Do We Really Need that Skip-Connection? Understanding Its Interplay with Task Complexity

Amith Kamath[1], Jonas Willmann[2,3], Nicolaus Andratschke[2], and Mauricio Reyes[1,4](✉)

[1] ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland
mauricio.reyes@unibe.ch

[2] Department of Radiation Oncology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

[3] Center for Proton Therapy, Paul Scherrer Institute, Villigen, Switzerland

[4] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

https://github.com/amithjkamath/to_skip_or_not

Abstract. The U-Net architecture has become the preferred model used for medical image segmentation tasks. Since its inception, several variants have been proposed. An important component of the U-Net architecture is the use of skip-connections, said to carry over image details on its decoder branch at different scales. However, beyond this intuition, not much is known as to what extent skip-connections of the U-Net are necessary, nor what their interplay is in terms of model robustness when they are subjected to different levels of task complexity. In this study, we analyzed these questions using three variants of the U-Net architecture (the standard U-Net, a "No-Skip" U-Net, and an Attention-Gated U-Net) using controlled experiments on varying synthetic texture images and evaluated these findings on three medical image data sets. We measured task complexity as a function of texture-based similarities between foreground and background distributions. Using this scheme, our findings suggest that the benefit of employing skip-connections is small for low-to-medium complexity tasks, and its benefit appears only when the task complexity becomes large. We report that such incremental benefit is non-linear, with the Attention-Gated U-Net yielding larger improvements. Furthermore, we find that these benefits also bring along robustness degradations on clinical data sets, particularly in out-of-domain scenarios. These results suggest a dependency between task complexity and the choice/design of noise-resilient skip-connections, indicating the need for careful consideration while using these skip-connections.

Keywords: Image segmentation · U-Net · robustness

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_29.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 302–311, 2023.
https://doi.org/10.1007/978-3-031-43901-8_29

1 Introduction

Due to the broad success of U-Nets [17] for image segmentation, it has become the go-to architecture in the medical image computing community. Since its creation in 2015, much research has been dedicated to exploring variants and improvements over the standard base model [3]. However, Isensee et al. [12] showed with their not-new-U-Net (nnU-Net) that the success of the U-Net relies on a well-prepared data pipeline incorporating appropriate data normalization, class balancing checks, and preprocessing, rather than on architecture changes. Arguably the two most important challenges at present for medical image segmentation are generalization and robustness. A lack of generalization decreases the performance levels of a model on data sets not well characterized by the training data set, while poor robustness appears when models under-perform on data sets presenting noise or other corruptions [13]. Modern neural networks have been shown to be highly susceptible to distribution shifts and corruptions that are modality-specific [6]. While the average accuracy of U-Net-based models has increased over the years, it is evident from the literature that their robustness level has not improved at the same rate [4,5,9].

One of the key elements of the U-Net are the skip-connections, which propagate information directly (i.e., without further processing) from the encoding to the decoding branch at different scales. Azad et al. [3] mention that this novel design propagates essential high-resolution contextual information along the network, which encourages the network to re-use the low-level representation along with the high-context representation for accurate localization. Nonetheless, there is no clear evidence supporting this intuition and moreover, there is limited knowledge in the literature describing to what extent skip-connections of the U-Net are necessary, and what their interplay is in terms of model robustness when they are subjected to different levels of task complexity.

Currently, the U-Net is used more as a “Swiss-army knife” architecture across different image modalities and image quality ranges. In this paper, we describe the interplay between skip-connections and their effective role of “transferring information” into the decoding branch of the U-Net for different degrees of task complexity, based on controlled experiments conducted on synthetic images of varying textures as well as on clinical data comprising Ultrasound (US), Computed tomography (CT), and Magnetic Resonance Imaging (MRI). In this regard, the work of [10] showed that neural networks are biased toward texture information. Recently, [19,20] similarly showed the impact of texture modifications on the performance and robustness of trained U-Net models. Contrary to these prior works analyzing the impact of data perturbation to model performance (e.g. [6,13]), in this study we focus on analyzing the role of skip-connections to model performance and its robustness. We hypothesize therefore that skip-connections may not always lead to beneficial effects across varying task complexities as measured with texture modifications. Our major contributions through this paper are:

(i) We describe a novel analysis pipeline to evaluate the robustness of image segmentation models as a function of the difference in texture between foreground and background.

(ii) We confirm the hypothesis that severing these skip-connections could lead to more robust models, especially in the case of out-of-domain (OOD) test data. Furthermore, we show that severing skip-connections could work better than filtering feature maps from the encoder with attention-gating.

(iii) Finally, we also demonstrate failure modes of using skip-connections, where robustness across texture variations appear to be sacrificed in the pursuit of improvements within domain.

2 Materials and Methods

2.1 Experiment Design

Figure 1 describes our experimental setup to assess the impact of skip-connections in U-Net-like architectures under varying levels of task complexity.

Given a set of N pairs of labeled training images $\{(I, S)_i : 1 \leq i \leq N\}, I \in \mathbb{R}^{H \times W}$ and $S \in \mathbb{Z} : \{0, 1\}^{H \times W}$, corresponding ground-truth segmentation, a deep learning segmentation model $M(I) \mapsto S$ is commonly updated by minimizing a standard loss term, such as the binary cross entropy or dice loss. To

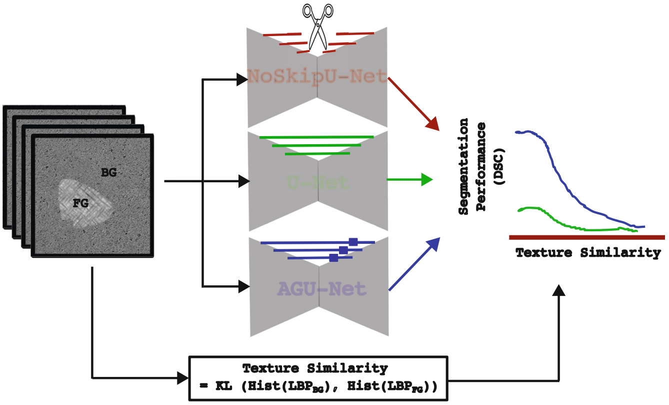
BG
FG
NoSkipU-Net
U-Net
AGU-Net
Segmentation Performance (DSC)
Texture Similarity
Texture Similarity = KL (Hist(LBP$_{BG}$), Hist(LBP$_{FG}$))

Fig. 1. Experimental design to evaluate the role of U-Net's skip-connections under different levels of task complexity. Given training images with controllable background (BG) and foreground (FG) textures, three variants of the U-Net were trained featuring no skip-connections (NoSkipU-Net), standard U-Net (U-Net) [17], and Attention-Gated U-Net (AGU-Net) [18], each characterizing a different strategy (zeroing information through skips, identity transform and filtering information through skips, respectively). Each model was trained with different levels of texture similarity between background and foreground, based on the Kullback-Leibler divergence of Local Binary Pattern (LBP) histograms for foreground and background regions. For each level of foreground-to-background texture similarity, the performance for each model was recorded in-domain, and robustness was measured with out-of-domain texture similarities.

evaluate how the model behaves at varying task complexities, we construct training data sets where each training sample is subjected to a linear transformation where its foreground is blended with the background: $I(x \mid Z(x) = 1) = \alpha I(x \mid Z(x) = 1) + (1 - \alpha)I(x \mid Z(x) = 0)$.

By increasing α from zero to one, more of the foreground texture is added in the foreground mask, which otherwise is made up of the background texture (See Fig. 2), while the background itself is unimpacted. We then quantify the similarity between foreground and background regions by measuring the Kullback-Leibler divergence between their local-binary-pattern (LBP) [16] histograms. We selected LBP since it is a commonly used and benchmarked texture descriptor in machine learning applications [8,15].

$$\mathcal{TS} = KL(\mathcal{H}(\mathcal{L}(I)_{BG})||\mathcal{H}(\mathcal{L}(I)_{FG}) \tag{1}$$

$$\mathcal{L}(I)_{BG} = LBP(I(x \mid Z(x) = 0) \tag{2}$$

$$\mathcal{L}(I)_{FG} = LBP(I(x \mid Z(x) = 1) \tag{3}$$

where $\mathcal{TS}$ refers to the level of texture similarity, $\mathcal{H}()$ corresponds to histogram, and $\mathcal{L}(\mathcal{I})_{\{BG,FG\}}$ refers to LBP calculated for BG or FG. The LBP histogram was computed using a 3×3 neighbourhood with 8 points around each pixel in the image. Three U-Net models were trained featuring three different skip-connection strategies: NoSkipU-Net, U-Net, and AGU-Net, representing the absence of skip-connections, the use of an identity transform (i.e., information through skips is kept as is), and filtering information via attention through skip-connections, respectively. Models were trained at different levels of $\mathcal{TS}$ between the foreground and background regions, determined based on the Kullback-Leibler divergence of Local Binary Pattern (LBP) histograms, Eq. 1. For each level of α used to create a training set, we trained a model to be evaluated on a synthetic test set using the same α to measure within-domain performance and across a range of α, to measure their out-of-domain robustness.

Next, using Eq. 1 and ground truth labels, we computed the $\mathcal{TS}$ of images from the test set of the medical data sets and applied corruptions by way of noise or blurring in order to increase and decrease $\mathcal{TS}$ depending on the imaging modality being analyzed. Then we evaluated the robustness of these models to texture changes in these data sets. We did this at two levels of task complexity (easier - where $\mathcal{TS}$ is higher, and harder, where $\mathcal{TS}$ is lower) and different from the original $\mathcal{TS}$. We report all model performances using dice scores.

2.2 Description of Data

Synthetic Textures: We took two representative grayscale textures from the synthetic toy data set described in [11] and used them as the background and foreground patterns. These patterns were chosen such that the $\mathcal{TS}$ values matched the range of medical data sets described next. We also generated synthetic segmentation masks using bezier curves setup such that the curvature and size of the foreground simulate clinical image segmentation problems. Examples of such images are shown in Fig. 2. We generated 100 such image-mask

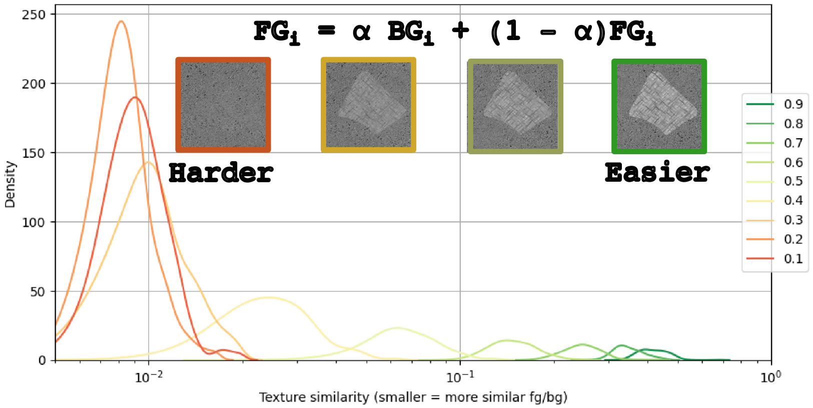

Fig. 2. Generation of synthetic data samples as a function of blending foreground texture into the background. Numbers in the legend indicate the proportion of foreground blended within the foreground mask.

pairs at 9 levels (for $\alpha \in \{0.1, 0.2, ..., 0.9\}$), so that we create training data sets at various task complexities. These images are generated by randomly cropping the grayscale textures to 256 × 256 pixels. 70 of these were used as the training set, 10 were reserved for validation, and the rest of the 20 formed the test set, identically split for all task complexities. Figure 2 show kernel density estimates of each of these 9 data sets along the texture similarity axis. The curve in orange ($\alpha = 0.1$) indicates that the foreground mask in this set contains only 10% of the actual foreground texture and 90% of the background texture blended together. This represents a situation where it is texturally hard for humans as well as for segmentation models. The data set in green ($\alpha = 0.9$) shows the reverse ratio - the foreground region now contains 90% of the foreground texture, thereby making it an easier task to segment.

Medical Data Sets: We tested the three variants of the U-Net architecture on three medical binary segmentation data sets: a Breast Ultrasound [1], a spleen CT and a heart MRI data set [2]. The breast ultrasound data set contained 647 images, 400 of which were used as training, 100 as validation and 147 as the test set. We used the benign and malignant categories in the breast ultrasound data and excluded images with no foreground to segment (i.e. the “normal” category). The spleen data set contained 899 images, 601 of which were used as training, 82 as validation and 216 as test set images. The heart data set contained 829 images, 563 of which were used as training, 146 as validation, and 120 as test set images. We selected 2D axial slices from the spleen, and sagittal slices from the heart data sets, both of which were originally 3D volumes, such that there is at least one pixel corresponding to the foreground. Care was taken to ensure that 2D slices were selected from 3D volumes and split at the patient level to avoid cross-contamination of images across training/test splits.

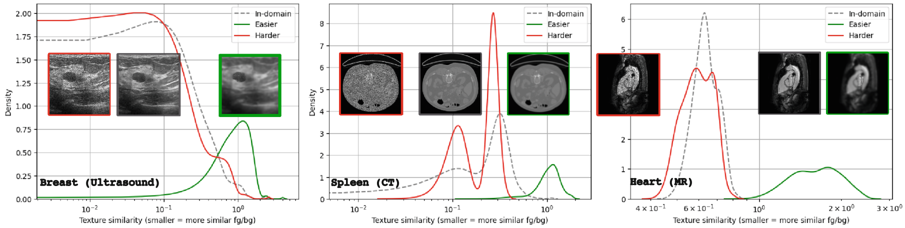

Fig. 3. Medical data test sets on the texture similarity ($\mathcal{TS}$) axis with in-domain (dashed gray), easier task (green, low similarity) and harder task (red, high similarity) distributions. Three modalities tested include US, CT, and MR, whose $\mathcal{TS}$ are in the same range as synthetic data in Fig. 2.

To vary $\mathcal{TS}$ of images in the test set, and to evaluate the robustness of the U-Net variants, speckle noise with variance 0.1 was added to both the foreground and background. This made the textures more similar, hence lowered $\mathcal{TS}$, and essentially rendered them harder to segment. This is shown in the red boxes in Fig. 3. We also created another test set with textures that are less similar by blurring the background using a Gaussian kernel of variance 3.0 while not blurring the foreground pixels. These are shown in the green boxes in Fig. 3, where it can be seen they are easier to segment.

2.3 Model Architecture and Training Settings

The network architectures were created using MONAI [7] v1.1 and were all trained with random weight initialization. The U-Net was implemented with one input and output channel, with input image size set to 256 × 256 pixels across all experiments. The model had five levels with $16, 32, 64, 128, 256$ channels each for synthetic experiments and six levels (an additional level with 512 channels) for medical image experiments, all intermediate channels with a stride of 2. The ReLU activation was used, and no residual units were included. To reduce stochasticity, no dropout was used in any variant.

The NoSkipU-Net was identical to the U-Net except for severed skip-connections. This led to the number of channels in the decoder to be smaller as there is no concatenation from the corresponding encoder level. The AGU-Net was setup to be the same as the U-Net, except with attention gating through the skip-connections.

The training parameters were kept constant across compared models for fair comparison. Our experiments[1] were implemented in Python 3.10.4 using the PyTorch implementation of the adam [14] optimizer. We set the learning rate to be $1e^{-3}$ for synthetic experiments (and $1e^{-2}$ for medical image experiments), maintaining it constant without using a learning rate scheduler. No early stopping criteria were used while training, and all models were allowed to train to 100

[1] Code to reproduce this is at https://github.com/amithjkamath/to_skip_or_not..

epochs. We trained our models to optimize the dice loss, and saved the model with the best validation dice (evaluated once every two epochs) for inference on the test set.

We did not perform any data augmentation that could change the scale of the image content, thereby also changing the texture characteristics. Therefore, we only do a random rotation by 90 degrees with a probability of 0.5 for training, and no other augmentations. We also refrained from fine-tuning hyperparameters and did not perform any ensembling as our study design is not meant to achieve the best possible performance metric as much as it attempts to reliably compare performance across architecture variants while keeping confounding factors to a minimum. We therefore trained each model using the same random seeds (three times) and report the dice score statistics. Training and testing were performed on an NVIDIA A5000 GPU with 24 GB RAM and CUDA version 11.4.

3 Results

3.1 On Synthetic Texture Variants

In-domain (Performance): Figure 4 (left) indicates the relative improvement in dice scores between the three U-Net variants using the NoSkipU-Net as the baseline. To make the interpretation easier, the α value is used as a proxy for $\mathcal{TS}$ on the horizontal axis. It is worth noting that for α values > 0.3, there is negligible difference between the dice score performances of all the U-Net variants, indicating their ability with or without the skip-connections to learn the distributions of the foreground and background textures at that level of task complexity. Below α values of 0.3, the benefits of using attention gating in the skip-connections start to appear. This indicates that the benefit of attention-gating as a function of complexity is non-linear: models do not benefit from skip-connections at lower ranges of task complexity, but at larger ones, filtering the information flowing through the skip connections is important. What is interesting is also how the standard U-Net performance is noisy compared to NoSkipU-Net, indicating that passing through the entire encoder feature map to be concatenated with the decoder feature maps may not always be beneficial.

Out-of-Domain (Robustness): Rows in Fig. 4 (right) includes a heatmap to represent the α values that the model was trained on, and columns correspond to the α value it was tested on. The entries in the matrix are normalized dice score differences between AGU-Net and NoSkipU-Net (comparisons between standard U-Net and NoSkipU-Net show similar trends). The diagonal entries here correspond to the AGU-Net plot in Fig. 4 (left). For α values 0.3 and 0.4 in training and 0.9 on testing (corresponding to an out-of-domain testing scenario), the NoSkipU-Net performs better than the AGU-Net, indicating that there indeed are situations where skip-connections cause more harm than benefit.

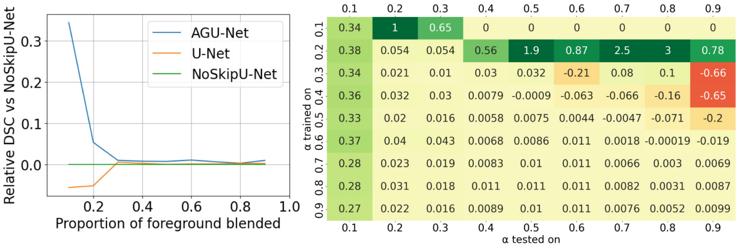

Fig. 4. Relative performance in-domain (left) across U-Net variants, and out-of-domain robustness metrics (right) for AGU-Net versus NoSkipU-Net.

3.2 On Medical Image Textures

In-Domain (performance): Looking at the "In-domain" rows in Table 1, on all three data sets, the AGU-Net outperforms both the other variants. However, the relative improvements in performance vary across modalities, with the performance differences on CT being the most stark. On the Ultrasound data set, the NoSkipU-Net performs as well as the standard U-Net, supporting our hypothesis that skip-connections may not always be beneficial.

Out-of-Domain (Robustness): Focusing on the rows "Harder" and "Easier" in Table 1, we observe for the Ultrasound data set that the AGU-Net improves in the easier task, but declines in performance in the harder one. The drop in performance is most pronounced for the U-Net, but moderate for the NoSkipU-Net. For the spleen data set, both the AGU-Net and the standard U-Net demonstrate severe drop in performance in the harder case. However, AGU-Net is better and

Table 1. Mean (standard deviation) of Dice scores for each of hard, in-domain and easy textures on the Breast (Ultrasound), Spleen (CT) and Heart (MR) data sets. Best performing model at each texture level is highlighted in bold.

Data set	Texture level	AGU-Net	U-Net	NoSkipU-Net
Breast (Ultrasound)	Harder	0.645 (0.291)	0.723 (0.281)	**0.735** (0.268)
	In-domain	**0.795** (0.206)	0.762 (0.261)	0.761 (0.258)
	Easier	**0.799** (0.200)	0.735 (0.244)	0.748 (0.243)
Spleen (CT)	Harder	0.310 (0.226)	0.074 (0.152)	**0.558** (0.265)
	In-domain	**0.927** (0.092)	0.745 (0.275)	0.606 (0.265)
	Easier	**0.809** (0.201)	0.394 (0.354)	0.486 (0.292)
Heart (MRI)	Harder	0.139 (0.242)	0.500 (0.316)	**0.815** (0.126)
	In-domain	**0.929** (0.055)	0.900 (0.080)	0.833 (0.111)
	Easier	**0.889** (0.073)	0.805 (0.129)	0.823 (0.103)

the standard U-Net is worse than the NoSkipU-Net in the easier texture situations. The heart data set shows the same trend as in the spleen data set.

4 Discussion and Conclusion

Through extensive experiments using synthetic texture images at various levels of complexity and validating these findings on medical image data sets from three different modalities, we show in this paper that the use of skip-connections can both be beneficial as well as harmful depending on what can be traded off: robustness or performance. A limitation of our work is that we vary only the foreground in synthetic experiments but background variations could demonstrate unexpected asymmetric behavior. We envision the proposed analysis pipeline to be useful in quality assurance frameworks where U-Net variants could be compared to analyse potential failure modes.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
2. Antonelli, M., et al.: The medical segmentation decathlon. Nature Commun. **13**(1), 4128 (2022)
3. Azad, R., et al.: Medical image segmentation review: the success of u-net. arXiv preprint arXiv:2211.14830 (2022)
4. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
5. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
6. Boone, L., et al.: Rood-MRI: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI. arXiv preprint arXiv:2203.06060 (2022)
7. Cardoso, M.J., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
8. Doshi, N.P., Schaefer, G.: A comprehensive benchmark of local binary pattern algorithms for texture retrieval. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 2760–2763. IEEE (2012)
9. Galati, F., Ourselin, S., Zuluaga, M.A.: From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review. Appl. Sci. **12**(8), 3936 (2022)
10. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
11. Hoyer, L., Munoz, M., Katiyar, P., Khoreva, A., Fischer, V.: Grid saliency for context explanations of semantic segmentation. Adv. Neural Inform. Process. Syst. **32** (2019)

12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: NNU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
13. Kamann, C., Rother, C.: Benchmarking the robustness of semantic segmentation models with respect to common corruptions. Int. J. Comput. Vision **129**(2), 462–483 (2021)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Liu, L., Fieguth, P., Wang, X., Pietikäinen, M., Hu, D.: Evaluation of LBP and deep texture descriptors with a new robustness benchmark. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 69–86. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_5
16. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**(1), 51–59 (1996)
17. 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)
18. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
19. Sheikh, R., Schultz, T.: Feature preserving smoothing provides simple and effective data augmentation for medical image segmentation. In: Martel, A.L., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I, pp. 116–126. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_12
20. You, S., Reyes, M.: Influence of contrast and texture based image modifications on the performance and attention shift of u-net models for brain tissue segmentation. Front. Neuroimag. **1**, 1012639 (2022)