Interpretability-Guided Data Augmentation for Robust Segmentation in Multi-centre Colonoscopy Data

Valentina Corbetta[1,2(✉)] , Regina Beets-Tan[1,2] , and Wilson Silva[1]

[1] Department of Radiology, The Netherlands Cancer Institute, Amsterdam,
The Netherlands
v.corbetta@nki.nl
[2] GROW School for Oncology and Developmental Biology, Maastricht University
Medical Center, Maastricht, The Netherlands

Abstract. Multi-centre colonoscopy images from various medical centres exhibit distinct complicating factors and overlays that impact the image content, contingent on the specific acquisition centre. Existing Deep Segmentation networks struggle to achieve adequate generalizability in such data sets, and the currently available data augmentation methods do not effectively address these sources of data variability. As a solution, we introduce an innovative data augmentation approach centred on interpretability saliency maps, aimed at enhancing the generalizability of Deep Learning models within the realm of multi-centre colonoscopy image segmentation. The proposed augmentation technique demonstrates increased robustness across different segmentation models and domains. Thorough testing on a publicly available multi-centre dataset for polyp detection demonstrates the effectiveness and versatility of our approach, which is observed both in quantitative and qualitative results. The code is publicly available at: https://github.com/nki-radiology/interpretability_augmentation.

1 Introduction

The adoption of Deep Learning (DL) techniques has significantly advanced medical image segmentation in recent years [4,12]. UNet and other U-shaped architectures have been pivotal in this revolution [11], remaining competitive even with the introduction of newer models.

However, when DL models are applied to unseen datasets acquired from different scanners or clinical centers, their performance at inference time declines noticeably [2,16]. This is due to *domain shifts*, caused by variations in data statistics between different clinical centers, resulting from varying patient populations, scanners, and scan settings [20,21]. These disparities in patient characteristics

and imaging settings can significantly affect the model's ability to generalize effectively [10,21].

To further integrate DL models into clinical practice, it is crucial for them to be robust against these changes and demonstrate a high level of generalizability. The most straightforward approach to address domain shifts is by collecting and annotating as many varied samples as possible. Nevertheless, acquiring and labeling enough data to encompass real-world variation is prohibitively time-consuming and costly.
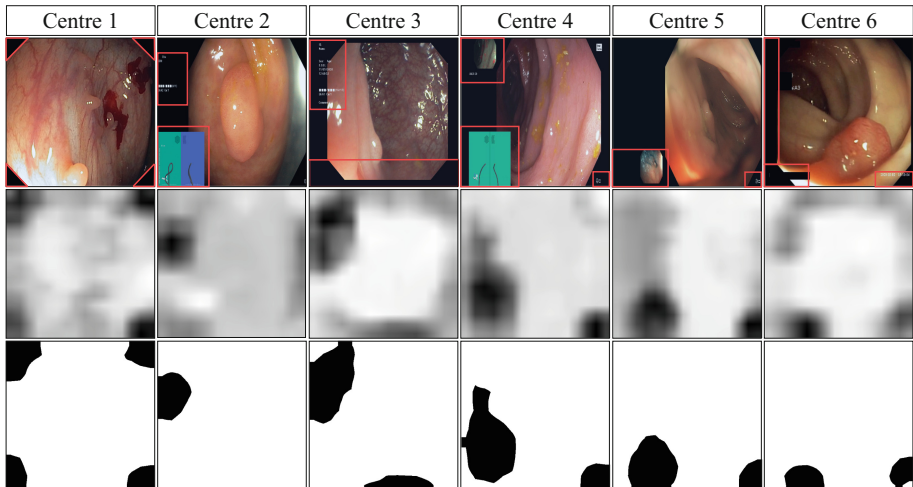
Fig. 1. Illustration of examples of the "extra"-anatomical content across the different centres in the PolypGen dataset and its impact on the GradCAM visualizations. The first row depicts the original image, the second and third row show the GradCAM visualizations and their binarization, respectively.

In recent years, Disentangled Representation Learning (DRL) has emerged as a promising solution to address the aforementioned limitations. This method encodes underlying variation factors into separate latent variables, capturing valuable information relevant to the task at hand. By adopting DRL, DL models gain increased robustness against domain shifts, reducing the need for a large number of meticulously labeled samples [13]. Various DRL models have been employed for segmentation in the context of multi-centre datasets, yielding state-of-the-art outcomes. One such model is the Spatial Decomposition Network (SDNet), which decomposes 2-dimensional (2D) medical images into spatial anatomical factors (content) and non-spatial modality factors (style) [3]. Expanding upon SDNet, Jiang et al. [8] have made additional advancements by further disentangling the pathology factor from the anatomy, particularly when the ground truth mask for anatomy is available. To further improve generalizability, Liu et al. [14] combined DRL with meta-learning, while Shin et al. [18]

have effectively disentangled intensity and non-intensity factors to enable domain adaptation in Computerised Tomography (CT) images.

Despite the significant advancements made by DRL methods in improving model generalizability, it is important to acknowledge that these methods assume that the shift introduced by unseen domains is embedded within the "style" features. However, this assumption does not always hold true, especially in scenarios like videos and images of colonoscopies, or other endoscopy applications. In such cases, various confounding factors affect the content of the images, depending on the domain, from hereon also referred to as *centre*, of acquisition. These factors include image miniaturization, anonymization, and depictions of the instrument's position during image acquisition, as shown in the first row of Fig. 1. As illustrated in Sect. 3, both traditional methods, like UNet and DeepLabV3+ [5], and DRL models like SDNet encounter challenges in generalizing to domains heavily characterized by this additional content that is unrelated to the anatomy.

To address these limitations, we propose an innovative data augmentation strategy based on interpretability techniques. Interpretability techniques have already been successfully applied to improve DL models' performance in medical image analysis tasks: Silva et al. [19] exploited interpretability methods to improve medical image retrieval in the radiological workflow; in [15], Gradient-tweighted Class Activation Mappings (GradCAMs) are used to improve generalized zero shot learning for medical image classification. Firstly, we pretrain a classifier network to identify the respective centres to which the images belong, using the same training set that will later be employed to train the segmentation module. Thus, we ensure a fair assessment of the segmentation model. During the training phase of the segmentation network, we employ the pre-trained classifier to generate visual explanations for the input batch. In this work, we use GradCAM [17] to produce the visual explanations, a widely adopted technique for visualizing and interpreting the decision-making process of Convolutional Neural Networks (CNNs) in a wide variety of computer vision tasks. By leveraging gradients of the predicted class, GradCAM assigns importance weights to different spatial locations within the last convolutional layer. This allows us to identify the regions in the input image that significantly contribute to the model's decision-making process. Figure 1, specifically the second row, showcases examples of the generated GradCAM visualizations for each centre. Notably, the classifier predominantly focuses on areas where the "extra"-anatomical content resides (indicated by darker regions). We binarize the generated GradCAMs, as depicted in the third row of Fig. 1, and multiply them with a probability p with the input to the segmentation network. Thus, this approach randomly blocks out the additional information, enabling the segmentation network to place greater emphasis on the anatomical regions.

The key contributions of our research can be summarized as follows:

– We introduce a novel data augmentation technique based on interpretability techniques. By incorporating visual explanations, we develop a robust technique that can be readily applied to different and multiple domains. This
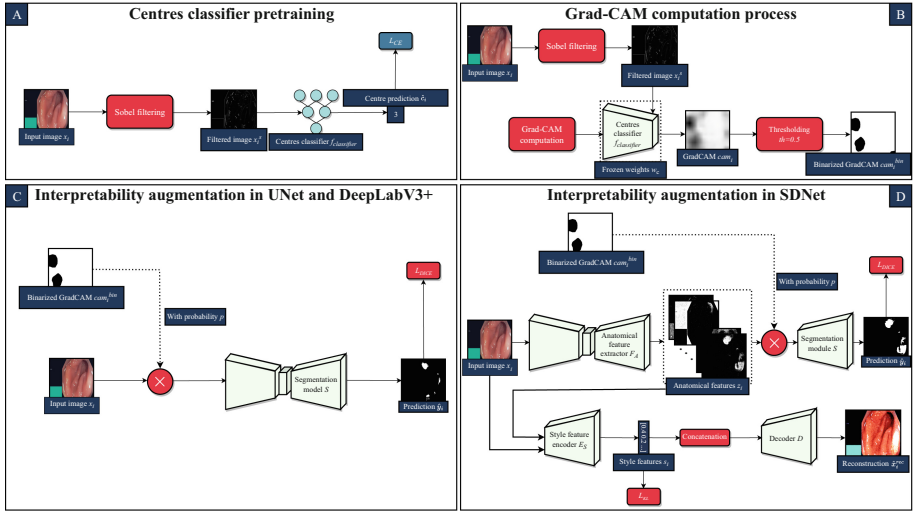
Fig. 2. Illustration of the proposed method. In **A**, the pretraining of the classifier is shown. **B** depicts the process of computing GradCAM visualizations. As for **C** and **D**, they demonstrate the incorporation of the interpretability-guided augmentation in the training phase of UNet and DeepLabV3+, and SDNet, respectively.

stands in contrast to standard augmentation techniques, which are limited in their ability to capture the variability of non-synthetic data [21].

– We apply and adapt the proposed methodology to both two conventional baseline models, UNet and DeepLabV3+ with ResNet101 as backbone, and a DRL model, specifically SDNet. This showcases the versatility of our augmentation strategy across different architectures.

– We conduct thorough testing of our method using an open-source multi-centre dataset, PolypGen [1], to demonstrate the robustness of our technique and its effectiveness in diverse domains.

2   Methodology

The proposed methodology is illustrated in Fig. 2, outlining the pretraining of the classifier module, the generation process of the GradCAM visualizations and their integration into the UNet, DeepLabV3+ and SDNet models.

2.1   Pretraining of Centres Classifier

For the classifier backbone, we employ a ResNet50 architecture [7] with pre-training on ImageNet [6]. The module is trained on the same training set used for subsequent training of the segmentation networks. Given an input image x_i, we initially apply Sobel filtering [9] to emphasize the edges characterizing

---

Algorithm 1. Gradient-weighted Class Activation Mapping (GradCAM)

1: **procedure** GRADCAM($\mathbf{x_i^s}$, $\mathbf{w_c}$)
2:     $\mathbf{g} \leftarrow$ Compute the gradient of the target class score c w.r.t the feature map a of the last convolutional layer l_c
3:     $\mathbf{i_w} \leftarrow$ Compute the importance vector of each feature channel by applying global average pooling to $\mathbf{g}$
4:     $\mathbf{cam_i} \leftarrow$ Compute the class activation map by combining the importance vector $\mathbf{i_w}$ with the corresponding feature map a
5:     $\mathbf{cam_i} \leftarrow$ ReLU($\mathbf{cam_i}$)   ▷ Apply ReLU activation to remove negative values
6:     $\mathbf{cam_i} \leftarrow$ normalize($\mathbf{cam_i}$)   ▷ Normalize the class activation map
7:     $\mathbf{cam_i} \leftarrow$ upsample($\mathbf{cam_i}$)   ▷ Upasample to match size of $\mathbf{x_i^s}$
8:     **return $\mathbf{cam_i}$**   ▷ Return the final GradCAM visualization

---

the extra-anatomical content and discard most of the anatomy. The resulting filtered image x_i^s is then passed through the ResNet50 to predict the original centre to which the image belongs. The classification network is trained using the Cross-Entropy (CE) Loss between the predicted $\hat{c}_i$ and the original centre label c_i.

2.2   GradCAM Visualizations Generation Process

During the training process of the segmentation network, the pretrained weights w_c of the centres classifier are loaded and kept frozen. The classifier is then used to perform inference on the Sobel-filtered input image x_i^s. Subsequently, the GradCAM visualization cam_i is generated following the steps outlined in Algorithm 1. The GradCAM visualizations provide a coarse representation of the areas in the input image x_i^s that the classifier focused on to make its prediction. To ensure that we do not inadvertently block useful content for the downstream segmentation task, we binarize cam_i using a threshold $th = 0.5$ to obtain cam_i^{bin}. This ensures that the augmentation procedure described in the subsequent paragraphs masks only the most relevant "extra"-anatomical content.

2.3   Interpretability-Guided Data Augmentation

We will now provide a detailed explanation of how the GradCAM visualizations are utilized as an augmentation technique to enhance the robustness and generalizability of segmentation models.

UNet and DeepLabV3+. The integration of the GradCAM visualizations into the UNet and DeepLabV3+ training process is straightforward. With a probability p, we multiply the input image x_i by the corresponding GradCAM visualization cam_i^{bin}. We introduce a probability p for the multiplication step to mitigate the risk of covering important areas for the downstream task. Figure 3 provides two examples of augmented samples in the UNet and DeepLabV3+
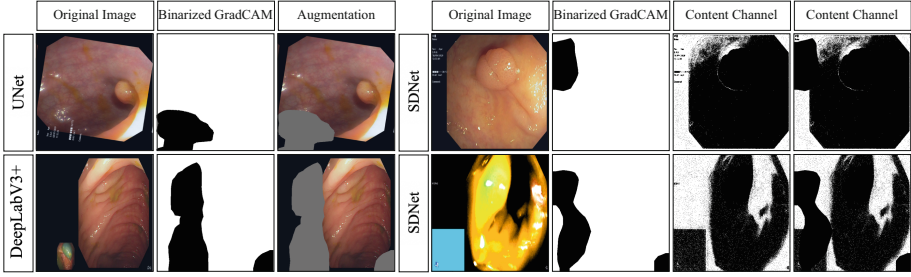
Fig. 3. Examples of augmented samples with our interpretability-guided augmentation, for UNet, DeepLabV3+ and SDNet.

training process. The models are trained by computing the Dice Loss between the predicted segmentation mask $\hat{y}_i$ and the ground truth y_i.

SDNet. To gain a better understanding of how interpretability-guided augmentation is integrated into the SDNet, it is necessary to provide a brief overview of the model's structure. Initially, the anatomy encoder $F_{anatomy}$ encodes input image x_i into a multi-channel spatial representation, the anatomical features z_i. It is important to note that z_i has shape $N \times H \times W$, where N represents a fixed number of channels (e.g. 8), while H and W correspond to the height and width of the original image, respectively. The modality encoder E_s uses factor z_i along with the input image x_i to produce the latent vector s_i, representing the style features. These two representations, s_i and z_i, are combined to reconstruct the input image through the decoder network D. The anatomical representation z_i is then fed into the segmentation network S to produce the segmentation mask $\hat{y}_i$. For a more detailed explanation of the SDNet and its associated losses, we refer the reader to the schematic in Fig. 2(D) and to the original paper [3]. To perform the interpretability-guided augmentation, we multiply the multi-channel anatomical representation z_i with the binary GradCAM visualization cam_i^{bin} using a probability p before feeding it as input to the segmentation module S. Figure 3 provides two examples of augmented samples in the SDNet training process, in particular we report the effect on only one of the N channels of z_i for synthesis purposes. We made the decision to apply the augmentation on the anatomy representation z_i and not on the input image x_i. This approach is intended to mimic the process used in the UNet and DeepLabv3+, where we directly manipulate the input to the module dedicated to the downstream segmentation task, while keeping the rest of the SDNet architecture intact.

3 Results

Dataset and Implementation Details. To evaluate the proposed methodology, we utilized the publicly available PolypGen dataset, which comprises colonoscopy data collected from 6 different centres, encompassing diverse patient populations. Our analysis focused on single frame samples, resulting in a total of 1537 images (Centre 1: 256, Centre 2: 301, Centre 3: 457, Centre 4: 227, Centre 5: 208, Centre 6: 88). Notably, as illustrated in Fig. 1, the centres exhibit substantial variability in image content, both within and across centres. To assess

Table 1. The comparison results of the proposed method with the corresponding baseline models. The best result concerning the specific baseline considered is highlighted in **bold**. In the Dice column, the value of probability p that produced the best result in terms of Dice score in the interpretability-guided augmentation is indicated in parentheses.

|	Dice	Recall	Accuracy	Dice	Recall	Accuracy
Out-dist set	UNet			UNet interpretability augmentation		
centre 1	0.7257	0.7725	0.9589	**0.7353 (60%)**	0.7716	0.9587
centre 2	0.5762	**0.6454**	0,9440	**0.6062 (60%)**	0.5948	**0.9530**
centre 3	0.7054	0.6724	0.9619	**0.7470 (40%)**	**0.6966**	**0.9658**
centre 4	0.4223	**0.3804**	0.9594	**0.4519 (40%)**	0.3757	**0.9607**
centre 5	0.4725	0.4934	0.9581	**0.4893 (40%)**	**0.5369**	0.9583
centre 6	0.6423	0.5597	0.9593	**0.6574 (60%)**	**0.6132**	0.9634
	Dice	Recall	Accuracy	Dice	Recall	Accuracy
Out-dist set	Deeplabv3+			Deeplabv3+ interpretability augmentation		
centre 1	0.6003	**0.6284**	0.9482	**0.6155 (60%)**	0.6195	**0.9545**
centre 2	0.5498	**0.5358**	0.9398	**0.5679 (60%)**	0.5159	**0.9496**
centre 3	0.5696	0.5407	**0.9585**	**0.6314 (40%)**	**0.6745**	0.9561
centre 4	0.3424	**0.2471**	0.9483	**0.3592 (50%)**	0.2166	**0.9493**
centre 5	0.3867	0.4171	**0.9543**	**0.4046 (60%)**	**0.4479**	0.9533
centre 6	0.6268	0.5932	0.9631	**0.6342 (40%)**	**0.6126**	**0.9658**
	Dice	Recall	Accuracy	Dice	Recall	Accuracy
Out-dist set	SDNet			SDNet interpretability augmentation		
centre 1	0.7130	0.7583	0.9551	**0.7226 (40%)**	**0.7726**	**0.9575**
centre 2	0.5489	**0.5794**	0.9328	**0.5579 (50%)**	0.5482	**0.9464**
centre 3	0.7151	0.7082	**0.9620**	**0.7208 (40%)**	**0.7245**	0.9603
centre 4	**0.3981**	**0.3254**	**0.9591**	0.3841 (50%)	0.3360	0.9557
centre 5	0.4312	0.4293	0.9587	**0.4546 (40%)**	**0.4722**	0.9588
centre 6	0.6398	**0.6195**	0.9610	**0.6626 (60%)**	0.5977	**0.9621**

the generalizability of the models, we conducted 6 distinct experiments for each tested model: we trained and validated the models using 5 centres while reserving one centre as an out-of-distribution test set. Throughout the experiments, a patient-level split was applied for the in-distribution frames, with 80% of patients allocated for training, 10% for validation, and 10% for the in-distribution test set.

Standard preprocessing techniques were applied to the images, including resizing them to a dimension of 256×256 and normalizing the pixel values. Furthermore, additional augmentations were performed exclusively on the training set. These augmentations comprised rotation, horizontal and vertical flips (with a 50% probability), which were also applied to the ground truth and binary GradCAM visualizations. Additionally, colour jitter was applied to the images with a 30% probability. The models were trained for 300 epochs, except for the centres classifier which was trained just for 10 epochs, on an NVIDIA RTX$^{\text{TM}}$ A6000 GPU, utilizing a batch size of 4 and a learning rate of 10^{-5}.

Results and Discussion. The experimental results are presented in Table 1, where we evaluate segmentation outcomes using as metrics the Dice Score (DSC), recall, and accuracy. To determine the best-performing augmented models based on DSC, we conduct a parameter study on the probability value p, as elaborated in the following paragraph, and subsequently compare them with the corresponding baseline models. Concerning the DSC, the models trained with interpretability-guided augmentation demonstrate superior performance in nearly all experiments. Particularly noteworthy is the substantial improvement in Centre 3 for the UNet, with an increase of 4.16%. Additionally, the SDNet exhibits an increment of 2.57% in Centre 5, and for DeepLabV3+, there is a significant 6.18% increase in Centre 3. The results also consistently demonstrate improvements in terms of accuracy and recall in nearly all experiments. Figure 4 presents several qualitative examples of the segmentation results. Notably, the masks obtained using our proposed methodology exhibit reduced noise levels, and our approach demonstrates greater performance in detecting smaller polyps.

Parameters Study. We delve deeper into the effectiveness of our proposed technique, conducting a detailed study of the parameters for the all the analyzed architectures. Our augmentation method was employed with probabilities of 40%, 50% and 60%. The results of this study are displayed in the Supplementary Material. The study proves that the fine-tuning of the probability value p within our augmentation approach plays a pivotal role in enhancing the models' generalizability. Indeed, when the suitable probability p is applied, the augmented architectures surpass the performance of the baseline models in nearly all the tests.
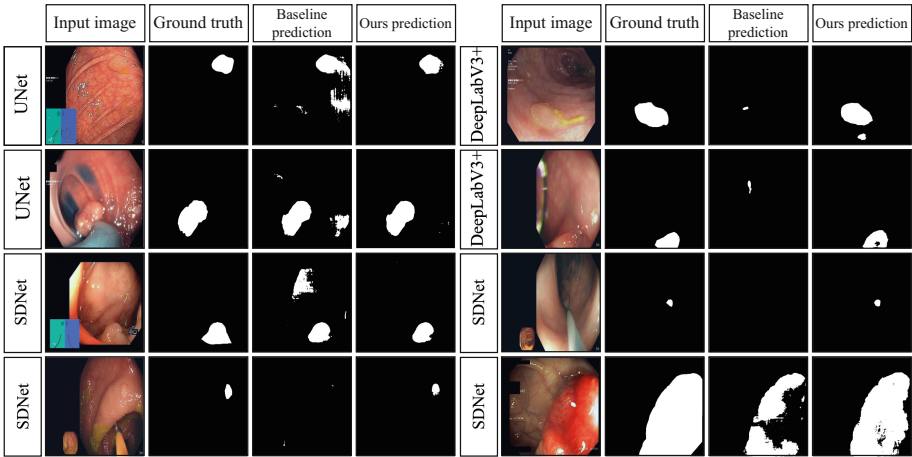
Fig. 4. Examples of qualitative results for the baselines and the interpretability-guided augmentation approach.

4 Conclusion

We have introduced an interpretability-guided augmentation technique aimed at improving the generalizability of DL models to unseen domains in colonoscopy segmentation tasks. We have demonstrated the strength and reliability of this proposed technique by successfully adapting it to three distinct architectures: UNet, SDNet and DeeplabV3+. Our future work will involve testing the proposed augmentation in low-data and semi-supervised settings, where DRL models, and in particular SDNet, significantly outperform conventional models. Moreover, we plan to investigate the adaptability of our methodology to other multi-centre endoscopy images, such as cystoscopy or laparoscopy, as they exhibit the same instrumentation-specific and User Interface (UI) overlays as colonoscopy data. Consequently, our method could improve models' generalizability to such data.

References

1. Ali, S., et al.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Sci. Data **10**(1), 75 (2023)
2. Campello, V.M., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging **40**(12), 3543–3554 (2021)
3. Chartsias, A., et al.: Disentangled representation learning in cardiac image analysis. Med. Image Anal. **58**, 101535 (2019)

4. Chen, C., et al.: Deep learning for cardiac image segmentation: a review. Front. Cardiovasc. Med. **7**, 25 (2020)
5. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
7. 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)
8. Jiang, H., et al.: Semi-supervised pathology segmentation with disentangled representations. In: Albarqouni, S., et al. (eds.) DART/DCL -2020. LNCS, vol. 12444, pp. 62–72. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60548-3_7
9. Kanopoulos, N., Vasanthavada, N., Baker, R.: Design of an image edge detection filter using the Sobel operator. IEEE J. Solid-State Circuits **23**(2), 358–367 (1988). https://doi.org/10.1109/4.996
10. Li, Y., Chen, J., Xie, X., Ma, K., Zheng, Y.: Self-loop uncertainty: a novel pseudo-label for semi-supervised medical image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 614–623. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_60
11. Liu, L., Cheng, J., Quan, Q., Wu, F.X., Wang, Y.P., Wang, J.: A survey on U-shaped networks in medical image segmentations. Neurocomputing **409**, 244–258 (2020)
12. Liu, X., Song, L., Liu, S., Zhang, Y.: A review of deep-learning-based medical image segmentation methods. Sustainability **13**(3), 1224 (2021)
13. Liu, X., Sanchez, P., Thermos, S., O'Neil, A.Q., Tsaftaris, S.A.: Learning disentangled representations in the imaging domain. Med. Image Anal. **80**, 102516 (2022)
14. Liu, X., Thermos, S., O'Neil, A., Tsaftaris, S.A.: Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 307–317. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_29
15. Mahapatra, D., Ge, Z., Reyes, M.: Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. IEEE Trans. Med. Imaging **41**(9), 2443–2456 (2022). https://doi.org/10.1109/TMI.2022.3163232
16. Prados, F., et al.: Spinal cord grey matter segmentation challenge. Neuroimage **152**, 312–329 (2017)
17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
18. Shin, S.Y., Lee, S., Summers, R.M.: Unsupervised domain adaptation for small bowel segmentation using disentangled representation. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part III 24, vol. 12903, pp. 282–292. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_27

19. Silva, W., Poellinger, A., Cardoso, J.S., Reyes, M.: Interpretability-guided content-based medical image retrieval. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 305–314. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_30

20. Tao, Q., et al.: Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology **290**(1), 81–88 (2019)

21. Zhang, L., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging **39**(7), 2531–2540 (2020)