SCORPION: Addressing Scanner-Induced Variability in Histopathology

Jeongun Ryu[1], Heon Song[1], Seungeun Lee[1], Soo Ick Cho[1], Jiwon Shin[1], Kyunghyun Paeng[1], and Sérgio Pereira[1]

Lunit Inc.
{rjw0205, heon.song, lsee1113, sooickcho, jwshin, khpaeng, sergio}@lunit.io

Abstract. Ensuring reliable model performance across diverse domains is a critical challenge in computational pathology. A particular source of variability in Whole-Slide Images is introduced by differences in digital scanners, thus calling for better scanner generalization. This is critical for the real-world adoption of computational pathology, where the scanning devices may differ per institution or hospital, and the model should not be dependent on scanner-induced details, which can ultimately affect the patient's diagnosis and treatment planning. However, past efforts have primarily focused on standard domain generalization settings, evaluating on unseen scanners during training, without directly evaluating consistency across scanners for the same tissue. To overcome this limitation, we introduce SCORPION, a new dataset explicitly designed to evaluate model reliability under scanner variability. SCORPION includes 480 tissue samples, each scanned with 5 scanners, yielding 2,400 spatially aligned patches. This scanner-paired design allows for the isolation of scanner-induced variability, enabling a rigorous evaluation of model consistency while controlling for differences in tissue composition. Furthermore, we propose SimCons, a flexible framework that combines augmentation-based domain generalization techniques with a consistency loss to explicitly address scanner generalization. We empirically show that SimCons improves model consistency on varying scanners without compromising task-specific performance. By releasing the SCORPION dataset[1] and proposing SimCons, we provide the research community with a crucial resource for evaluating and improving model consistency across diverse scanners, setting a new standard for reliability testing.

Keywords: Computational Pathology · Scanner Generalization.

1 Introduction

Computational pathology [2] has been increasingly adopted in modern healthcare, offering high-throughput and precise analysis of histopathological samples for disease diagnosis, prognosis [14], and treatment planning [11]. Advances in

[1] SCORPION dataset is available at https://doi.org/10.5281/zenodo.16517924

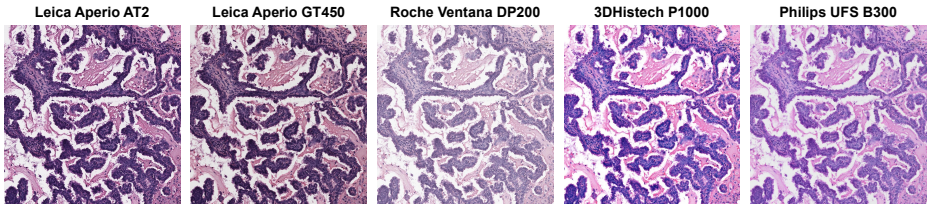


Fig. 1: **Example of scanner-induced variability in SCORPION.** The same tissue region is digitized using five different scanners. Despite capturing identical histological structures, variations in color, contrast, and texture highlight the challenges of inter-scanner variability in histopathology.

computational models have revolutionized traditional workflows, significantly enhancing diagnostic accuracy and efficiency [15,17]. However, a critical challenge hindering the deployment of these models in real-world clinical settings is the variability introduced by different scanners used to digitize tissue slides into whole-slide images (WSI).

Scanners vary significantly in the manufacturer, hardware, and image acquisition settings, introducing scanner-induced variability that can affect WSI, even for the same tissue slide (see Fig. 1). While human pathologists can intuitively disregard such differences, computational models are susceptible to these discrepancies, leading to inconsistent predictions [3,4]. This is particularly problematic in clinical workflows, where inconsistent predictions could result in varying diagnoses or treatment recommendations, jeopardizing patient outcomes [7].

Prior efforts, including initiatives such as Camelyon17 [4] and MIDOG2021 [3], have contributed benchmarks for domain generalization in histopathology. However, these datasets do not provide scanner-paired images, thus making it impossible to directly assess inter-scanner consistency. More recently, datasets [13,22,1] have introduced paired scans of the same tissue acquired with different scanners. While these datasets enable the possibility of consistency evaluation, prior work has predominantly adopted conventional domain generalization settings, evaluating on unseen scanners during training, without explicitly analyzing consistency across scanners for the same tissue. In contrast, we directly leverage the scanner-paired nature of the data to rigorously evaluate inter-scanner consistency, which is critical to ensure clinical reliability of the model.

Building upon these recent datasets that include scanner-paired images, we introduce SCORPION, a novel H&E dataset designed explicitly to evaluate model consistency under scanner variability. SCORPION comprises 480 tissue regions, each scanned using 5 scanners, resulting in 2,400 spatially aligned patches. By isolating scanner variability from tissue heterogeneity, SCORPION enables rigorous evaluation of model consistency across scanners. This dataset establishes a new benchmark for scanner generalization, facilitating the development of robust and reliable computational models. Therefore, SCORPION will help accelerate research in this critical area.

Further, we propose SimCons, a flexible framework that explicitly incorporates consistency as a core objective. SimCons combines style-based augmentation with a consistency loss to encourage consistent predictions on style variation. By emphasizing consistency, we address the unique challenges posed by scanner variability, ensuring the clinical reliability of the model without compromising task-specific performance. Together, SCORPION and SimCons provide a critical resource and methodology for advancing scanner generalization research, setting a new standard for reliability testing in CPath.

In summary, our contributions are threefold: (1) we publicly release SCORPION, a comprehensive dataset of spatially aligned patches scanned using 5 different scanners; (2) we establish a novel problem setting and evaluation protocol specifically addressing inter-scanner variability; and (3) we introduce SimCons, a straightforward yet effective framework that mitigates inter-scanner variability while maintaining or enhancing task-specific performance.

2 SCORPION dataset

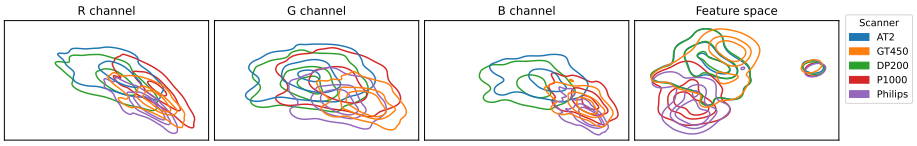
SCORPION is an H&E-stained histopathology dataset specifically built to enable the evaluation of model consistency across scanners. Each sample of the SCORPION is composed of 5 patches, $(x_{AT2}, x_{GT450}, x_{DP200}, x_{P1000}, x_{B300})$, where patches in a sample share the same tissue content, differing only in the scanner used. Fig.1 shows a sample of the SCORPION dataset.

2.1 Dataset Collection.

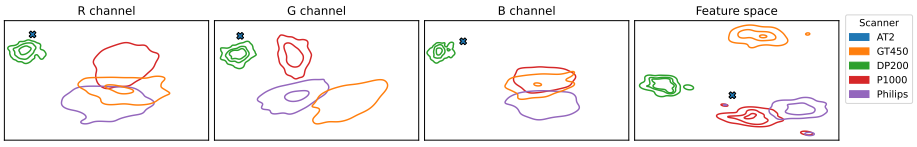
We collected 48 H&E-stained tissue slides, where each slide was digitized using the following scanners: *Leica Aperio AT2, Leica Aperio GT450, Roche Ventana DP200, 3DHistech P1000,* and *Philips UFS B300.* To ensure spatial consistency across the scans, we aligned the images of the same tissue slide through a registration algorithm based on ORB [16] feature extraction and affine transformation. From each aligned slide, we extracted 10 regions, each measuring $800\mu m \times 800\mu m$, and resized them as 1024×1024 pixel patches. This process resulted in 480 samples, each consisting of 5 patches from the same tissue region but captured by different scanners, leading to a total of 2,400 patches.

2.2 Dataset Analysis.

To analyze scanner-induced variability in the SCORPION dataset, we first investigate the input-level distributions. For each RGB channel, the mean and standard deviation of the pixel values are computed for each patch, which quantifies the distribution of pixel intensities across scanners. Then, we further investigate scanner-induced differences in the feature space. Features are extracted for each patch using ResNet50 [9], pre-trained on ImageNet [6]. For visualization, we project the features into a 2D space by UMAP [12]. The density contours of each scanner are displayed in Fig. 2.



(a) Unpaired Analysis.



(b) Paired Analysis.

Fig. 2: **Input- and feature-level analysis across scanners in SCORPION.** Density contours highlight the distribution of patches across different scanners. In the left three columns, (x,y) denote the mean and standard deviation of pixel values for each RGB channel. In the last column, features extracted using a ResNet50 pre-trained on ImageNet are projected into a 2D space using UMAP. (a) In unpaired analysis, all patches are plotted without considering the paired-patch relationship. (b) On the other hand, in paired analysis, AT2 (blue) is set as the reference scanner, and deviations are computed for each paired patch relative to its corresponding AT2 patch, positioning all AT2 patches at (0,0).

Unpaired analysis We begin with an *unpaired analysis*, where the paired-patch relationship is ignored, and all patches are plotted with scanner labels assigned. As shown in Fig. 2a, it reveals substantial overlap between scanners, making it challenging to identify scanner-specific characteristics, potentially leading to mistakenly consider there are no major inter-scanner differences.

Paired analysis To overcome the limitations of the *unpaired analysis*, we leverage the paired nature of the SCORPION dataset for a *paired analysis*. In this analysis, AT2 is selected as the reference scanner, and deviations are computed for each paired patch relative to the corresponding AT2 patch, ensuring that all AT2 patches are positioned at (0,0) in Fig. 2b. As a result, we found that the distributions of each scanner are easily separated in feature space, which indicates that the pre-trained encoder captures scanner-specific information from the input. Also in the RGB channels, except for the GT450 (orange), all the scanner distributions are well separated. This shows the superiority of *paired analysis*, which provides a clearer view of scanner-induced variability compared to *unpaired analysis*, by utilizing the scanner-paired structure.

2.3 Inter-Scanner Consistency Evaluation Protocol.

Prior efforts [1,3,4,13,22] that studied inter-scanner variability have typically followed the standard domain generalization protocol by evaluating the model on unseen scanners during training. However, this evaluation approach does not

assess whether the model produces consistent outputs when only the scanner changes while the underlying tissue remains the same. In real-world clinical settings, where scanner types can vary between hospitals, this limitation poses a significant risk: a patient could receive different clinical decisions depending solely on the scanner used. Motivated by this, we leverage the scanner-paired design of SCORPION to propose a new evaluation protocol that rigorously quantifies model consistency across scanners, addressing a critical gap in existing evaluation methods.

In this evaluation protocol, we compute the consistency score (e.g. Dice score in tissue segmentation task) between predictions from scanner-paired patches, for each scanner pair. Since SCORPION includes 5 scanners, there are 10 unique scanner pairs, resulting in 10 consistency scores. Finally, we compute two measurements: the average and minimum of the 10 consistency scores. The average score represents the overall consistency of the model across different scanner pairs, while the minimum score captures the lower bound, ensuring that the model maintains a certain level of consistency, regardless of which scanner pair is used. This minimum metric is particularly important for assessing the worst-case scenario in real-world clinical settings, where scanner-induced variability can affect medical treatment.

3 Method

To enhance the model consistency across diverse scanners, we propose SimCons, a simple framework combining style-based augmentation (SA) with a consistency loss. SimCons leverages an SA to synthesize style-altered images while preserving the content of the original image. Then, the consistency loss explicitly encourages consistent model predictions between original and style-altered images. An overview of the SimCons framework is illustrated in Fig.3.

3.1 Style-based Augmentation (SA)

Due to the differences among scanners, including optical components and sensor characteristics, each scanner introduces unique color profiles, texture patterns, and contrast properties. As a result, the same tissue scanned by multiple scanners can have varying styles for the same content. In prior work, various SA methods [8,18,20,23] were studied to improve the robustness to these style variations. In Tab.1, we observe that SA methods improve the model consistency on scanner variations while maintaining or improving performance in primary tasks.

3.2 Consistency Loss

While SA can enhance model performance on unseen scanners by generating images with various styles during training, it alone cannot ensure consistent predictions across scans with identical tissue content. To explicitly encourage the model to produce consistent output when only the scanner changes, we apply a

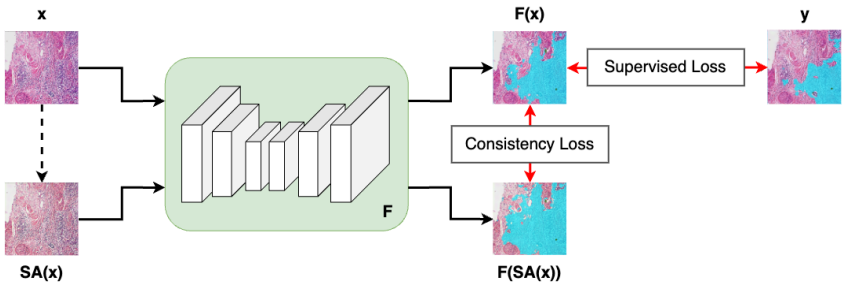


Fig. 3: **Overview of SimCons framework.** SimCons improves model consistency across scanners by integrating style-based augmentation (SA) with a consistency loss. The SA module transforms a training image into a style-altered version while preserving its content. Both the original and style-altered images are fed into the model (F), and their predictions are aligned using a consistency loss. Simultaneously, a supervised loss ensures task-specific learning.

consistency loss between the model's predictions on the original image and its style-altered counterpart. The final loss of SimCons aggregates the supervised and consistency losses, formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{supervised}}(F(x), y) + \lambda \cdot \mathcal{L}_{\text{consistency}}(F(x), F(\text{SA}(x))) \quad (1)$$

where (x, y) represents a training patch with a corresponding label, and F denotes the model to be trained. The hyperparameter λ controls the impact of the consistency loss. Intuitively, a higher λ can reduce the scanner-induced variability of the model predictions. But, if set too large, it can hinder the primary task performance. We conduct an in-depth study on the effect of λ in Section 4.3.

4 Experiments

In this section, we assess the model consistency over scanners and show the effectiveness of the proposed SimCons framework. We use tissue segmentation as the primary task; accordingly, Dice score is chosen for both consistency and primary task metrics. For this, we first acquire tissue segmentation models by training and validating on the internal HTS dataset. Then, we measure two types of metrics: (1) scanner-paired Dice as a consistency metric following Section 2.3 with the SCORPION dataset and (2) primary task Dice with the HTS dataset.

HTS is an internal dataset containing 8,327 patches extracted from 3,399 H&E-stained WSIs, decomposed into training, validation, and test sets with 4,027, 3,947, and 353 patches, respectively. Each patch is a 1024 $\times$ 1024 pixel image that represents a tissue region of approximately $800\mu m \times 800\mu m$. Per-pixel manual annotations for semantic segmentation consider 3 classes: cancer area, cancer stroma, and background. All the WSIs are scanned with *Leica Aperio AT2* and *3DHistech P1000*. Importantly, this is only a subset of the scanners used in the SCORPION dataset, thus allowing us to assess generalization.

Table 1: **Tissue segmentation performance with consistency score.** The Baseline has no style-based augmentation.

	scanner-paired Dice		primary task Dice	
Method	**Avg**	**Min**	**Val**	**Test**
Baseline	84.89±0.56	75.63±0.92	68.71±0.06	80.04±0.40
ColorJitter	86.77±0.46	78.65±0.85	68.93±0.07	80.54±0.42
+ SimCons	89.64±0.20	83.89±0.40	69.45±0.13	81.32±0.26
RandStainNA [18]	88.47±0.37	81.93±0.75	68.75±0.06	80.56±0.49
+ SimCons	90.22±0.21	84.92±0.34	69.45±0.14	81.75±0.44
FDA [24]	88.10±0.32	81.98±0.56	68.41±0.05	80.31±0.54
+ SimCons	90.32±0.27	85.43±0.35	68.87±0.07	82.08±0.28

4.1 Experimental Setup

We use DeepLabV3+ [5] with a ResNet34 [9] backbone for training the models on the HTS dataset. Models are trained for 150 epochs with the Adam optimizer [10]. The Dice loss [19] is used for both supervised and consistency loss. During training, four data augmentations are randomly applied, including two photometric (Gaussian blur, Gaussian noise) and two geometric (horizontal flip, vertical flip) transformations. For each experiment, the best epoch is chosen based on the performance on the HTS validation set. All experiments are repeated five times, and the mean and standard deviation of the metrics are reported.

4.2 Main Results

We adopt three SA methods with consistency loss to validate the efficacy of the SimCons framework. The adopted SA methods are the following:

ColorJitter. It is a commonly used data augmentation technique that randomly alters the brightness, contrast, saturation, and hue of an image. Despite its simplicity, it has been proven to improve the generalizability of the model, especially in color-related variations [21].

RandStainNA. It is designed to address stain variations in histopathology images by combining stain normalization and stain augmentation [18].

FDA. It is an empirically proven technique to improve model generalization. It generates diverse style images by transferring low-frequency information between training images, utilizing the content-preserving characteristic of the Fourier phase component [24].

From Tab.1, we find that all SA methods show improved consistency over scanners compared to the Baseline, which does not have a style-based augmentation. Adding a consistency constraint on top of the SA methods further boosts

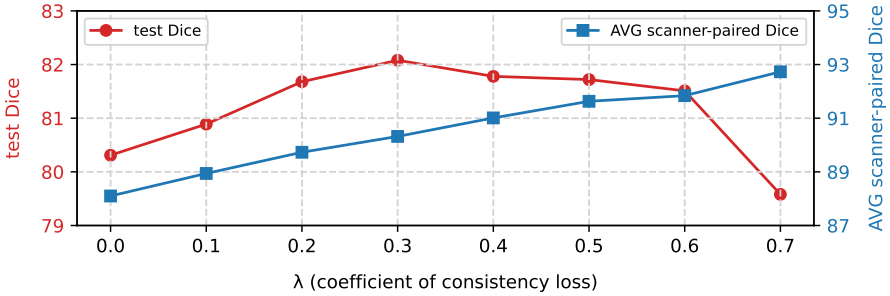


Fig. 4: **Consistency vs. Primary Task.** Adjusting λ can effectively balance improved scanner consistency with primary task performance.

the consistency metric. This implies that consistency loss is the key to enhancing the model consistency over scanners by enforcing the model to output style-invariant predictions. In addition, adding a consistency constraint also improves the primary task performance, especially in the test set. This suggests that while the consistency loss is aimed at reducing scanner-induced variability, it also improves generalization, further showing the benefits of the SimCons framework.

4.3 Trade-off: Scanner Robustness vs Task Performance

As described in Eq.1, the loss function of SimCons consists of two terms: a supervised loss for learning the primary task and a consistency loss to encourage consistent predictions over style variation. The total loss is weighted with a coefficient λ controlling the influence of the consistency loss. In Fig.4, we observe a trade-off between the primary task performance (red) and consistency score (blue). When $\lambda = 0.0$, the model only considers the primary task, leading to a low consistency score. With increased λ, the consistency score improves consistently. Additionally, until $\lambda = 0.3$, we observe that it helps improving the primary task. However, the performance of the primary task degrades for $\lambda > 0.3$. This is explained by a mode collapse phenomenon, where the model can reduce the consistency loss by generating a trivial but consistent prediction, thus harming the primary task's learning. Therefore, it is important to balance the supervised and consistency losses to achieve satisfactory consistency while showing strong performance in the primary task. This can be done by manipulating λ, where the optimal range was 0.3 to 0.5 within our experiment setting.

5 Conclusion

This paper introduces SCORPION, a new dataset specifically designed to isolate scanner-induced variability in histopathology. By providing scanner-paired patches from five distinct scanners, SCORPION enables precise evaluation of

model consistency across scanners while controlling for tissue differences. Furthermore, we propose an evaluation protocol that leverages SCORPION's unique structure to quantify inter-scanner variability, establishing a new benchmark for scanner generalization research. Beyond dataset contributions, we present SimCons, a flexible framework that integrates style-based augmentation with consistency loss to explicitly mitigate scanner-induced prediction variability. Our results demonstrate that SimCons enhances model consistency across scanners without compromising task performance. These findings underscore the critical role of inter-scanner variability in real-world deployment, where reliable and consistent model predictions are essential for clinical decision-making. By releasing SCORPION and its evaluation framework, we aim to provide the research community with a foundational resource for improving model robustness and advancing the adoption of AI-driven pathology in clinical practice.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cosas challenge. `https://cosas.grand-challenge.org/` (2024)
2. Abels, E., et al.: Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. The Journal of Pathology **249**(3), 286–294 (2019)
3. Aubreville, M., et al.: Mitosis domain generalization in histopathology images - the midog challenge. Medical Image Analysis **84**, 102699 (2023)
4. Bandi, P., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE Transactions on Medical Imaging (2018)
5. Chen, L.C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 801–818 (2018)
6. Deng, J., et al.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
7. Duenweg, S.R., et al.: Whole slide imaging (wsi) scanner differences influence optical and computed properties of digitized prostate cancer histology. Journal of Pathology Informatics **14**, 100321 (2023)
8. Faryna, K., et al.: Tailoring automated data augmentation to he-stained histopathology. In: Proceedings of the Fourth Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 143, pp. 168–178. PMLR (2021)
9. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
10. Kingma, D.P., et al.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)

11. Lee, H.J., et al.: Artificial intelligence (ai)–powered spatial analysis of tumor-infiltrating lymphocytes (til) for prediction of response to neoadjuvant chemotherapy (nac) in triple-negative breast cancer (tnbc). (2022)
12. McInnes, L., et al.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29), 861 (2018)
13. Ochi, Y., et al.: Registered multi-device/staining histology image dataset for domain-agnostic machine learning models. Scientific Data **11**(330) (2024)
14. Park, C., et al.: Tumor-infiltrating lymphocyte enrichment predicted by ct radiomics analysis is associated with clinical outcomes of non-small cell lung cancer patients receiving immune checkpoint inhibitors. Frontiers in Immunology **13**, 1038089 (2023)
15. Park, S., et al.: Artificial intelligence-powered spatial analysis of tumor-infiltrating lymphocytes as complementary biomarker for immune checkpoint inhibition in non-small-cell lung cancer. Journal of Clinical Oncology **40**(17), 1916–1928 (2022)
16. Rublee, E., et al.: Orb: An efficient alternative to sift or surf. In: International Conference on Computer Vision. pp. 2564–2571 (2011)
17. Ryu, J., et al.: Ocelot: Overlapped cell on tissue dataset for histopathology. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23902–23912 (2023)
18. Shen, Y., et al.: Randstainna: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 212–221. Springer (2022)
19. Sudre, C.H., et al.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 240–248 (2017)
20. Tellez, D., et al.: H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection. In: Medical Imaging (2018)
21. Tellez, D., et al.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis **58**, 101544 (2019)
22. Wilm, J., et al: Multi-scanner canine cutaneous squamous cell carcinoma histopathology dataset. In: Bildverarbeitung für die Medizin 2023. pp. 199–204. Informatik aktuell, Springer Vieweg, Wiesbaden (2023)
23. Yang, S., et al.: Sk-unet model with fourier domain for mitosis detection. In: International conference on medical image computing and computer-assisted intervention. pp. 86–90. Springer (2021)
24. Yang, Y., et al.: Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)