





Evolutionary Normalization Optimization Boosts Semantic Segmentation Network Performance

Luisa Neubig^(✉)  and Andreas M. Kist 

Department Artificial Intelligence in Biomedical Engineering,
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
{luisa.e.neubig, andreas.kist}@fau.de

Abstract. Semantic segmentation is an important task in medical imaging. Typically, encoder-decoder architectures, such as the U-Net, are used in various variants to approach this task. Normalization methods, such as Batch or Instance Normalization are used throughout the architectures to adapt to data-specific noise. However, it is barely investigated which normalization method is most suitable for a given dataset and if a combination of those is beneficial for the overall performance. In this work, we show that by using evolutionary algorithms we can fully automatically select the best set of normalization methods, outperforming any competitive single normalization method baseline. We provide insights into the selection of normalization and how this compares across imaging modalities and datasets. Overall, we propose that normalization should be managed carefully during the development of the most recent semantic segmentation models as it has a significant impact on medical image analysis tasks, contributing to a more efficient analysis of medical data. Our code is openly available at <https://github.com/neuluna/evoNMS>.

Keywords: Semantic segmentation · Normalization · Evolutionary Algorithm

1 Introduction

Semantic segmentation, i.e., assigning a semantic label to each pixel in an image, is a common task in medical computer vision nowadays typically performed by fully convolutional encoder-decoder deep neural networks (DNNs). These DNNs usually incorporate some kind of normalization layers which are thought to reduce the impact of the internal covariate shift (ICS) [6]. This effect describes the adaption to small changes in the feature maps of deeper layers rather than learning the real representation of the target structures [21]. The understanding of Batch Normalization (BN) is very controversial, namely that its actual success is to use higher learning rates by smoothing the objective function instead of reducing the ICS [2]. Instance Normalization (IN), Layer Normalization (LN)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_67.

and Group Normalization (GN) are examples of developments of BN to overcome its shortcomings, like reduced performance using smaller batch sizes [2]. As an alternative, Scaled Exponential Linear Units (SELUs) can act as self-normalization activation functions [9]. All normalization methods have different strengths and weaknesses, which influence the performance and generalizability of the network. Commonly, only a single normalization method is used throughout the network, and studies involving multiple normalization methods are rare.

Neural architecture search (NAS) is a strategy to tweak a neural architecture as such to discover efficient combinations of architectural building blocks for optimal performance on given datasets and tasks [12]. NAS strategies involve, for example, evolutionary algorithms to optimize an objective function by evaluating a set of candidate architectures and selecting the “fittest” architectures for breeding [12]. After several generations of training, selection, and breeding, the objective function of the evolutionary algorithm should be maximized.

In this study, we propose a novel evolutionary NAS approach to increase semantic segmentation performance by optimizing the spatiotemporal usage of normalization methods in a baseline U-Net [17]. Our study provides a uniquely and thorough analysis of the most effective layer-wise normalization configuration across medical datasets, rather than proposing a new normalization method. In the following, we refer to our proposed methodology as evoNMS (evolutionary Normalization Method Search).

We evaluated the performance of evoNMS on eleven biomedical segmentation datasets and compared it with a state-of-the-art semantic segmentation method (nnU-Net [7]) and U-Nets with constant normalization such as BN, IN, and no normalization (NoN). Our analysis demonstrates that evoNMS discovers very effective network architectures for semantic segmentation, achieves better or similar performance to state-of-the-art architectures, and guides the selection of the best normalization method for a specific semantic segmentation task. In addition, we analyze the normalization pattern across datasets and modalities and compare the normalization methods regarding their layer-specific contribution.

2 Related Works

To gain optimal performance in semantic segmentation tasks, it is important to optimize data preprocessing and architectural design. Popat et al. [15] and Wei et al. [19] concurrently developed an evolutionary approach to determine the best-performing U-Net architecture variant, considering the depth, filter size, pooling type, kernel type, and optimizer as hyperparameter-genes coding for a specific U-Net phenotype. When applied to retinal vessel segmentation, both showed that their approach finds a smaller U-Net configuration while achieving competitive performance with state-of-the-art architectures. Liu et al. [10] proved that not only the architecture has a huge impact on the generalizability of a neural network, but also the combination of normalization.

Various studies show that neural networks (NNs) benefit from normalization to enhance task performance, generalizability, and convergence behavior. Zhou et

al. [22] showed the benefit of batch normalization, which focuses on the data bias in the latent space by introducing a dual normalization for better domain generalization. Dual normalization estimates the distribution from source-similar and source-dissimilar domains and achieves a more robust model for domain generalization. Domain-independent normalization also helps to improve unsupervised adversarial domain adaptation for improved generalization capability as shown in [16]. In [21], the authors analyzed the influence of different normalization methods, such as BN, IN, LN, and GN. Although many segmentation networks rely on BN, they recommend using normalization by dividing feature maps, such as GN (with a higher number of groups) or IN [21]. Normalization methods have been well discussed in the literature [3, 5, 9, 18, 20]. In a systematic review, Huang et al. [5] concluded that normalizing the activations is more efficient than normalizing the weights. In addition to the efficiency of normalizing the activations, Luo et al. [13] demonstrated a synergistic effect of their advantages by introducing switchable normalization (SN). SN alternates between the normalization strategies IN, LN, and BN according to their respective importance weights [13]. With an evolutionary approach, Liu et al. [11] also showed that the combination of normalization and activation functions improves the performance of the NN.

3 Methods

We investigated the impact of normalization on semantic segmentation using eleven different medical image datasets. Eight datasets were derived from the Medical Segmentation Decathlon (MSD) [1]. In this study, we selected subsets for segmenting the hippocampus, heart, liver, lung, colon, spleen, pancreas, and hepatic vessels. We only considered the datasets with 3D volumes of the MSD. In addition, we used the BAGLS [4] dataset (segmentation of glottal area), the Kvasir-SEG [8] dataset (gastrointestinal polyp images), and an in-house dataset for bolus segmentation in videofluoroscopic swallowing studies (VFSS). Table 1 lists the datasets used regarding their region of interest/organ, modality, number of segmented classes, and number of images. To minimize the differences between the datasets and to gain comparable results, all datasets were analyzed in 2D. The images were converted to grayscale images, resized, and cropped to a uniform size of 224×224 px. Their pixel intensity was normalized to a range from 0 to 1. The datasets were divided into training, validation, and test subsets, with percentages of 70%, 10%, and 20% for the BAGLS, Kvasir-SEG, and bolus datasets. If a test set was explicitly given, the split for training and validation was 80% and 20%, respectively.

We implemented our evolutionary optimization algorithm and the NN architectural design in TensorFlow 2.9.1/Keras 2.9.0 and executed our code on NVIDIA A100 GPUs. Each individual U-Net variant was trained for 20 epochs using the Adam optimizer, a constant learning rate of 1×10^{-3} , and a batch size of 64. All segmentation tasks were optimized using the Dice Loss (DL). To evaluate the performance of the trained network, we calculated the Dice Coefficient (DC), Intersection over Union (IoU) of the fitted bounding boxes (BBIoU), and Hausdorff Distance with a percentile of 95% (HD95) of the validation set after

Table 1. Overview of different datasets regarding their medical objective, number of segmentation labels, modality, and number of train, validation, test images, and the average evoNMS wall time (Time).

Dataset	ROI	Labels	Modality	Train	Validation	Test	Time
BAGLS	Glottal Area	1	Endoscope	16,277	2,325	4,650	96 h
Kvasir-SEG	Polyp	1	Endoscope	700	100	200	4 h
Bolus	Bolus	1	VFSS	7,931	1,133	2,266	39 h
Task02 MSD	Heart	1	MRI	1,215	135	1,297	7 h
Task03 MSD	Liver	2	CT	17,238	1,915	27,041	102 h
Task04 MSD	Hippocampus	2	MRI	5,960	662	4,499	40 h
Task06 MSD	Lung	1	CT	1,480	164	8,888	12 h
Task07 MSD	Pancreas	2	CT	7,907	878	13,544	49 h
Task08 MSD	Hepatic Vessel	2	CT	11,448	1,272	10,519	75 h
Task09 MSD	Spleen	1	CT	945	105	1,327	12 h
Task10 MSD	Colon	1	CT	1,152	128	6,616	8 h

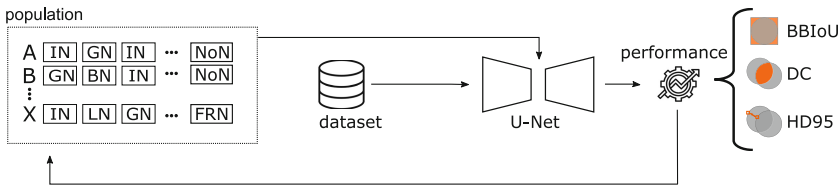


Fig. 1. Process of evolutionary optimization of normalization layers in a U-Net architecture. The sequence of normalization layers was randomly determined for a population of 20 individuals in the first generation. After training each U-Net variant on a given task/dataset, its performance was evaluated using a fitness function based on BBIOU, DC, and HD95. Based on the fitness value, normalization sequences were selected to breed the population of the next generation.

20 epochs. In addition, we included an early stopping criterion that is activated when the validation loss changes less than 0.1% to avoid unnecessary training time without further information gain. To compare our approach to state-of-the-art segmentation networks, we considered nnU-Net [7] as a baseline which was similarly trained as the above-mentioned U-Net variants.

Our proposed evoNMS approach is based on evolutionary optimization with leaderboard selection and is executed for 20 generations. Each generation's population consists of 20 individuals, i.e., U-Net variants, meaning that we train 400 variants for one evoNMS execution (duration 4 h (polyp) to 5 days (glottal area) on one A100 GPU). The first generation contains individuals with random sequences drawn from our gene pool containing either a BN, IN, FRN, GN2, GN4 layer or skips normalization (no normalization, NoN). Other U-Net-specific hyperparameters, such as initial filter size and activation functions were set across datasets to a fixed value (initial filter size of 16, and ReLU as activa-

tion function). In general, we kept all other hyperparameters fixed to focus only on the influence of normalization and infer whether it exhibits decoder/encoder dependence or even dependence on the underlying modality. After training each architecture, the fitness F_i (Eq. (1)) is evaluated for each individual i

$$F_i = \frac{1}{3} \cdot (\text{DC}_i + \text{BBIoU}_i + \frac{1}{\text{HD95}_i}) \quad , \text{ where } \frac{1}{\text{HD95}_i} \in [0, 1], \quad (1)$$

where we compute the mean validation DC, validation IoU of the Bounding Box, and reciprocal of the validation HD95. We use the reciprocal of HD95 to balance the influence of each metric on the fitness value by a value ranging from 0 to 1. After each generation, the top ten individuals with the highest fitness F were bred. To breed a new individual, we selected two random individuals from the top ten candidates and combined them with a randomly selected crossing point across the normalization layer arrays of the two parental gene pools. Next, we applied mutations at a rate of 10%, which basically changes the normalization method of a random position to any normalization technique available in the gene pool.

Table 2. Overview of performance across datasets and normalization configurations. For multiclass segmentation, we define the highest validation DC as the mean across the segmentation labels. Each value is the mean value of five individual runs. For evoNMS, we selected the best normalization pattern w.r.t. the fitness value and trained this configuration five times. The ranking represents the average behavior of each network across the datasets (1-best, 6-worst). The bottom rows show the average behavior of each network across the eleven datasets regarding DC, BBIoU, and HD95 on the validation dataset.

Dataset	Labels	BN	IN	NoN	nnU-Net	Gen1 (Ours)	Gen20 (Ours)
glottal area	1	0.919	0.933	0.005	0.862	0.928	0.931
polyp	1	0.251	0.639	0.442	0.804	0.672	0.704
bolus	1	0.833	0.841	0.000	0.829	0.836	0.838
heart	1	0.016	0.844	0.029	0.910	0.243	0.885
liver	2	0.898	0.932	0.350	0.656	0.918	0.921
hippocampus	2	0.819	0.829	0.503	0.786	0.829	0.829
lung	1	0.701	0.827	0.008	0.672	0.663	0.833
pancreas	2	0.653	0.722	0.031	0.429	0.689	0.695
hepatic vessel	2	0.482	0.721	0.000	0.474	0.718	0.716
spleen	1	0.054	0.786	0.223	0.934	0.920	0.953
colon	1	0.022	0.717	0.074	0.695	0.735	0.741
Ranking		4.64	2.00	5.63	3.82	2.91	1.72
DC (avg)		0.514	0.799	0.151	0.732	0.741	0.823
BBIoU (avg)		0.446	0.770	0.099	0.686	0.691	0.773
HD95 (avg)		185.992	4.121	367.448	7.802	94.599	4.323

4 Results

We first evaluated the influence of different normalization methods on medical-image segmentation. We report the performance of neural architectures defined by our proposed evoNMS, which followed an evolutionary approach to determine the potentially best normalization method for each bottleneck layer, at generations 1 and 20. For each dataset, we evaluated the DC across different baselines of our default U-Net implementation with a fixed normalization method across layers (BN, IN, or NoN) and a state-of-the-art semantic segmentation network (nnU-Net) against our proposed evoNMS approach. Table 2 provides an overview of the mean validation DC as the main performance metric.

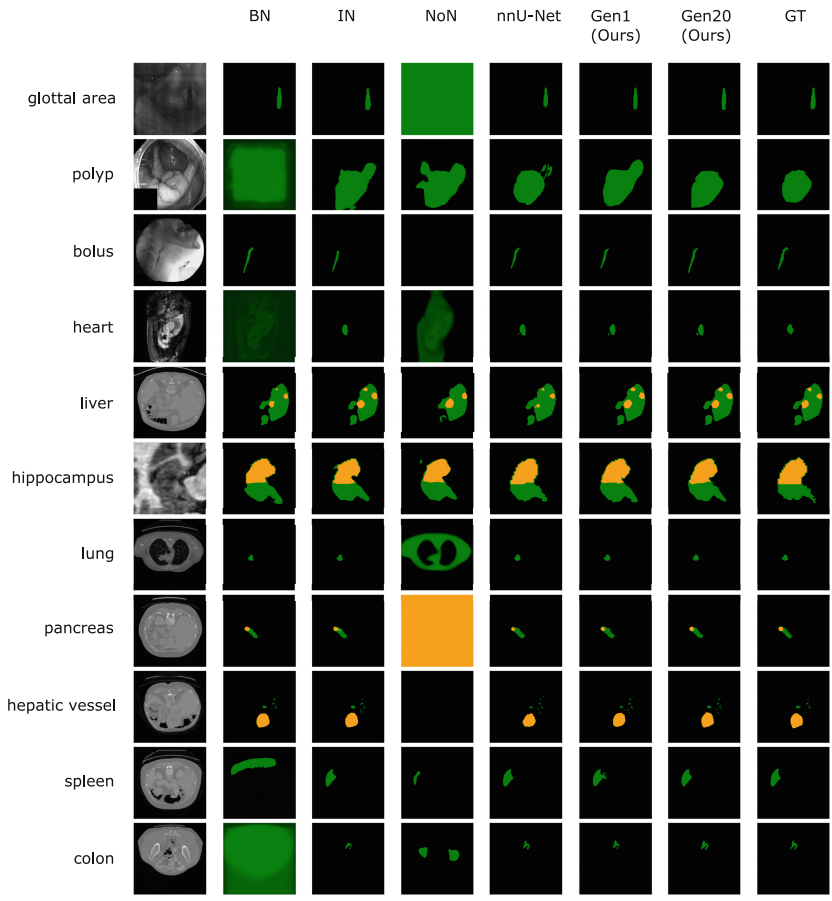


Fig. 2. Qualitative comparison of baseline architectures and the proposed evoNMS approach. Each horizontal set shows exemplary performance for a given dataset.

Overall, the architecture configurations with solely IN (six datasets), nnU-Net (two datasets), or our proposed evoNMS (one first generation, four in the last generation) achieved the highest mean validation DC across all datasets. Noteworthy, our evoNMS approach achieved competitive performance across all datasets, which is clearly shown at the ranking in Table 2, where the last generation of our approach achieved the best grade of 1.72. The results of our evoNMS demonstrate that our approach achieves superior performance in terms of the average DC and BBIOU scores on the validation dataset and yields comparable results in terms of HD95 to the U-Net trained with IN. In contrast, architectures that rely on BN or NoN consistently produced poor results due to exploding gradients across multiple datasets ([14]), questioning the broad application of BN. When comparing qualitative results of all baselines and evoNMS, we find that our approach accurately reflects the ground truth across datasets, especially after several generations of optimization (Fig. 2). We found that an evolutionary search without prior knowledge of the required hyperparameters and properties of the medical data can perform as well as, or in some cases, even better than, the best baseline of a U-net trained with IN or nnU-Net, showing the importance of normalization in semantic segmentation architectures.

We next were interested in the optimization behavior of evoNMS. We found that random initialization of normalization methods yielded poorly and highly variant converging behavior overall (Supplementary Fig. 1). However, after evolutionary optimization, the converging behavior is clearly improved in terms of convergence stability and lower variability. In Supplementary Fig. 2, we can exemplary show that evoNMS also improves all fitness-related metrics across generations. This highlights the ability of evoNMS to converge on multiple objectives. These findings suggest that our approach can also be used as a hyperparameter optimization problem to improve convergence behavior.

As we initialize the first population randomly, we determined whether the evoNMS algorithm converges to the same set of normalization methods for each dataset. We found for three independent evoNMS runs, that evoNMS converges for four out of eleven datasets on very similar patterns (Supplementary Fig. 3) across runs, with an overall average correlation of 36.3%, indicating that our algorithm is able to find relatively quickly a decent solution for a given dataset. In the case of the polyp dataset, top evoNMS-performing networks correlate with a striking 61.1% (Supplementary Table 1).

We next identified the final distribution of normalization methods across the encoder and decoder U-Net layers to determine dataset-specific normalization. In Fig. 3, we show the distribution of the top 10% performers in the last generation of evoNMS. We found consistent patterns across individuals especially in the last layer: in four out of eleven datasets, no normalization was preferred. In the colon dataset, the encoder was mainly dominated by IN, especially in the first encoding layer. In contrast, the decoder in the polyp, liver, and hippocampus datasets showed more consistency in the normalization methods suggesting that normalization methods could be encoder- and decoder-specific and are dataset dependent.

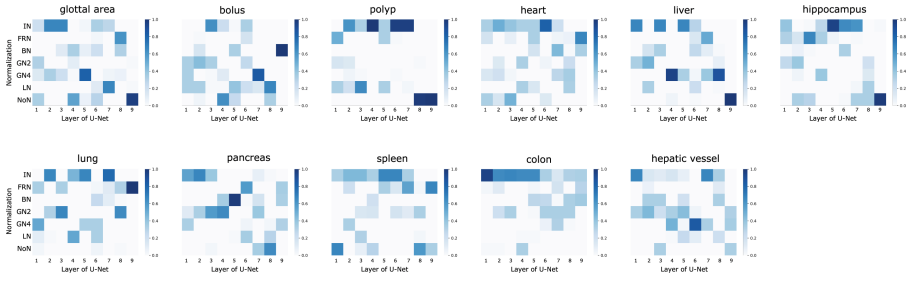


Fig. 3. Visualization of the relative frequency of selected normalizations over the U-Net layer for all datasets. This average was calculated for all individuals in the last generation of each dataset.

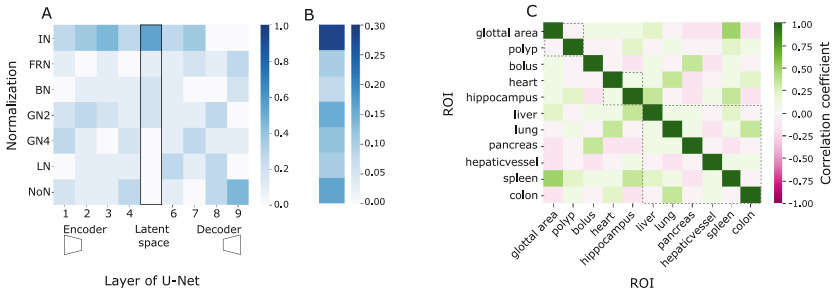


Fig. 4. Visualization of the best normalization pattern in the last generation and the selection for each individual. A) Distribution of normalizations across U-Net layers and datasets. B) Relative frequency of normalization patterns. C) shows the normalization pattern correlation of U-Net variants with the highest fitness F for each dataset in the last generation. The boxes indicated the same imaging modality.

To understand good normalization design across datasets in semantic segmentation tasks, we determined which normalization methods are more abundant at specific U-Net layers. IN is across layers a preferred choice by evoNMS except for the last two decoding layers (Fig. 4 A and B). Our results suggest that the latent space embedding heavily relies on IN (Fig. 4 A). Other normalization methods are less abundant in top-performing architectures, such as FRN, BN, and LN. However, NoN is mainly used in the last layer. FRN and especially BN seem to be inferior choices in semantic segmentation architectures.

Finally, we investigated if any set of normalization methods can be derived by the imaging modality, such as endoscopy, CT and MRI. In Fig. 4, we show the cross-correlation of all evoNMS top performers across datasets. In general, there are only weak correlations at the global level; a stronger correlation can be seen by correlating encoder and decoder separately (Supplementary Fig. 3). These results provide evidence, that *a priori* knowledge of the imaging modality does not hint towards a specific normalization pattern but rather has to be optimized for any given dataset.

5 Discussion and Conclusion

Our approach shows that using normalization methods wisely has a powerful impact on biomedical image segmentation across a variety of datasets acquired using different imaging modalities. Due to its inherent property of always finding an optimal solution, evoNMS is potentially capable of providing the best-performing set of normalization patterns for any given data set. For feasibility reasons, we considered only 20 epochs for each training set and 20 generations. However, we show that evoNMS with these constrained settings provides competitive results and outperforms or performs on par compared to all baselines.

Our results suggest the superior performance of IN and GN (Fig. 4) as overall very useful normalization strategies when used at their preferred location, in contrast to FRN, BN, and LN. State-of-the-art architectures, such as nnU-Net also rely on IN throughout the network [7], as well as evolutionary optimized U-Net architectures [19]. This established use of normalization confirms our findings for high-performing evoNMS networks and can be extrapolated to other semantic segmentation architectures that incorporate normalization methods. The advantage of evoNMS is its ability to also include non-learnable objectives, such as architectural scaling and reducing inference time, crucial for point-of-care solutions. In this study, we constrained the search space, but we will incorporate multiple hyperparameters in the future. On the other hand, approaches that use main building blocks and optimize mainly convolutional layer parameters and activation functions [15, 19] would benefit from incorporating normalization strategies as well.

References

1. Antonelli, M., Reinke, A., Bakas, S., et al.: The medical segmentation decathlon. *Nat. Commun.* **13**(1), 4128 (2022). <https://doi.org/10.1038/s41467-022-30695-9>
2. Awais, M., Iqbal, M.T.B., Bae, S.H.: Revisiting internal covariate shift for batch normalization. *IEEE Trans. Neural Networks Learn. Syst.* **32**(11), 5082–5092 (2021). <https://doi.org/10.1109/tnnls.2020.3026784>
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016). <https://doi.org/10.48550/ARXIV.1607.06450>
4. Gómez, P., Kist, A.M., Schlegel, P., Berry, D.A., Chhetri, D.K., Dórr, S., Echter-nach, M., Johnson, A.M., Kniesburges, S., Kunduk, M., et al.: Bagls, a multihospital benchmark for automatic glottis segmentation. *Sci. Data* **7**(1), 186 (2020). <https://doi.org/10.1038/s41597-020-0526-3>
5. Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., Shao, L.: Normalization techniques in training DNNs: methodology, analysis and application (2020). <https://doi.org/10.48550/ARXIV.2009.12836>
6. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift (2015). <https://doi.org/10.48550/ARXIV.1502.03167>
7. Isensee, F., Jaeger, P.F., Kohl, S.A.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 (2020). <https://doi.org/10.1038/s41592-020-01008-z>

8. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
9. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). <https://proceedings.neurips.cc/paper/2017/file/5d44ee6f2c3f71b73125876103c8f6c4-Paper.pdf>
10. Liu, C., et al.: Auto-DeepLab: hierarchical neural architecture search for semantic image segmentation, pp. 82–92 (2019). <https://doi.org/10.1109/CVPR.2019.00017>
11. Liu, H., Brock, A., Simonyan, K., Le, Q.: Evolving normalization-activation layers. **33**, 13539–13550 (2020). <https://doi.org/10.48550/arXiv.2004.02967>
12. Liu, Y., Sun, Y., Xue, B., Zhang, M., Yen, G.G., Tan, K.C.: A survey on evolutionary neural architecture search. IEEE Trans. Neural Networks Learn. Syst. **34**(2), 550–570 (2023). <https://doi.org/10.1109/TNNLS.2021.3100554>
13. Luo, P., Ren, J., Peng, Z., Zhang, R., Li, J.: Differentiable learning-to-normalize via switchable normalization (2018). <https://doi.org/10.48550/ARXIV.1806.10779>
14. Philipp, G., Song, D., Carbonell, J.G.: The exploding gradient problem demystified - definition, prevalence, impact, origin, tradeoffs, and solutions (2017). <https://doi.org/10.48550/ARXIV.1712.05577>
15. Popat, V., Mahdinejad, M., Cedeño, O., Naredo, E., Ryan, C.: GA-based U-Net architecture optimization applied to retina blood vessel segmentation. In: Proceedings of the 12th International Joint Conference on Computational Intelligence. SCITEPRESS - Science and Technology Publications (2020). <https://doi.org/10.5220/0010112201920199>
16. Romijnders, R., Meletis, P., Dubbelman, G.: A domain agnostic normalization layer for unsupervised adversarial domain adaptation, pp. 1866–1875, January 2019. <https://doi.org/10.1109/WACV.2019.00203>
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Singh, S., Krishnan, S.: Filter response normalization layer: eliminating batch dependence in the training of deep neural networks (2019). <https://doi.org/10.48550/ARXIV.1911.09737>
19. Wei, J., et al.: Genetic U-Net: automatically designed deep networks for retinal vessel segmentation using a genetic algorithm. IEEE Trans. Med. Imaging **41**(2), 292–307 (2022). <https://doi.org/10.1109/TMI.2021.3111679>
20. Wu, Y., He, K.: Group normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_1
21. Zhou, X.Y., Yang, G.Z.: Normalization in training U-Net for 2-D biomedical semantic segmentation. IEEE Robot. Autom. Lett. **4**(2), 1792–1799 (2019). <https://doi.org/10.1109/lra.2019.2896518>
22. Zhou, Z., Qi, L., Yang, X., Ni, D., Shi, Y.: Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20824–20833 (2022). <https://doi.org/10.1109/CVPR52688.2022.02019>