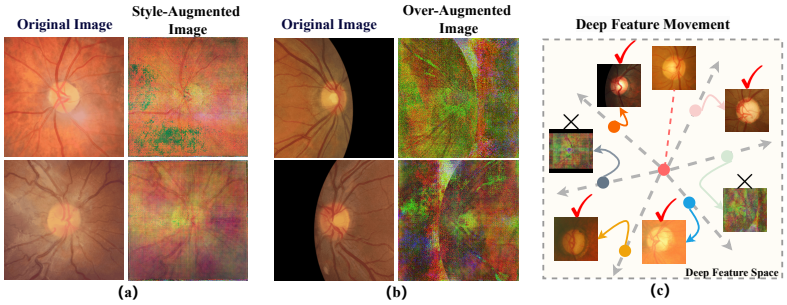


Fig. 1. Visualization of original and augmented images. (a) Original images (left) and style-augmented images (right). (b) Original images (left) and over-augmented images (right). (c) In the deep feature space, when the feature vector of a sample moves along a specific direction (gray dashed arrow), the resulting feature vector corresponds to an augmented image (✓: well-augmented images or ×: over-augmented images).

Existing DG methods can be categorized into decoupling-based [4,13,7] and domain randomization-based methods [24,26,8]. Decoupling-based methods attempt to extract domain-invariant features from data by normalizing the features or constructing dedicated modules within the model. However, these approaches may compromise the semantic information within the features, as there is no guarantee that only domain-specific features are eliminated, leading to a degradation in the model's discriminative capability. In contrast, domain randomization-based DG methods aim to simulate unseen domains using source domain data. This line of approaches can be divided into two branches: 1) image perturbation via image reconstruction [2,24] or generation [8], and 2) feature perturbation by transferring statistical information [26,10,21,23]. However, these methods only augment image style [Fig. 1(a)], limiting the diversity of augmented data distributions. Moreover, due to the uncontrollability of the perturbation process, it is possible to produce over-augmented images with corrupted semantic information and unreal image appearance [Fig. 1(b)], which degrade the model training.

To address these issues, we propose a novel content style augmentation method (ConStyX) for generalizable medical image segmentation. Our method is built upon the following assumption: moving a deep feature along a certain direction produces a new feature that corresponds to another sample of the same class but contains different content[5] and style information [16,18] [Fig. 1(c)]. Specif-

[5] "Content" refers to substructures within a specific class region. For example, different optic discs (class) may contain distinct optic vessels (content) in fundus images.

ically, ConStyX consists of two components: 1) Deep Feature Augmentation algorithm (DFA) conducts content and style augmentation, ensuring the augmented data covers a wide scope for unseen target domains and 2) Augmented Feature Utilization strategy (AFU) quantifies the contributions of augmented features to the model training, thereby mitigating the negative impacts of over-augmented features while fully leveraging the well-augmented ones. We evaluated our method on five public fundus datasets, corresponding to five different domains. Extensive experiments demonstrate that our method achieves better generalization performance on unseen target domains compared with state-of-the-art domain generalization methods.

Our contributions are three-fold:

- We propose a content style augmentation method that generates augmented deep features by moving the original deep features toward correct directions with appropriate degrees, thus enabling the training data to cover a wider range of unseen domains.
- We devise an augmented feature utilization strategy to quantify the contributions of augmented features to model training, for exploiting well-augmented features while mitigating the negative effects of over-augmented ones.
- Our ConStyX outperforms the baseline and five state-of-the-art DG methods on the joint optic disc (OD) and optic cup (OC) segmentation benchmarks.

2 Method

Notions & Notations. Given a source domain $D = \{(X_i, Y_i)_{i=1}^{M}\}$, for single domain generalized medical segmentation, the objective is to learn a segmentation model $f(\cdot;\theta)$ from D that shows strong generalization capability on unseen domains. The segmentation model $f(\cdot;\theta)$ consists of an encoder $E(\cdot;\theta_1)$ and a segmentation head $H(\cdot;\theta_2)$. Let $Z_i \in R^{N\times H\times W}$ be the deep features, where N, H, and W denote the channel, height, and width respectively, and $\mathbf{z}_i^j \in R^{N\times 1\times 1}$ be the deep feature of the j^{th} pixel for X_i with class c.

Overview. Considering the limited diversity in style augmentation and the detrimental effects of over-augmented samples, this work aims to develop an effective and controllable content-style augmentation method for generalizable medical image segmentation. Our core assumption is that: moving a deep feature along a specific direction generates a new feature corresponding to an augmented image of the same class but with distinct style and content characteristics. Based on this, we propose a content and style augmentation framework (ConStyX). As illustrated in Fig. 2, ConStyX consists of two key components: 1) Deep Feature Augmentation algorithm (DFA) for intra-domain and cross-domain feature movements under the guidance of intra-class variation and feature gradient and 2) Augmented Feature Utilization strategy (AFU) for augmented feature re-weighting according to their contributions to model training quantified by feature similarity and prediction confidence.

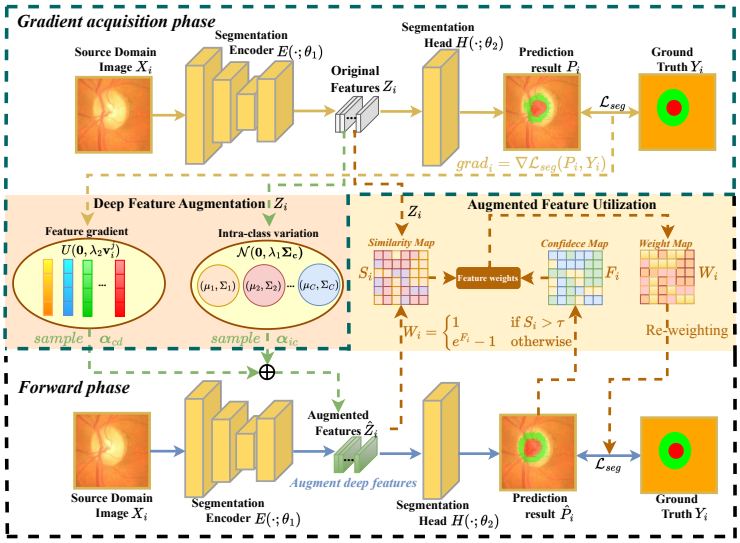


Fig. 2. Overview of the proposed Content Style Augmentation framework (ConStyX). It includes: 1) Deep Feature Augmentation algorithm (DFA) for content-style augmentation under the guidance of intra-class variation and feature gradient and 2) Augmented Feature Utilization strategy (AFU) for augmented feature re-weighting.

2.1 Deep Feature Augmentation (DFA)

The fundamental concept of DFA lies in *How to appropriately move the deep features while preserving their original semantic information.* To this end, we determine the direction and degree of movement by **intra-class variation** and **feature gradient**. Specifically, we construct two augmentation distributions, i.e., $\mathcal{N}(\mathbf{0}, \lambda_1\boldsymbol{\Sigma}_c)$ and $U(\mathbf{0}, \lambda_2\mathbf{v}_i^j)$, from which movement vectors can be sampled for feature augmentation:

$$\hat{\mathbf{z}}_i^j = \mathbf{z}_i^j + \boldsymbol{\alpha}_{ic} + \boldsymbol{\alpha}_{cd}, \quad \boldsymbol{\alpha}_{ic} \sim \mathcal{N}(\mathbf{0}, \lambda_1\boldsymbol{\Sigma}_c), \quad \boldsymbol{\alpha}_{cd} \sim U(\mathbf{0}, \lambda_2\mathbf{v}_i^j) \tag{1}$$

where $\boldsymbol{\alpha}_{ic}$ represents the intra-domain augmentation vector determined by capturing the maximum **intra-domain variation**, and $\boldsymbol{\alpha}_{cd}$ denotes the cross-domain augmentation factor guided by **feature gradients**, $\boldsymbol{\Sigma}_c$ refers to a class-conditional covariance, $\mathbf{v}_i^j$ indicates feature mask, and λ_1 and λ_2 are two scaled factors. Through these movements, deep features are appropriately augmented along the correct directions with proper degrees, generating augmented features that correspond to content-style augmented samples in the image space.

Intra-class variation. It yields the movement vector $\boldsymbol{\alpha}_{ic} \sim \mathcal{N}(\mathbf{0}, \lambda_1\boldsymbol{\Sigma}_c)$ to enable features to move along the maximum intra-class variation direction. Concretely, we extract a feature map Z_i from image X_i and establish a zero-mean multivariate normal distribution for the feature points corresponding to a certain

class based on the label Y_i. Let $\boldsymbol{\mu}_c^{(t)}$ and $\boldsymbol{\Sigma}_c^{(t)}$ represent the mean and covariance matrix of the c^{th} class features in the first t iterations, which are estimated using online weighted averaging; $\bar{\boldsymbol{\mu}}_c^{(t)}$ and $\bar{\boldsymbol{\Sigma}}_c^{(t)}$ are the mean and covariance matrix of the features of c^{th} class in the t^{th} iteration; $n_c^{(t)}$ denotes the total number of deep features belonging to c^{th} class in the first t iterations, and $m_c^{(t)}$ denotes the number of deep features belonging to c^{th} class in the t^{th} iteration. For the c^{th} class feature points, we establish the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \lambda_1 \boldsymbol{\Sigma}_c)$, where the class-conditional covariance $\boldsymbol{\Sigma}_c$ in the first t iterations is obtained by:

$$\boldsymbol{\Sigma}_c^{(t)} = \frac{n_c^{(t-1)} \boldsymbol{\Sigma}_c^{(t-1)} + m_c^{(t)} \bar{\boldsymbol{\Sigma}}_c^{(t)}}{n_c^{(t-1)} + m_c^{(t)}} + \frac{n_c^{(t-1)} m_c^{(t)} \Delta\boldsymbol{\mu}_c \Delta\boldsymbol{\mu}_c^{\top}}{(n_c^{(t-1)} + m_c^{(t)})^2} \tag{2}$$

where $\boldsymbol{\mu}_c^{(t)} = \frac{n_c^{(t-1)} \boldsymbol{\mu}_c^{(t-1)} + m_c^{(t)} \bar{\boldsymbol{\mu}}_c^{(t)}}{n_c^{(t-1)} + m_c^{(t)}}$, $\Delta\boldsymbol{\mu}_c = \boldsymbol{\mu}_c^{(t-1)} - \bar{\boldsymbol{\mu}}_c^{(t)}$, and $n_c^{(t)} = n_c^{(t-1)} + m_c^{(t)}$.

Feature gradient. By capturing the intra-class variations, the features can be forced to move along the direction of the most significant intra-class variation. However, this strategy limits the augmentation to be operated within the source domain. To further expand the augmented training data distribution over unseen domains, we introduce gradient-guided feature movement to generate another moving vector $\boldsymbol{\alpha}_{cd} \sim U(\mathbf{0}, \lambda_2 \mathbf{v}_i^j)$, which is applied along the minimum gradient directions to ensure cross-domain and semantic-invariant content style augmentation. Specifically, we first obtain the gradient of $\mathbf{z}_i^j$: $grad_i^j = \nabla \mathcal{L}_{seg}(P_i^j, Y_i^j)$, where $\mathcal{L}_{seg}$ denote the segmentation loss and P_i^j represents the model prediction. Note that model parameters are not updated in this process. Then, we define the feature mask $\mathbf{v}_i^j \in R^{N \times 1 \times 1}$ as:

$$(\mathbf{v}_i^j)_n = \begin{cases} 1 & \text{if } n \in pos_i^j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where the position of the minimum k partial derivatives in $\mathbf{z}_i^j$ are obtained by: $pos_i^j = \text{MinSort}(grad_i^j)[:k]$. $\text{MinSort}(\cdot)$ is an ascending sorting function.

Based on the feature mask $\mathbf{v}_i^j$, we construct the uniform distribution $U \sim (\mathbf{0}, \lambda_2 \mathbf{v}_i^j)$ to for further feature augmentation, where λ_2 denotes scaled factor.

2.2 Augmented Feature Utilization (AFU)

Although DFA aims to move deep features in appropriate directions for content style augmentation, it inevitably generates new features with distinct characteristics. We categorize these features into three types: 1) **trivial-augmented features**: maintain similar information with the original features, 2) **over-augmented features** lose the original information and even become noise, and 3) **well-augmented features** augment the content and style information appropriately while maintaining their original semantic information. It is crucial to suppress the adverse effects of over-augmented features during model training while sufficiently exploiting the well-augmented ones.

To this end, we introduce AFU, which leverages both **feature similarity** and **prediction confidence** to determine the contributions of the three types of augmented features to model training. Specifically, we compute the **cosine similarity** map S_i between an original deep feature map Z_i and its augmented counterpart $\hat{Z}_i$, as well as the **prediction confidence** $F_i = 1 - \text{MinMax}(-\sum_{m=1}^{C}(\hat{P}_i)_m log((\hat{P}_i)_m))$ for augmented feature $\hat{Z}_i$ (where MinMax$(\cdot)$ denotes a Min-Max normalization operation, and F_i is normalized to $[0, 1]$). Formally, based on S_i and F_i, the three type augmented features are defined as: 1) the augmented feature points with a cosine similarity higher than threshold τ are considered trivial-augmented features, 2) those with a similarity lower than τ and larger prediction confidence are divided as well-augmented features, and 3) those with a similarity lower than τ and lower confidence are regarded as over-augmented features. During the training process, the weight W_i is employed to determine the contributions of the augmented feature $\hat{Z}_i$ to model training:

$$W_i^j = \begin{cases} 1 & \text{if } S_i^j > \tau \\ e^{F_i^j} - 1 & \text{otherwise} \end{cases} \tag{4}$$

which serves as a pixel-wise weight for the segmentation loss (combining cross-entropy and Dice loss).

3 Experiment

Datasets and Evaluation Metrics. *Datesets:* We use five fundus datasets [11,9,22,15] corresponding to five different domains for joint segmentation of optic cup (OD) and optic disc (OC): BinRushed (195 images), Magrabia (95 images), REFUGE (400 images), ORIGA (650 images), and Drishti-GS (101 images); each image is resized to 512×512. We adopt the extremely challenging single-domain generalization setting, where one of the datasets is considered the source domain and divided into training and validation sets with a 9:1 ratio, while the remaining datasets serve as a test set. *Evaluation metrics:* Dice similarity coefficient (DSC, %).

Implementation Details. *Experimental environment:* All experiments are conducted using PyTorch [12] on a Tesla V100 with 32GB GPU memory. U-Net [14] with a modified ResNet-34 encoder [6] is used as the segmentation backbone for both our ConStyX and all compared methods. *Hyperparameter setting:* We set $k = 5$, $\lambda_1 = 1$, and $\lambda_2 = 0.5$ for the DFA module, and the threshold $\tau = 0.6$ for the AFU module. ConStyX is optimized using the SGD optimizer with a momentum of 0.99 and an initial learning rate of 0.001 decayed according to a polynomial rule. The batch size and the number of training epochs are 8 and 100, respectively.

3.1 Comparison with other DG methods

We conduct comparative experiments over various state-of-the-arts, including: statistics transfer-based (MixStyle [26], DSU [10], EFDM [21], TriD [3], and

Table 1. Comparative results of cross-domain segmentation performance (OD, OC). The best results are highlighted in **bold**.

Method	Domain1	Domain2	Domain3	Domain4	Domain5	Average
	DSC ↑	DSC ↑	DSC ↑	DSC ↑	DSC ↑	DSC ↑
MixStyle [26]	(86.67, 64.86)	(87.34, 73.78)	(86.83, 67.34)	(78.03, 59.49)	(83.52, 67.27)	75.51
DSU [10]	(86.64, 65.99)	(87.44, 74.20)	(86.38, 66.80)	(78.77, 57.98)	(83.44, 66.00)	75.36
EFDM [21]	(86.32, 65.11)	(88.03, 75.62)	(86.55, 67.34)	(77.02, 58.22)	(83.66, 67.60)	75.55
TriD [3]	(84.03, 63.73)	(85.75, 70.16)	(86.93, **67.70**)	(75.46, 66.96)	(84.85, 57.38)	74.30
CSU [23]	(87.58, 69.19)	(87.50, 74.91)	(86.87, 67.55)	(79.58, 57.93)	(83.50, 66.46)	76.11
RandConv [20]	(85.87, 65.99)	(88.10, 76.05)	(86.47, 65.77)	(79.33, 57.79)	(83.30, 67.52)	75.62
MoreStyle[24]	(80.38, 59.60)	(88.47, 64.32)	(82.63, 63.85)	(77.07, 51.91)	(78.14, 51.63)	69.80
CCSDG [7]	(86.21, 63.55)	(**90.34**, **78.24**)	(87.18, 65.44)	(81.00, 63.16)	(82.21, 64.57)	76.19
ConStyX (ours)	(**88.95**, **72.55**)	(89.86, 77.61)	(**88.17**, 67.22)	(**81.09**, **67.50**)	(**86.67**, **69.19**)	**78.88**

CSU [23]), random convolution-based (RandConv [20]), adversarial noise-based (MoreStyle [24]), and feature disentanglement-based (CCSDG [7]) methods.

Overall, both the segmentation performance in Table 1 and the qualitative results in Fig. 3 suggest that our ConStyX consistently outperforms other DG methods, showing its superior generalization capability for cross-domain medical image segmentation. Specifically, the five statistical information transferred-based methods and the random convolution-based method yield similar segmentation results. This phenomenon is attributed to their limited augmentation efficacy (only augmenting image style), resulting in augmented samples that only cover a narrow range of unseen domains. Due to the adversarial noise generating augmented samples that deviate significantly from real images, MoreStyle achieved unsatisfactory results. Meanwhile, CCSDG attains the second-highest average DSC across domains, i.e., 76.19 %, by integrating feature disentanglement and various style augmentation techniques. In contrast, our ConStyX outperforms the second-highest result by 2.09% in terms of DSC, due to the advantages of the content-style augmentation to generate diverse augmented samples and the feature re-weighting strategy to fully leverage well-augmented features.

3.2 Ablation study

Analysis of the proposed components. We conduct an ablation study on the five domains to evaluate the effect of our proposed modules including DFA and AFU. In Table 2, one can observe that the segmentation performance gradually increases as each component is incorporated into our method. Compared with the baseline, our final model obtains average segmentation improvements of 9.17% and 9.78% across five distinct domains, significantly improving the generalizability of the baseline model.

Investigation of distribution forms. To investigate the influence of distribution forms on feature movement operations, we conduct a comparative experiment on the feature gradient-guided movement by constructing two gradient-guided augmentation distributions: a uniform distribution $\boldsymbol{\alpha}_{cd} \sim U(\mathbf{0}, \lambda_2 \mathbf{v}_i^j)$ and a normal distribution $\boldsymbol{\alpha}_{cd} \sim \mathcal{N}(\mathbf{0}, \lambda_2 \mathbf{v}_i^j \mathbf{I})$. As shown in Table 3, the segmentation

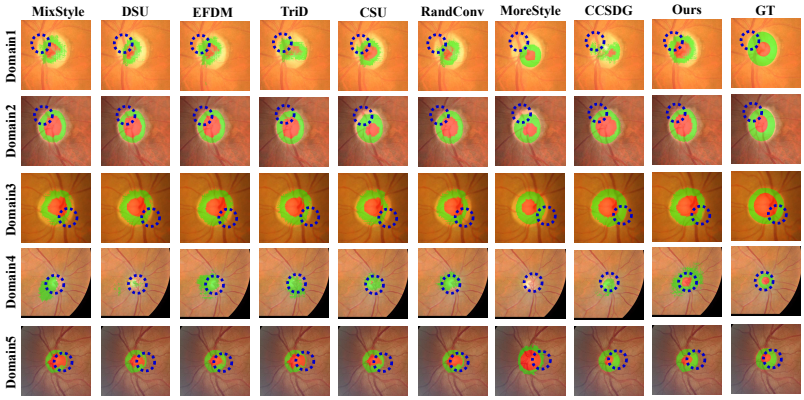


Fig. 3. Visualization of OD and OC segmentation results. Blue dashed circles highlight some regions with significant differences in the segmentation results.

Table 2. Performance (OD, OC) of ablation study on joint segmentation of OD and OC. The best results are highlighted in **bold**.

Method	Domain1	Domain2	Domain3	Domain4	Domain5	Average
	DSC ↑	DSC ↑	DSC ↑	DSC ↑	DSC ↑	DSC ↑
Baseline	(86.21, 53.02)	(85.81, 58.05)	(78.73, 50.80)	(78.41, 57.56)	(81.72, 60.72)	69.10
Baseline+DFA	(87.87, 69.39)	(**89.90**, 76.85)	(88.11, 66.89)	(79.98, **68.03**)	(**86.67**, 69.04)	78.27
Baseline+DFA+AFU	(**88.95**, **72.55**)	(89.86, **77.61**)	(**88.17**, **67.22**)	(**81.09**, 67.50)	(**86.67**, **69.19**)	**78.88**

Table 3. The influence of different distribution and position schemes for feature augmentation. The best results are highlighted in **bold**.

Method	Domain1	Domain2	Domain3	Domain4	Domain5	Average
	DSC ↑	DSC ↑	DSC ↑	DSC ↑	DSC ↑	DSC ↑
Normal Distribution	(88.74, 67.56)	(89.75, **77.76**)	(87.49, 65.67)	(**81.69**, 67.37)	(**87.66**, 68.80)	78.25
Uniform Distribution	(**88.95**, **72.55**)	(**89.86**, 77.61)	(**88.17**, **67.22**)	(81.09, **67.50**)	(86.67, **69.19**)	**78.88**
Random position	(87.32, 66.16)	(**90.19**, 77.00)	(87.22, 66.85)	(**81.39**, 67.63)	(85.86, 68.56)	77.82
Maximum k position	(85.58, 56.55)	(88.76, 74.71)	(**88.46**, 65.76)	(80.22, 63.40)	(**86.96**, 67.86)	75.83
Minimum k position	(**88.95**, **72.55**)	(89.86, **77.61**)	(88.17, **67.22**)	(81.09, **67.50**)	(86.67, **69.19**)	**78.88**

results of these two distribution forms are comparable, validating the robustness of our method to different distributions.

Analysis of perturbation positions. To verify that perturbations at the positions with the minimum partial derivatives have minimal impact on feature semantics, we conducted a perturbation position analysis experiment across five domains. As shown in Table 3, perturbing the k positions with the minimum partial derivatives resulted in the best performance of the model, confirming the importance of considering direction during the feature moving process.

4 Conclusion

We propose a novel domain randomization-based DG method, called ConStyX, for generalizable medical image segmentation. To enable the source domain data to cover a wider range of unseen domains, ConStyX augments the content and style of the training data by moving the deep features toward correct directions with proper degrees. Besides, we define three types of augmented features and assign different weights to them to mitigate the negative impact of over-augmented features on model training, while fully leveraging well-augmented features. Extensive experiments on five fundus datasets demonstrate that ConStyX achieves compelling performance over the state-of-the-art methods.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (No.62076059), the Science and Technology Joint Project of Liaoning province (2023JH2/101700367, ZX20240193). Osmar Zaiane acknowledges the funding from NSERC and the Canada CIFAR AI Chairs Program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep semantic segmentation of natural and medical images: a review. Artificial intelligence review **54**, 137–178 (2021)
2. Chen, C., Li, Z., Ouyang, C., Sinclair, M., Bai, W., Rueckert, D.: Maxstyle: Adversarial style composition for robust medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 151–161. Springer (2022)
3. Chen, Z., Pan, Y., Ye, Y., Cui, H., Xia, Y.: Treasure in distribution: a domain randomization based multi-source domain generalization for 2d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 89–99. Springer (2023)
4. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11580–11590 (2021)
5. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
6. 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)
7. Hu, S., Liao, Z., Xia, Y.: Devil is in channels: Contrastive single domain generalization for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–23. Springer (2023)
8. Jia, Y., Hoyer, L., Huang, S., Wang, T., Van Gool, L., Schindler, K., Obukhov, A.: Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In: European Conference on Computer Vision. pp. 91–109. Springer (2025)

9. Latha, G., Priya, P.A.: Glaucoma retinal image detection and classification using machine learning algorithms. In: Journal of Physics: Conference Series. vol. 2335, p. 012025. IOP Publishing (2022)
10. Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., Duan, L.Y.: Uncertainty modeling for out-of-distribution generalization. arXiv preprint arXiv:2202.03958 (2022)
11. Orlando, J.I., Fu, H., Barbosa Breda, J., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., Lee, J., Li, X., Liu, P., Lu, S., Murugesan, B., Naranjo, V., Phaye, S.S.R., Shankaranarayana, S.M., Sikka, A., Son, J., van den Hengel, A., Wang, S., Wu, J., Wu, Z., Xu, G., Xu, Y., Yin, P., Li, F., Zhang, X., Xu, Y., Bogunović, H.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Analysis p. 101570 (Jan 2020). https://doi.org/10.1016/j.media.2019.101570, http://dx.doi.org/10.1016/j.media.2019.101570
12. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
13. Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W.: Semantic-aware domain generalized segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2594–2605 (2022)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
15. Sivaswamy, J., Krishnadas, S.R., Datt Joshi, G., Jain, M., Syed Tabish, A.U.: Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI). pp. 53–56 (2014). https://doi.org/10.1109/ISBI.2014.6867807
16. Upchurch, P., Gardner, J., Pleiss, G., Pless, R., Snavely, N., Bala, K., Weinberger, K.: Deep feature interpolation for image content changes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7064–7073 (2017)
17. Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Philip, S.Y.: Generalizing to unseen domains: A survey on domain generalization. IEEE transactions on knowledge and data engineering **35**(8), 8052–8072 (2022)
18. Wang, Y., Huang, G., Song, S., Pan, X., Xia, Y., Wu, C.: Regularizing deep networks with semantic data augmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3733–3748 (2021)
19. Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S.: A survey on incorporating domain knowledge into deep learning for medical image analysis. Medical Image Analysis **69**, 101985 (2021)
20. Xu, Z., Liu, D., Yang, J., Raffel, C., Niethammer, M.: Robust and generalizable visual representation learning via random convolutions. arXiv preprint arXiv:2007.13003 (2020)
21. Zhang, Y., Li, M., Li, R., Jia, K., Zhang, L.: Exact feature distribution matching for arbitrary style transfer and domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8035–8045 (2022)
22. Zhang, Z., Yin, F., Liu, J., Wong, W., Tan, N., Lee, B., Cheng, J., Wong, T.: Origa: An online retinal fundus image database for glaucoma analysis and research. In: Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. pp. 3065–3068

23. Zhang, Z., Wang, B., Jha, D., Demir, U., Bagci, U.: Domain generalization with correlated style uncertainty. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2000–2009 (2024)
24. Zhao, H., Dong, W., Yu, R., Zhao, Z., Du, B., Xu, Y.: Morestyle: relax low-frequency constraint of fourier-based image reconstruction in generalizable medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 434–444. Springer (2024)
25. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4396–4415 (2022)
26. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: ICLR (2021)