Probabilistic Image Diversification to Improve Segmentation in 3D Microscopy Image Data

Dennis Eschweiler[1⋆], Justus Schock[1], and Johannes Stegmaier[1]

[1] Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany {dennis.eschweiler,johannes.stegmaier}@lfb.rwth-aachen.de

Abstract. The lack of fully-annotated data sets is one of the major limiting factors in the application of learning-based segmentation approaches for microscopy image data. Especially for 3D image data, generation of such annotations remains a challenge, increasing the demand for approaches making most out of existing annotations. We propose a probabilistic approach to increase image data diversity in small annotated data sets without further cost, to improve and evaluate segmentation approaches and ultimately contribute to an increased efficacy of available annotations. Different experiments show utilization for benchmarking, image data augmentation and test-time augmentation on the example of a deep learning-based 3D segmentation approach.

Keywords: Augmentation · Segmentation · 3D Microscopy

1 Introduction

In recent years, automated machine learning-based and deep learning-based approaches have become the state-of-the-art for segmentation in biomedical image data [3]. Since these approaches often require large amounts of annotated training data to become generalist, there is a high demand for fully-annotated image data sets. Due to the very costly, time-consuming and, especially for 3D image data, often infeasible creation of annotations, fully-annotated data sets are rarely available. To overcome this limitation and increase the efficacy of existing annotated data sets, different types of augmentations (such as, *e.g.*, rotation, translation, noise injection) can be applied to existing image data, increasing image diversity and, therefore, segmentation robustness without the need to acquire further expensive annotations [5]. Cost-free specialization to specific image domains or data sets can be achieved by using augmentations in an autoencoder-based pretraining [8] and resilience towards data diversity during inference can be gained by using test-time augmentation techniques [4].

Image augmentation approaches, however, often enrich image data sets with overly artificial modifications that do not represent the real diversity found in image data sets. This includes, *e.g.*, inpainting, local-shuffling, non-linear transforms or strong additive noise. Although those methods have been proven to

⋆ This work was funded by the German Research Foundation DFG with the grant STE2802/2-1 (DE)

enhance segmentation performance by encouraging robustness to strong image alterations [8], we want to propose a simple approach that focuses on transforming image intensities using real image data statistics for increased segmentation robustness and efficacy of annotated image data sets. We envision that this approach encourages robustness to the diversity present in real image data, while being complementary to existing methods. Our approach relies on basic local image intensity statistics and allows to alter the appearance of existing image data in a controlled and realistic way, virtually without any cost. Moreover, altered image data can be generated on-the-fly, once local intensity statistics have been determined in a preprocessing step. We test the usability of this approach on different experiments, including benchmarking, image data augmentation and test-time augmentation.

2 Probabilistic Image Diversification

In the proposed approach, each image voxel is represented as a distribution of possible image intensities, rather than a single fixed scalar, which allows to randomly create new samples from a range of appearances. In contrast to image simulation approaches, the distribution representation does not aim at modelling the image formation and degradation process, but it rather determines the range in which a specific image intensity can be realistically changed, given a predefined neighborhood region. Therefore, a standard normal distribution is re-parametrized to represent intensities, following

$$\tilde{I}(\mathbf{x}) = \mathcal{N}(0,1) \cdot w_v \cdot \sigma_{\mathbf{x}} + w_i \cdot I_{\mathbf{x}}, \tag{1}$$

with $\mathcal{N}(0,1)$ being the standard normal distribution, $\mathbf{x} = (x, y, z)$ representing a given voxel position, and $I_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ being the intensity value and determined variance statistics for position $\mathbf{x}$. Both, intensity and variance are weighted by global scalar values w_i and w_v, respectively, to be able to influence the determined intensity statistics. The variance is estimated from a small cubic neighborhood with size s_σ centered at position $\mathbf{x}$, while $I_{\mathbf{x}}$ represents the original voxel intensity at position $\mathbf{x}$. This serves two purposes, since it allows for the aforementioned realistic change of intensity values rather than modelling it, and it introduces a neutral element to this concept with $w_i = 1$ and $w_v = 0$. The definition of a neutral element would not be possible with a mean intensity value estimated from a given neighborhood and it helps to straightforwardly and smoothly control the degree of abstraction of the applied diversification.

During application, a fixed set of global statistic weights $\mathbf{w} = (w_i, w_v)$ is chosen and for each voxel $\mathbf{x}$ a random intensity is drawn from the assigned distribution $\tilde{I}(\mathbf{x})$. The intensity weight w_i allows to control the brightness of the resulting image, while the variance weight w_v gives control over the noise content within the image, as shown in Fig. 1. To preserve the normalized intensity value range, the sampled image is value-clipped to (0,1). Once the variance statistics have been determined in a preprocessing step and are available during application, transformation of a given image can be performed fast and on-the-fly, since intensities only have to be drawn from the assigned distributions.

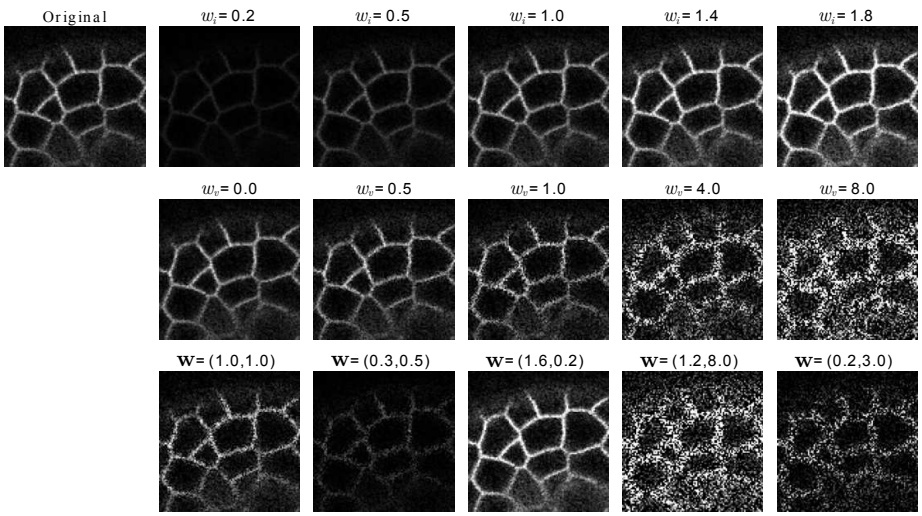

Fig. 1. Examples of an image slice from a publicly available data set [7] (left), and diverse appearances obtained with the proposed method. The upper row shows results for different intensity weights w_i, while $w_v = 0$ is kept neutral. The middle row shows results for different variance weights w_v, while $w_i = 1$ is kept neutral. The lower row shows random examples when jointly changing the statistic weights $\mathbf{w} = (w_i, w_v)$.

3 Experiments and Results

We experimented with different applications of the proposed method, which range from benchmarking and data augmentation to test-time augmentation approaches. With these experiments we demonstrate the range of usability, while we note that the method is not limited to this choice of applications. We used a publicly available annotated data set of 3D image stacks showing fluorescently labeled cell membranes in *A. thaliana* [7], which was split into plants 2, 4, 8 and 13 for training, and plants 15 and 18 for testing. Furthermore, to impose a more complex challenge, we used the same test data set (plants 15 and 18) with synthetically decreased quality, which was published in [1]. As segmentation approach we employed a 3D extension [2] of the Cellpose instance segmentation approach [6] and used the intersection-over-union (IoU) score as a metric to assess the segmentation quality. Models were trained for 500 epochs with patches of size (128, 128, 64). For each experiment we extracted variance statistics as a preprocessing step, allowing to use the proposed method on-the-fly. Extracting variance statistics from larger neighborhood sizes s_σ, resulted in decreasing peak signal-to-noise ratios (PSNR) and correlation scores between the sampled image and the original image, exemplary assessed for default weights $w_i = 1$ and $w_v = 1$ (Fig. 2). We empirically chose $s_\sigma = 5$ as a reasonable trade-off between loss of similarity to the original image and size of the influencing neighborhood, *i.e.*, significance of the estimated intensity variance $\sigma_{\mathbf{x}}$. Consequently, the intensity

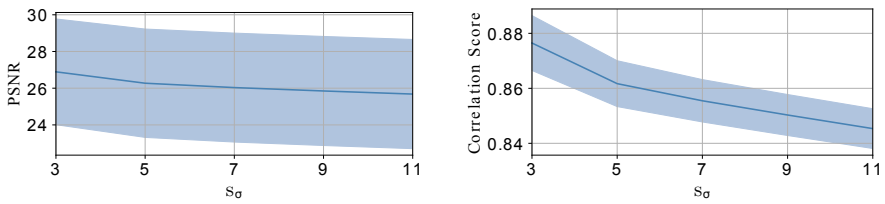

Fig. 2. Peak signal-to-noise ratio (PSNR, left) and correlation score (right) for different neighborhood extents s_σ used to determine the local intensity variance $\sigma_{\mathbf{x}}$. Default weights $w_i = w_v = 1$ were used to augment images from a publicly available data set [7], and shaded regions show the score standard deviation.

variance for each voxel was estimated from 125 local intensity values for the used 3D image data set.

3.1 Data Augmentation

In order to render segmentation approaches more robust to difficult image characteristics, such as low PSNR and low intensity (or a combination of both), the proposed diversification can be utilized as data augmentation method. Therefore, we designed three different experimental setups, which range from a straightforward augmentation to a curriculum learning strategy. To assess more details in the evaluation of the obtained instance segmentations, IoU thresholds determined if a segmentation was accurately predicted or not. This allows to calculate an accuracy score for multiple IoU thresholds, ultimately leading to insights into how segmentations of different precision were influenced by the proposed approaches. Details of each experiment are explained in the following.

Single Augmentation Each training patch was augmented using the proposed method, by randomly choosing statistic weights $\mathbf{w} = (w_i, w_v)$. During application, weights were limited to predefined ranges and individually drawn from a normal distribution centered around the respective neutral elements ($w_i = 1$, $w_v = 0$). Furthermore, the normal distribution was scaled, such that the weight limit furthest away from the neutral element was located at the distance that matches three times the standard deviation. This is followed by a truncation of the distribution to the predefined, potentially asymmetric, weight range, to prevent drawing of outliers. Different experiments with multiple weight limits were conducted, with results being shown in Fig. 3.

Stacked Diversification Since the proposed method transforms intensities in a given image without imposing geometrical changes, the ground truth segmentation mask remains unchanged. In this experiment we wanted to exploit this fact and create diverse appearances of the training patch. The resulting stack

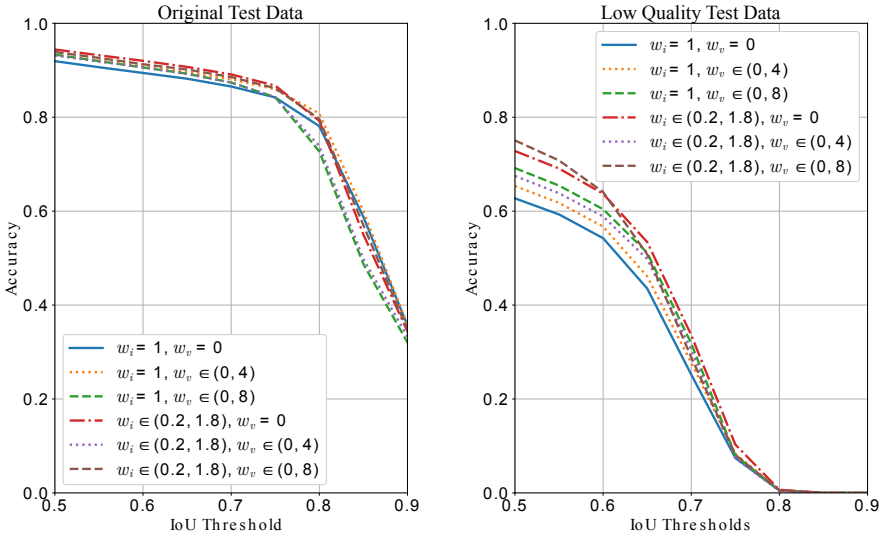

Fig. 3. Obtained accuracy over different IoU thresholds for the single augmentation experiment. Results are shown for the original (left) and low quality (right) test data.

included the original patch and a predefined number n_{div} of diversified patch appearances, leading to a total number of $b = n_{div} + 1$ patches, while the statistic weights $\mathbf{w} = (w_i, w_v)$ were chosen with two different strategies. As a first experiment, $w_i \in [0.2, 1.8]$ and $w_v \in [0, 8]$ were randomly and uniformly drawn from the specified value ranges, which we refer to as *random strategy*. As another experiment, $w_i \in [0.2, 1.8]$ and $w_v \in [0, 8]$ were selected, such that the resulting appearance stack included patches transformed by the full range of possible weight values, *i.e.*, the weight limits and uniformly distributed weights in between, depending on the predefined number of total appearances. In the special case when only one additional patch is desired ($b = 2$), the weights were fixed to default values $w_i = w_v = 1$. We refer to this experiment as *structured strategy*. In addition to the usual loss, comparing prediction and ground truth target for each patch, we introduced a robustness loss to both strategies, assessing consistency among predictions for the diverse patch appearances during training. Therefore, the prediction of an augmented patch was additionally compared to the prediction obtained for the previous input patch within the stack, starting with the original version. Since the resulting stack of patches effectively increases the batch size to b, we conducted a *baseline* experiment by simply increasing the batch size without using any transformations. Results for all experiments on the original and low quality test data are shown in Fig. 4.

Curriculum Learning By choosing the value ranges for the statistic weights $\mathbf{w} = (w_i, w_v)$, the abstraction of appearances can be smoothly controlled. In

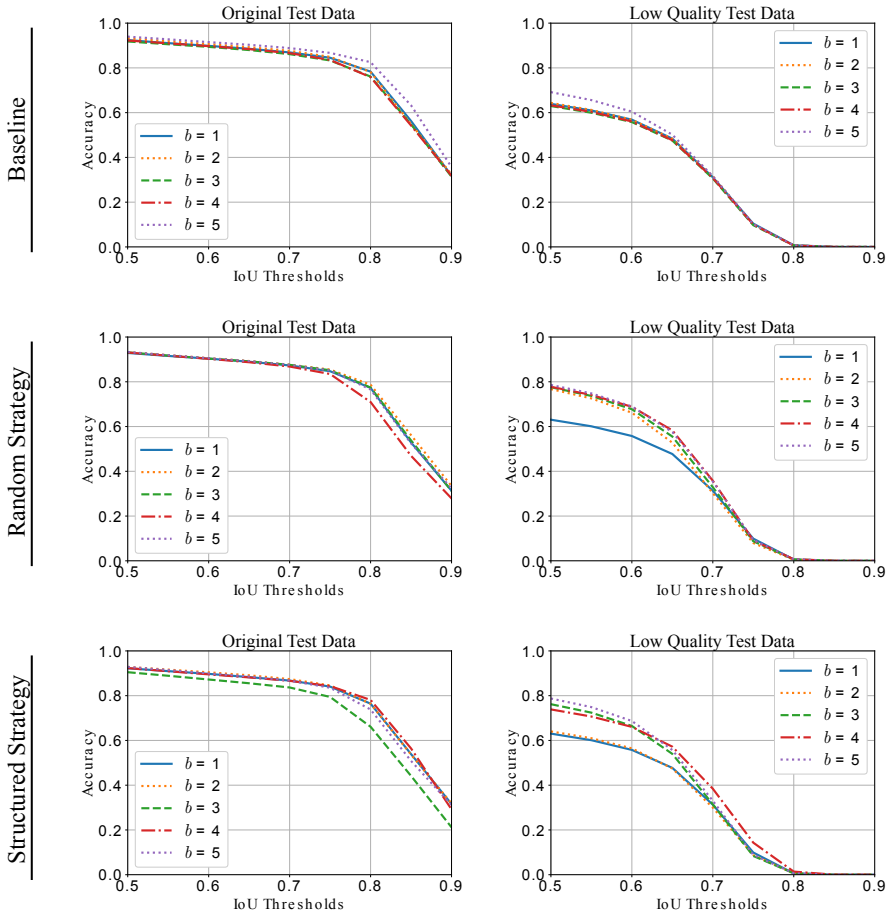

Fig. 4. Obtained accuracy over different IoU thresholds for the stacked diversification experiments, including the *baseline* setup (top), *random strategy* (middle) and *structured strategy* (bottom). Results are shown for the original (left) and low quality (right) test data sets. Parameter b denotes the effective number of patches per batch.

order to guide a segmentation model to be more robust towards challenging appearances, the weight ranges can be continuously increased during training. This allows to start training with the original image diversity of the given data set, while adding more complex challenges as the training progresses. To configure this process we defined a parameter δ_{full}, indicating the number of epochs until the maximum weight range should be reached. Within this interval, the weights are continuously moved away from the neutral element to the predefined limits. This experiment was limited to the *random strategy*, since this proved to reliably increase robustness of the segmentation model. Results obtained for the original and low quality test data using different δ_{full} are shown in Fig. 5. The case $\delta_{full} = 0$ equals the regular random strategy.

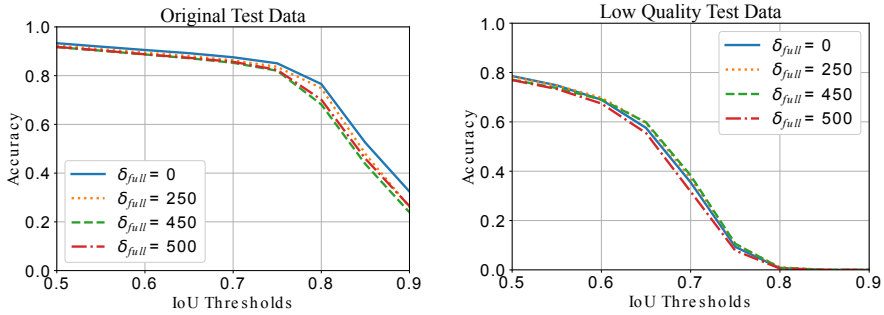

Fig. 5. Obtained accuracy over different IoU thresholds for the curriculum learning experiment. Results are shown for the original (left) and low quality (right) test data sets. Parameter δ_{full} denotes the number of epochs to reach the full weight range.

3.2 Benchmarking

The proposed weights w_i and w_v allow to influence the determined local intensity statistics, to ultimately generate qualitatively different appearances of the same image. This can be utilized to generate benchmark data sets, that allow to assess the sensitivity of an approach to data with different characteristics, such as decreasing PSNR (Fig. 6, top). For demonstration, we trained the 3D Cellpose extension [2] as a baseline experiment on the original training set and applied it to the original test data split, which was transformed with increasing w_v (Fig. 6, bottom). Furthermore, models trained with the previously mentioned strategies (single augmentation, stacked diversification and curriculum learning) were applied to the same transformed test set (Fig. 6, bottom). Although we note that these results are biased, due to using the same transformations during training and for generation of the test set, these experiments demonstrate the improvements gained from those strategies and the possibility to generate

benchmark data. This benchmark concept can be extended by including changes in intensity or a combination of both, changes in intensity and PSNR.

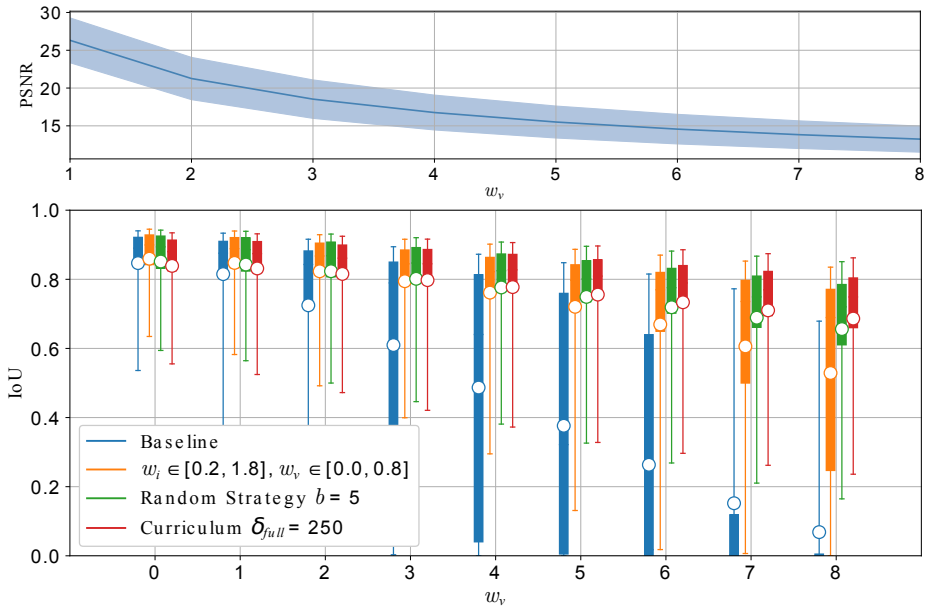

Fig. 6. PSNR (top) and IoU scores of obtained segmentations (bottom) for images augmented with increasing w_v. Segmentation approaches include the baseline approach without using any augmentations, the single augmentation approach with $w_i \in [0.2, 1.8]$ and $w_v \in [0, 8]$, the stacked diversification approach using the *random strategy* with $b = 5$ and the curriculum approach with $\delta_{full} = 250$. IoU scores are shown as boxplots with whiskers showing the 5th and 95th percentile, boxes indicating the inter-quartile range, and mean scores being highlighted as circles.

3.3 Test-time Augmentation

Different local intensity characteristics impose different challenges for segmentation approaches. Since the proposed method aims at diversifying image characteristics, we experimented with a test-time augmentation strategy to offer a variety of challenges, which in combination help to find an overall robust segmentation result. During inference of the segmentation model that was trained with the original training data split, each test patch was diversified into n_{tta} different appearances. Weights $w_i \in [0.2, 1.8]$ and $w_v \in [0, 8]$ were chosen similar to the previously mentioned *structured strategy*, *i.e.*, they included the weight limits and uniformly distributed values in between. The final consensus was computed as an average of the raw network outputs, and the postprocessing to obtain

instance segmentations (as explained in [2]) was performed using the final aggregated prediction. Experimental results for different n_{tta} are shown in Fig. 7.

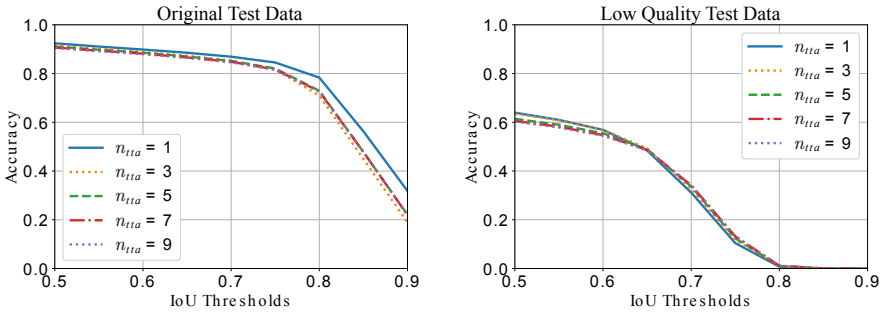

Fig. 7. Obtained accuracy over different IoU thresholds for the test-time augmentation experiment. Results are shown for the original (left) and low quality (right) test data sets. Parameter n_{tta} denotes the number of different appearances processed to generate the averaged output. For $n_{tta} = 1$ only the original appearance was used.

4 Discussion and Conclusion

All experiments showed how local image intensity statistics can be utilized to alter the appearance of image data in a realistic and controlled way. Altering intensity values and variance statistics as augmentation strategy increased robustness to challenging image regions and, specifically, improved inaccurate segmentations as shown for the low quality test data set (Fig. 3, 4, 5). Already accurate segmentations obtained on the original test data set were only slightly influenced, while we note that the data diversity in the original test data is similar to the diversity in the training data set. Consequently, adding more diversity to the training set prevented the model from specializing to the original test set diversity, in exchange for increased robustness. All proposed augmentation strategies, including single augmentation, stacked augmentation and curriculum learning, proved to outperform the baseline results, specifically on the low quality test data set. For the special case of $b = 2$ for the structured strategy in the stacked diversification experiment (Fig. 4, lower right), the original patch appearance and a patch transformed with default parameters $w_i = w_v = 1$ were used. Since the obtained results match those obtained for the baseline experiment, we claim this as evidence that the default case of the proposed transformation creates realistic image appearances similar to the original image data. Including a curriculum strategy to the training process, however, did not improve results upon the stacked augmentation training strategy (Fig. 5), but it improved upon the baseline results and is valued as an incentive for further research.

Furthermore, we demonstrated that statistic weights can be smoothly altered to, *e.g.*, continuously decrease the PSNR of images (Fig. 6, top). Although we note, that using the same transformation strategy for the training and test data leads to biased results, we interpret the results obtained for the test data with decreasing PSNR (Fig. 6, bottom) as evidence that the proposed transformation method can be used to generate realistic benchmark data sets to assess the robustness of automated approaches towards certain challenges. All tested augmentation strategies proved to outperform the baseline experiment, most significantly for image data with low PSNR. Moreover, the curriculum strategy performed best on the benchmark data, although it did not show further improvements on the previous experiments. Due to the smooth and interpretable control of appearances, benchmark data sets can be adjusted to test automated approaches for problem-tailored and realistic challenges. As a last application example, we experimented with test-time augmentation, which, however, did not have a large impact on the final results, but it, nevertheless, states another useful application case up for further research.

In conclusion, we demonstrated with different applications, that the proposed transformation allows to realistically increase image data diversity to enhance the efficacy of available annotations and that it helps to assess robustness of automated approaches to realistic image data challenges. Since the transformation does not come with further annotation costs, we envision that it is straightforwardly applicable to other image processing tasks besides automated segmentation and we will further test applicability to different data sets.

References

1. Eschweiler, D., Rethwisch, M., Jarchow, M., Koppers, S., Stegmaier, J.: 3D Fluorescence Microscopy Data Synthesis for Segmentation and Benchmarking. PLOS One **16**(12), e0260509 (2021)
2. Eschweiler, D., Stegmaier, J.: Robust 3D Cell Segmentation: Extending the View of Cellpose. In: IEEE International Conference in Image Processing (2022)
3. Meijering, E.: A Bird's-Eye View of Deep Learning in Bioimage Analysis. Computational and Structural Biotechnology Journal **18**, 2312 (2020)
4. Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., Horvath, P.: Test-time Augmentation for Deep Learning-based Cell Segmentation on Microscopy Images. Scientific reports **10**(1), 1–7 (2020)
5. Shorten, C., Khoshgoftaar, T.M.: A Survey on Image Data Augmentation for Deep Learning. Journal of Big Data **6**(1), 1–48 (2019)
6. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: A Generalist Algorithm for Cellular Segmentation. Nature Methods **18**(1), 100–106 (2021)
7. Willis, L., Refahi, Y., Wightman, R., Landrein, B., Teles, J., Huang, K.C., Meyerowitz, E.M., Jönsson, H.: Cell Size and Growth Regulation in the Arabidopsis Thaliana Apical Stem Cell Niche. Proceedings of the National Academy of Sciences **113**(51), E8238–E8246 (2016)
8. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models Genesis. Medical Image Analysis **67**, 101840 (2021)