S^3-TTA: SCALE-STYLE SELECTION FOR TEST-TIME AUGMENTATION IN BIOMEDICAL IMAGE SEGMENTATION

Kangxian Xie[1,2*] *Siyu Huang*[3] *Sebastian Cajas Ordonez*[4] *Hanspeter Pfister*[4] *Donglai Wei*[2]

1. Boston University 2. Boston College 3. Clemson University 4. Harvard University

ABSTRACT

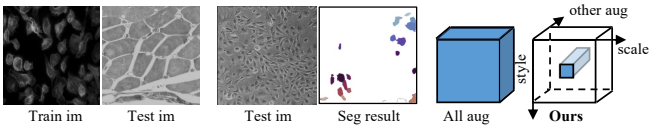
Deep-learning models have been successful in biomedical image segmentation. To generalize for real-world deployment, test-time augmentation (TTA) methods are often used to transform the test image into different versions that are hopefully closer to the training domain. However, due to the diversity of instance scale and styles, many augmented test images produce undesirable results, lowering the overall performance. This work proposes a new TTA framework, S^3-TTA, which selects the suitable image scale and style for each test image based on a transformation consistency metric. In addition, S^3-TTA constructs an end-to-end augmentation-segmentation joint-training pipeline to ensure a task-oriented augmentation. On public benchmarks for cell and lung segmentation, S^3-TTA demonstrates improvements over the prior art by 3.4% and 1.3%, respectively, by simply augmenting the input data in testing phase. Code and models are available at `https://github.com/kangxian97/S3-TTA`.

***Index Terms*—** Test-time augmentation, Style transfer, Instance segmentation, Biomedical images.

1. INTRODUCTION

Segmentation is central to biomedical image analysis [1], which generates object masks, *e.g.*, cell or lung segments, for downstream statistical and morphological analysis. However, learning-based methods often do not generalize well to unseen test images due to their mismatch in morphology and appearance with training ones. To improve domain generalization and prevent test performance degradation, test-time augmentation (TTA) is a popular approach [2, 3] to adapt the segmentation model to different domains.

However, due to the vast diversity of biomedical images [4], a mismatched augmentation in style or scale can lead to poor segmentation results (Fig. 1a-b). Thus, previous approaches suffer from aggregating results from all versions of the augmented test image (Fig. 1c). In this work, we propose a new TTA framework, scale-style selection for test-time augmentation (**S^3-TTA**), to automatically select the suitable image scale and style for each test image. S^3-TTA constructs a task-oriented, augmentation-segmentation joint-training pipeline for image segmentation. A scale-style selection unit adopts a self-consistency metric to select the best augmentated image as an intermediate feature suited for the following segmentation. We evaluate S^3-TTA on public benchmarks of two biomedical domains, *i.e.*, cell and lung segmentation, to demonstrate that it achieves the state-of-the-art performance.



(a) Mismatched style (b) Mismatched scale (c) TTA aggregation

Fig. 1: **Scale-and-style-aware TTA.** (a-b) Due to the diversity of biomedical images, a pre-trained segmentation model may fail significantly if the test image has an unexpected style or scale. (c) Thus, instead of aggregating over all augmentations, we propose to select the suitable style and scale before the aggregation.

1.1. Related Works

Biomedical image segmentation. Mask-RCNN [5] takes the detect-to-segment approach and achieves decent performance for cells with simple shapes [3, 1]. For more complicated cell shapes, Cellpose [4] creates a reversible mapping from vector flow and foreground prediction to an instance segmentation mask. For the out-of-the-domain setting, Lauenburg *et al*. [6] combines image translation and instance segmentation as a unitary design, while Keaton *et al*. [7] proposes a contrastive learning-based domain adaptation approach to adapt the pre-trained model. Distinct from these works, this paper focuses on adapting images in testing phase, where the domain adaptation methods are inapplicable.

Test-time augmentation (TTA) for segmentation. TTA aims to improve the model performance on test image by augmenting it with geometric and appearance changes and then aggregating all results as a prediction. Moshkov*et al*. [3] adopts geometric transformations, while Huang *et al*. [2] employs style transfer for appearance modifications. Instead of aggregating predictions from all augmented samples, this

*The first author performed work during research internship at Boston College.

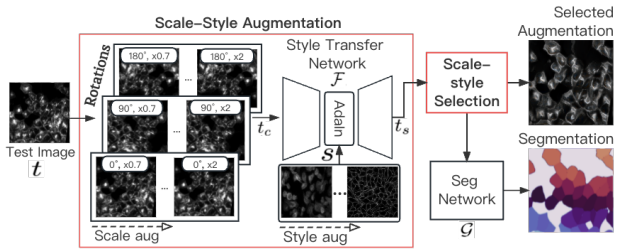


Fig. 2: **Model overview.** For a test image, we first apply scale, style augmentations at different angles. We then employ a consistency-based metric to select the best augmented images for segmentation.

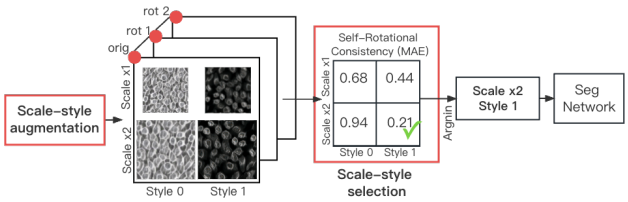


Fig. 3: **Scale-style selector.** This module measures the rotational-consistency of all augmentations and picks the best for segmentation. Two scales/styles are used for illustration.

work selects a single effective one for segmentation.

Image style transfer. Gatys *et al*. [8] introduce a neural network model to transfer the artistic styles of images with a neural network. Subsequent works such as adaptive instance normalization (AdaIN) [9] enable real-time and arbitrary style transfer. For data augmentation, [10, 11, 12] adopts style transfer to generate new training images in other styles. Specifically for test time augmentation, [2] transfers the styles of test images by randomly selecting styles from source data. Alternatively, Ma *et al*. [13] use the Wasserstein metric for better style selection while Liu *et al*. [14] employ a feature histogram matching strategy. In addition, Liu *et al*. [15] formulates a 1-stage framework by directly conducting style transfer in segmentation with parameter injection. This work also adopts neural style transfer for style augmentation and jointly trains the style transfer with the segmentation network to preserve segmentation-related details.

2. METHOD

In this work, we propose an end-to-end test-time augmentation framework for image segmentation. As illustrated in Fig. 2, the framework is composed of a scale-style augmentation module, a scale-style selector, and a U-Net-based segmentation network. The scale-style augmentation module resizes the input to different scales and applies style transfers as visual adaptation for task-oriented domain alignment. Then, the scale-style selector based on rotational-consistency picks the best augmented version for instance-segmentation. In the following, we discuss the proposed framework in detail.

2.1. Scale-Style Image Augmentation

We propose applying combinations of augmentations to the input, including scale augmentation and style augmentation, at multiple rotated versions (Sec. 2.2) of the original image.

For scale augmentation, we adopt a series of different scaling-up ratios, e.g., $\times 1.5$ and $\times 2$, which rescale the input image into different spatial sizes. For style augmentation, we adopt a parametric method dubbed neural style transfer. More specifically, we adopt AdaIN [9] as the style transfer network $\mathcal{F}$ due to its lightweight and adaptability to arbitrary styles. Take t_c as image content and s as style input, AdaIN operation is formulated as

$$\mathcal{F}(t_c, s) = \sigma(s) * (\frac{t_c - \mu(t_c)}{\sigma(t_c)}) + \mu(s). \quad (1)$$

In short, the ordered augmentation procedure involves rotations, resizing, and style transfer, forming a series of the input augmented to different versions, at different angles.

2.2. Scale-Style Selector

While style transfer reduces the visual gap, it damages the content given unexpected style at the wrong scale. As Fig. 1 shows, the segmentation network may fail to recognize images augmented to improper styles or scales. Therefore, instead of aggregating over all augmented segmentation results (Fig. 1c)[3, 2], in this paper, we propose to select one augmentation policy from all combinations of augmentations.

More specifically, we devise a consistency-based scale-style selector. As illustrated in Fig. 3, for each augmentation policy, the transformation is conducted on the image at different angles. We measure its self-rotational-consistency on the image across multiple angles based on mean absolute error (MAE). Essentially, the augmented images at different angles are rotated back to original orientation, for pair-wise MAE measurements. The augmentation with the lowest averaged MAE, *i.e*., best consistency, is selected to proceed with the subsequent pipeline. The assumption behind this method is that the stylized image would be derandomized if the scale and style of cells is properly recognized by style transfer, and as the ST and segmentation are trained end-to-end (Sec. 2.3), they establish recognition on similar domain, which hints the plausibility of an augmentation policy for segmentation. This method is non-parametric and task-agnostic that it can be easily incorporated into existing TTA frameworks.

2.3. Optimization

Prior to training, the ST network $\mathcal{F}$ is pre-trained on random content-style image pairs for training stability. We adopt the

Table 1: Benchmark results on cell segmentation with Cellpose[4] as baseline, using F1 score (%) with IOU matching thresholds .5, .6, and .7.

		Baseline[4]	TTA[3]	StyleInv[2]	Ours (1S/1S)	Ours
CP-Full[4]	.5	78.8	79.2	71.4	81.7	**81.9**
	.6	74.1	74.9	67.8	77.4	**77.9**
	.7	66.9	67.5	59.5	67.7	**68.8**
CP-Hard[4]	.5	65.6	66.3	63.8	75.0	**76.9**
	.6	59.9	60.8	57.4	69.6	**70.3**
	.7	52.6	53.4	49.8	58.6	**59.3**
DSB2018[16]	.5	71.2	72.5	68.0	**80.6**	79.9
	.6	60.9	61.6	58.8	**73.0**	69.4
	.7	44.0	44.9	41.5	**56.7**	52.5

Cellpose model [4] as the segmentation network $\mathcal{G}$. During training, the ST network encoder is fixed while the decoder is jointly trained with $\mathcal{G}$, which enables the following two benefits: (1) The style transfer network will be supervised by the boundary or foreground and background information to preserve more segmentation-related content details; (2) The segmentation network will be finetuned on the style-transferred images to further narrow the visual gap between the training and test phases. Additionally, we incorporate the selection process into training, where only the selected augmented image passes through segmentation and optimization.

For the segmentation network, we employ the conventional practice and optimize it using the Cross-Entropy loss, written as $\mathcal{L}_{Seg}$. The style transfer network is pre-train, and subsequently jointly-trained under content and style loss. The content loss computes MSE between the AdaIN output and the embedded feature of the stylized image. Let stylized image be t_s and style input be s. The content loss is formulated as

$$\mathcal{L}_c = ||\mathcal{F}(t_s, s) - t||. \tag{2}$$

The style loss measures the layer-wise difference in mean and variance between the style image and the stylized output. Let $\phi_i(x)$ be the output of the i-th ReLU layer of ST network $\mathcal{F}$, the style loss is written as

$$\mathcal{L}_s = \sum_{i=1}^{L} ||\mu(\phi_i(t)) - \mu(\phi_i(s))|| + \sum_{i=1}^{L} ||\sigma(\phi_i(t)) - \sigma(\phi_i(s))|| \tag{3}$$

The whole model is optimized with a weighted sum of the content loss, style loss, and segmentation loss with weights 1, 2, 5, respectively as w_c, w_s, and w_{seg}:

$$\mathcal{L}_{total} = \mathcal{L}_{Seg} * w_{seg} + (\mathcal{L}_c * w_c + \mathcal{L}_s * w_s), \tag{4}$$

3. EXPERIMENT

3.1. Datasets and implementation details

Cell image benchmarks. We evaluate our proposed framework with datasets in two tasks: Instance segmentation for microscopy cell images and semantic segmentation for chest x-ray lung segmentation. We apply instance-segmentation evaluation on the following 3 datasets: (1) CP-Full: The full testset (68 images) of the Cellpose dataset [4], a visually diverse multi-centered instance segmentation dataset with 616 cell images. (2) CP-Hard: We observed that 38 of the 68 Cellpose [4] test images are in the training domain, i.e. from the same subset of training images. The remaining 30 images are from un-exposed domains. To assess the model's performance on unseen domains only, we evaluate the 30 images independently, which formulate the Cellpose-Hard Testset. (3) DSB2018 [16]: Data Science Bowl 2018 presents a cell nuclei instance segmentation dataset with 670 training samples. We directly test methods on the training set, as it has mask labels for cell instances.

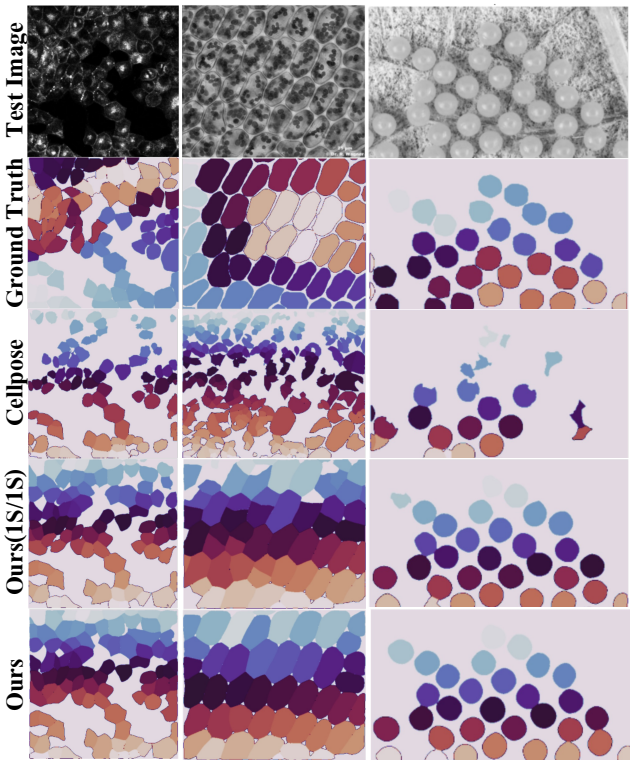


Fig. 4: **Qualitative results on cell image segmentation.** The fourth row represents our co-training pipeline without augmentation, while the final row presents results visualization with our multi-scale/style method.

X-ray image benchmarks. We test our framework for chest X-ray lung segmentation. We adopted 3 chest X-ray datasets for evaluation: (1) Shenzhen [17]: Collected by Shenzhen No.3 People's Hospital, China, containing 662 frontal chest X-rays.(2) Montgomery [17]: The dataset has 138 frontal chest X-rays.(3) Darwin [18]: The Darwin dataset is a sizable collection of 6106 diverse X-ray images.

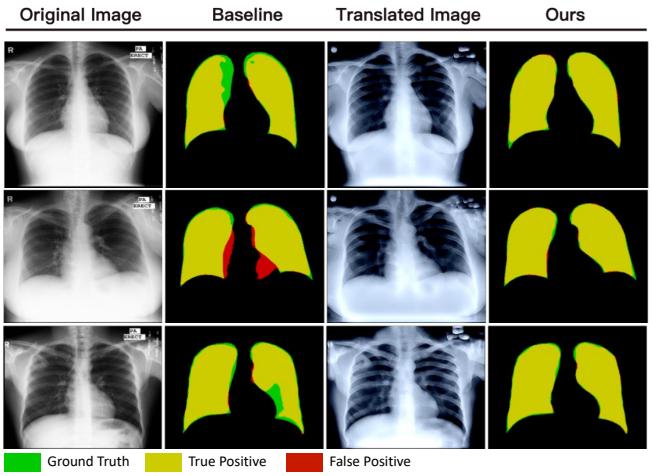


Fig. 5: Qualitative results on X-ray lung segmentation.

Table 2: Benchmark results on X-ray lung segmentation, where the metric is Dice score/Jaccard index in percentage (%).

Train	Shenzhen		Montg.	
Test	Darwin	Montg.	Darwin	Shenzhen
Baseline[19]	84/73	96/93	88/79	97/95
Ours	**86/76**	**98/96**	**90/80**	97/95

3.2. Benchmark results

Cell image results. We present quantitative comparison results on three cell instance segmentation datasets in Table 1. We apply the F1 score as an evaluation metric, representing the ins-segmentation quality with multiple IOU matching thresholds. We compare our methods (1 scale/style and multi-scale/style) with the original Cellpose model and 2 baseline methods: test-time augmentation [3] and style invariance [2]. As a result, our pipeline outperforms all baselines in F1 score metrics with all IOU thresholds. Fig. 4 is the qualitative results of cell instance segmentation. Our method contributes to better signal recognition as well as border matching.

X-ray image results. We present results on chest x-ray lung semantic segmentation task. We employ Deeplab-V3 [19] as the Baseline network. As shown in Table 2, our method performs best in both metrics. We show the qualitative results of lung semantic segmentation in Fig. 5. The lung areas in the image after task-oriented style transfer are more prominent. Therefore, the task-oriented style transfer not only aligns visual differences between source and target domains but also helps accentuate the region of interest.

3.3. Ablation studies and Analysis

Number of Scale and Style. Style input and scaling options compose our entire augmentation space. We separately evaluate the effectiveness of scale and style on Cp-Full with different numbers of scale-style options and present results in F1 score. Table 3 shows that both components contribute to performance improvement. While incorporating more styles results in incremental F1 score gain, scaling options greatly improve results on images with small cells.

Table 3: F1 score (%) with matching threshold 0.5 on CP-Full, with different sets of scale and style augmentations.

# of Styles/# of Scales	{1}	{1,2}	{1,1.5,2}	{0.7,1,1.5,2}
1	78.2	80.1	81.3	81.7
3	79.3	81.0	81.6	81.9

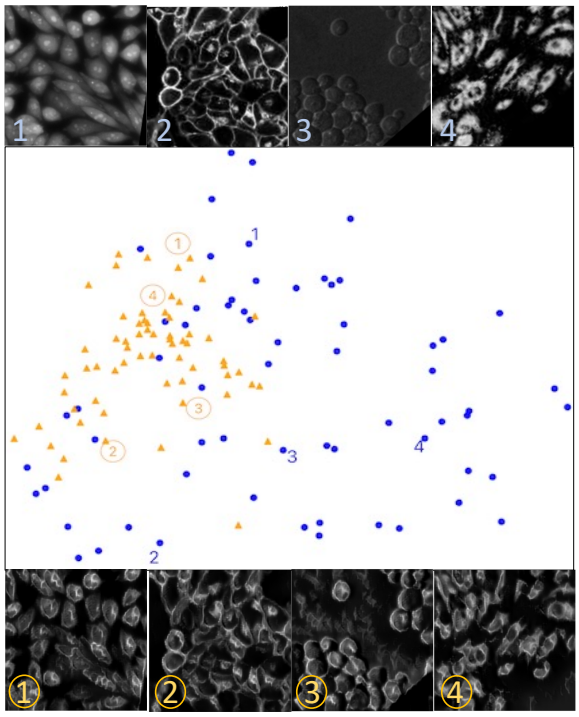


Fig. 6: VGG feature embedding for original (top, blue) and stylized images (bottom, orange).

Style embedding visualization. In Fig. 6, we visualize the distributions of Cellpose test images before and after style transfer in VGG-19 embedded space. Compared to the original images, the stylized images are located in a more condensed area. It demonstrates the effectiveness of style transfer, *i.e.*, aligning the visual styles of images to smaller domains to enhance the segmentation performance.

4. CONCLUSION

In this paper, we have proposed a novel TTA framework, dubbed S^3-TTA, to select the suitable style and scale for test images with a consistency metric. Furthermore, S^3-TTA consists of an end-to-end augmentation-segmentation training pipeline to ensure a task-oriented augmentation. In two biomedical image domains, the proposed framework significantly outperforms the prior art.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using open access human subject data from MC set [17] and Shenzhen set [17], made available by the U.S. National Library of Medicine. Ethical approval was not required.

6. ACKNOWLEDGMENT

This research is supported in part by the NSF award IIS-2239688.

7. REFERENCES

[1] Reka Hollandi et al, "nucleaizer: a parameter-free deep learning framework for nucleus segmentation using image style transfer," *Cell Systems*, vol. 10, no. 5, pp. 453–458, 2020. 1

[2] Xiaoqiong Huang et al, "Style-invariant Cardiac Image Segmentation with Test-time Augmentation," *arXiv e-prints*, p. arXiv:2009.12193, Sept. 2020. 1, 2, 3, 4

[3] Nikita Moshkov et al, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," *Scientific Reports*, vol. 10, 2020. 1, 2, 3, 4

[4] Carsen Stringer et al, "Cellpose: a generalist algorithm for cellular segmentation," *Nature methods*, vol. 18, no. 1, pp. 100–106, 2021. 1, 3

[5] Kaiming He et al, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017. 1

[6] Leander Lauenburg et al, "Instance segmentation of unlabeled modalities via cyclic segmentation gan," *arXiv preprint arXiv:2204.03082*, 2022. 1

[7] Matthew R Keaton et al, "Celltranspose: Few-shot domain adaptation for cellular instance segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 455–466. 1

[8] Gatys Leon A. et al, "A neural algorithm of artistic style," *ArXiv*, vol. abs/1508.06576, 2015. 2

[9] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017. 2

[10] Reka Hollandi et al, "A deep learning framework for nucleus segmentation using image style transfer," *bioRxiv*, 2019. 2

[11] Philip T. G. Jackson et al, "Style augmentation: Data augmentation via style randomization," in *CVPR Workshops*, 2018. 2

[12] Lei Li et al, "Random style transfer based domain generalization networks integrating shape and spatial information," *ArXiv*, vol. abs/2008.12205, 2020. 2

[13] Chunwei Ma et al, "Neural style transfer improves 3d cardiovascular mr image segmentation on inconsistent data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019. 2

[14] Zhendong Liu et al, "Remove appearance shift for ultrasound image segmentation via fast and universal style transfer," *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1824–1828, 2020. 2

[15] Zhendong Liu et al, "Generalize ultrasound image segmentation via instant and plug & play style transfer," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 419–423, 2021. 2

[16] Juan C Caicedo et al, "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nature methods*, vol. 16, no. 12, pp. 1247–1253, 2019. 3

[17] Stefan Jaeger et al, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, 2014. 3, 5

[18] Viacheslav Danilov et al, "Chest x-ray dataset for lung segmentation," `http://dx.doi.org/10.17632/8gf9vpkhgy.2`. 3

[19] Liang-Chieh Chen et al, "Rethinking atrous convolution for semantic image segmentation," *ArXiv*, vol. abs/1706.05587, 2017. 4