TPRO: Text-Prompting-Based Weakly Supervised Histopathology Tissue Segmentation

Shaoteng Zhang[1,2,4], Jianpeng Zhang[2], Yutong Xie[3](✉), and Yong Xia[1,2,4](✉)

[1] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
yxia@nwpu.edu.cn

[2] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China

[3] Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia
yutong.xie678@gmail.com

[4] Research and Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China

Abstract. Most existing weakly-supervised segmentation methods rely on class activation maps (CAM) to generate pseudo-labels for training segmentation models. However, CAM has been criticized for highlighting only the most discriminative parts of the object, leading to poor quality of pseudo-labels. Although some recent methods have attempted to extend CAM to cover more areas, the fundamental problem still needs to be solved. We believe this problem is due to the huge gap between image-level labels and pixel-level predictions and that additional information must be introduced to address this issue. Thus, we propose a text-prompting-based weakly supervised segmentation method (TPRO), which uses text to introduce additional information. TPRO employs a vision and label encoder to generate a similarity map for each image, which serves as our localization map. Pathological knowledge is gathered from the internet and embedded as knowledge features, which are used to guide the image features through a knowledge attention module. Additionally, we employ a deep supervision strategy to utilize the network's shallow information fully. Our approach outperforms other weakly supervised segmentation methods on benchmark datasets LUAD-HistoSeg and BCSS-WSSS datasets, setting a new state of the art. Code is available at: https://github.com/zhangst431/TPRO.

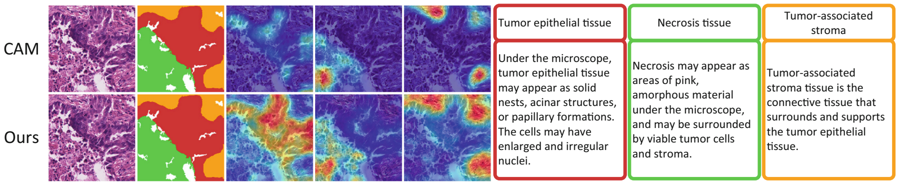
Keywords: Histopathology Tissue Segmentation · Weakly-Supervised Semantic Segmentation · Vision-Language

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_11.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 109–118, 2023.
https://doi.org/10.1007/978-3-031-43907-0_11

1 Introduction

Automated segmentation of histopathological images is crucial, as it can quantify the tumor micro-environment, provide a basis for cancer grading and prognosis, and improve the diagnostic efficiency of clinical doctors [6,13,19]. However, pixel-level annotation of images is time-consuming and labor-intensive, especially for histopathology images that require specialized knowledge. Therefore, there is an urgent need to pursue weakly supervised solutions for pixel-wise segmentation. Nonetheless, weakly supervised histopathological image segmentation presents a challenge due to the low contrast between different tissues, intra-class variations, and inter-class similarities [4,11]. Additionally, the tissue structures in histopathology images can be randomly arranged and dispersed, which makes it difficult to identify complete tissues or regions of interest [7].



Tumor epithelial tissue

Under the microscope, tumor epithelial tissue may appear as solid nests, acinar structures, or papillary formations. The cells may have enlarged and irregular nuclei.

Necrosis tissue

Necrosis may appear as areas of pink, amorphous material under the microscope, and may be surrounded by viable tumor cells and stroma.

Tumor-associated stroma

Tumor-associated stroma tissue is the connective tissue that surrounds and supports the tumor epithelial tissue.

Fig. 1. Comparison of activation maps extracted from CAM and our method, from left to right: origin image, ground truth, three activation maps of tumor epithelial (red), necrosis (green), and tumor-associated stroma (orange) respectively. On the right side, there are some examples of the related language knowledge descriptions used in our method. It shows that CAM only highlights a small portion of the target, while our method, which incorporates external language knowledge, can encompass a wider and more precise target tissue. (Color figure online)

Recent studies on weakly supervised segmentation primarily follow class activation mapping (CAM) [20], which localizes the attention regions and then generates the pseudo labels to train the segmentation network. However, the CAM generated based on the image-level labels can only highlight the most discriminative region, but fail to locate the complete object, leading to defective pseudo labels, as shown in Fig. 1. Accordingly, many attempts have been made to enhance the quality of CAM and thus boost the performance of weakly supervised segmentation. Han *et al.* [7] proposed an erasure-based method that continuously expands the scope of attention areas to obtain rich content of pseudo labels. Li *et al.* [11] utilized the confidence method to remove any noise that may exist in the pseudo labels and only included the confident pixel labels for the segmentation training. Zhang *et al.* [18] leveraged the Transformer to model the long-distance dependencies on the whole histopathological images to improve the CAM's ability to find more complete regions. Lee *et al.* [10] utilized the ability of an advanced saliency detection model to assist CAM in locating more precise targets. However, these improved variants still face difficulties in capturing the complete tissues. The primary limitation is that the symptoms and manifestations of histopathological subtypes cannot be comprehensively described by an abstract semantic category. As a result, the image-level label supervision may not be sufficient to pinpoint the complete target area.

To remedy the limitations of image-level supervision, we advocate for the integration of language knowledge into weakly supervised learning to provide reliable guidance for the accurate localization of target structures. To this end, we propose a text-prompting-based weakly supervised segmentation method (TPRO) for accurate histopathology tissue segmentation. The text information originates from the task's semantic labels and external descriptions of subtype manifestations. For each semantic label, a pre-trained medical language model is utilized to extract the corresponding text features that are matched to each feature point in the image spatial space. A higher similarity represents a higher possibility of this location belonging to the corresponding semantic category. Additionally, the text representations of subtype manifestations, including tissue morphology, color, and relationships to other tissues, are extracted by the language model as external knowledge. The discriminative information can be explored from the text knowledge to help identify and locate complete tissues accurately by jointly modeling long-range dependencies between image and text. We conduct experiments on two weakly supervised histological segmentation benchmarks, LUAD-HistoSeg and BCSS-WSSS, and demonstrate the superior quality of pseudo labels produced by our TPRO model compared to other CAM-based methods.

Our contributions are summarized as follows: (1) To the best of our knowledge, this is the first work that leverages language knowledge to improve the quality of pseudo labels for weakly-supervised histopathology image segmentation. (2) The proposed text prompting models the correlation between image representations and text knowledge, effectively improving the quality of pseudo labels. (3) The effectiveness of our approach has been effectively validated by two benchmarks, setting a new state of the art.

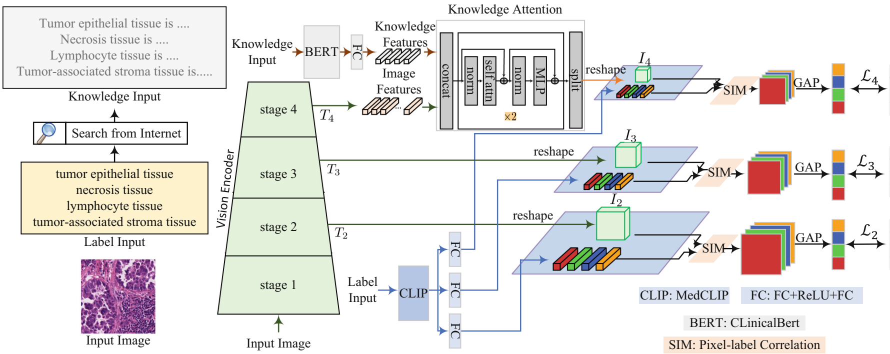


Fig. 2. The framework of the proposed TPRO.

2 Method

Figure 2 displays the proposed TPRO framework, a classification network designed to train a suitable model and extract segmentation pseudo-labels. The framework comprises a knowledge attention module and three encoders: one vision encoder and two text encoders (label encoder and knowledge encoder).

2.1 Classification with Deep Text Guidance

Vision Encoder. The vision encoder is composed of four stages that encode the input image into image features. The image features are denoted as $T_s \in R^{M_s \times C_s}$, where $2 \leq s \leq 4$ indicates the stage number.

Label Encoder. The label encoder encodes the text labels in the dataset into N label features, denoted as $L \in R^{N \times C_l}$, where N represents the number of classes in the dataset and C_l represents the dimension of label features. Since the label features will be used to calculate the similarity with image features, it is important to choose a language model that has been pre-trained on image-text pairs. Here we use MedCLIP[1] as our label encoder, which is a model fine-tuned on the ROCO dataset [12] based on CLIP [14].

Knowledge Encoder. The knowledge encoder is responsible for embedding the descriptions of subtype manifestations into knowledge features, denoted as $K \in R^{N \times C_k}$. The knowledge features guide the image features to focus on regions relevant to the target tissue. To encode the subtype manifestations description into more general semantic features, we employ ClinicalBert [2] as our knowledge encoder. ClinicalBert is a language model that has been fine-tuned on the MIMIC-III [8] dataset based on BioBert [9].

Adaptive Layer. We freeze the label and knowledge encoders for training efficiency but add an adaptive layer after the text encoders to better tailor the text features to our dataset. The adaptive layer is a simple FC-ReLU-FC block that allows for fine-tuning of the features extracted from the text encoders.

Label-Pixel Correlation. After the input image and text labels are embedded. We employ the inner product to compute the similarity between image features and label features, denoted as F_s. Specially, we first reshape the image features from a token format into feature maps. We denote the feature map as $I_s \in R^{H_s \times W_s \times C_s}$, where H_s and W_s mean the height and width of the feature map. F_s is computed with the below formula

$$F_s[i, j, k] = I_s[i, j] \cdot L[k] \in R^{H_s \times W_s \times N}. \tag{1}$$

Then, we perform a global average-pooling operation on the produced similarity map to obtain the class prediction, denoted as $P_s \in R^{1 \times N}$. We then calculate the binary cross-entropy loss between the class label $Y \in R^{1 \times N}$ and the class prediction P_s to supervise the model training, which is formulated as:

[1] https://github.com/Kaushalya/medclip.

$$\mathcal{L}_s = -\frac{1}{N}\sum_{n=1}^{N} Y[n] log\, \sigma(P_s[n]) + (1 - Y[n]) log[1 - \sigma(P_s[n])] \tag{2}$$

Deep Supervision. To leverage the shallow features in the network, we employ a deep supervision strategy by calculating the similarity between the image features from different stages and the label features from different adaptive layers. Class predictions are derived from these similarity maps. The loss of the entire network is computed as:

$$\mathcal{L} = \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 + \lambda_4\mathcal{L}_4. \tag{3}$$

2.2 Knowledge Attention Module

To enhance the model's understanding of the color, morphology, and relationships between different tissues, we gather text representations of different subtype manifestations from the Internet and encode them into external knowledge via the knowledge encoder. The knowledge attention module uses this external knowledge to guide the image features toward relevant regions of the target tissues.

The knowledge attention module, shown in Fig. 2, consists of two multi-head self-attention modules. The image features $T_4 \in R^{M_4 \times C_4}$ and knowledge features after adaptive layer $K \in R^{N \times C_4}$ are concatenated in the token dimension to obtain $T_{fuse} \in R^{(M_4+N) \times C_4}$. This concatenated feature is then fed into the knowledge attention module for self-attention calculation. The output tokens are split, and the part corresponding to the image features is taken out. Noted that the knowledge attention module is added only after the last stage of the vision encoder to save computational resources.

2.3 Pseudo Label Generation

In the classification process, we calculate the similarity between image features and label features to obtain a similarity map F, and then directly use the result of global average pooling on the similarity map as a class prediction. That is, the value at position (i, j, k) of F represents the probability that pixel (i, j) is classified into the k_{th} class. Therefore we directly use F as our localization map. We first perform min-max normalization on it, the formula is as follows

$$F^c_{fg} = \frac{F^c - \min(F^c)}{\max(F^c) - \min(F^c)}, \tag{4}$$

where $1 \leq c \leq N$ means c_{th} class in the dataset. Then we calculate the background localization map by the following formula:

$$F_{bg}(i, j) = \{1 - \max_{c\in[0,C)} F^c_{fg}(i, j)\}^{\alpha}, \tag{5}$$

where $\alpha \geq 1$ denotes a hyper-parameter that adjusts the background confidence scores. Referring to [1] and combined with our own experiments, we set α to

10. Then we stitch together the localization map of foreground and background, denoted as $\hat{F}$. In order to make full use of the shallow information of the network, we perform weighted fusion on the localization maps from different stages by the following formula:

$$F_{all} = \gamma_2 \hat{F}_2 + \gamma_3 \hat{F}_4 + \gamma_4 \hat{F}_4. \tag{6}$$

Finally, we perform argmax operation on F_{all} to obtain the final pseudo-label.

3 Experiments

3.1 Dataset

LUAD-HistoSeg[2] [7] is a weakly-supervised histological semantic segmentation dataset for lung adenocarcinoma. There are four tissue classes in this dataset: tumor epithelial (TE), tumor-associated stroma (TAS), necrosis (NEC), and lymphocyte (LYM). The dataset comprises 17,258 patches of size 224×224. According to the official split, the dataset is divided into a training set (16,678 patch-level annotations), a validation set (300 pixel-level annotations), and a test set (307 pixel-level annotations). **BCSS-WSSS**[3] is a weakly supervised tissue semantic segmentation dataset extracted from the fully supervised segmentation dataset BCSS [3], which contains 151 representative H&E-stained breast cancer pathology slides. The dataset was randomly cut into 31826 patches of size 224 × 224 and divided into a training set (23422 patch-level annotations), a validation set (3418 pixel-level annotations), and a test set (4986 pixel-level annotations) according to the official split. There are four foreground classes in this dataset, including Tumor (TUM), Stroma (STR), Lymphocytic infiltrate (LYM), and Necrosis (NEC).

3.2 Implementation Details

For the classification part, we adopt MixTransformer [17] pretrained on ImageNet, MedCLIP, and ClinicalBert [2] as our vision encoder, label encoder, and

Table 1. Comparison of the pseudo labels generated by our proposed method and those generated by previous methods.

Dataset	LUAD-HistoSeg					BCSS-WSSS				
Method	TE	NEC	LYM	TAS	mIoU	TUM	STR	LYM	NEC	mIoU
CAM [20]	69.66	72.62	72.58	66.88	70.44	66.83	58.71	49.41	51.12	56.52
Grad-CAM [15]	70.07	66.01	70.18	64.76	67.76	65.96	56.71	43.36	30.04	49.02
TransWS (CAM) [18]	65.92	60.16	73.34	69.11	67.13	64.85	58.17	44.96	50.60	54.64
MLPS [7]	71.72	76.27	73.53	67.67	72.30	70.76	61.07	50.87	52.94	58.91
TPRO (Ours)	**74.82**	**77.55**	**76.40**	**70.98**	**74.94**	**77.18**	**63.77**	**54.95**	**61.43**	**64.33**

[2] https://drive.google.com/drive/folders/1E3Yei3Or3xJXukHIybZAgochxfn6FJpr.

[3] https://drive.google.com/drive/folders/1iS2Z0DsbACqGp7m6VDJbAcgzeXNEFr77.

knowledge encoder, respectively. The hyperparameters during training and evaluation can be found in the supplementary materials. We conduct all of our experiments on 2 NVIDIA GeForce RTX 2080 Ti GPUs.

3.3 Compare with State-of-the-Arts

Comparison on Pseudo-Labels. Table 1 compares the quality of our pseudo-labels with those generated by previous methods. CAM [20] and Grad-CAM [15] were evaluated using the same ResNet38 [16] classifier, and the results showed that CAM [20] outperformed Grad-CAM [15], with mIoU values of 70.44% and 56.52% on the LUAD-HistoSeg and BCSS-WSSS datasets, respectively. TransWS [18] consists of a classification and a segmentation branch, and Table 1 displays the pseudo-label scores generated by the classification branch. Despite using CAM [20] for pseudo-label extraction, TransWS [18] yielded inferior results compared to CAM [20]. This could be due to the design of TransWS [18] for single-label image segmentation, with the segmentation branch simplified to binary segmentation to reduce the difficulty, while our dataset consists of multi-label images. Among the compared methods, MLPS [7] was the only one to surpass CAM [20] in terms of the quality of the generated pseudo-labels, with its proposed progressive dropout attention effectively expanding the coverage of target regions beyond what CAM [20] can achieve. Our proposed method outperformed all previous methods on both LUAD-HistoSeg and BCSS-WSSS datasets, with improvements of 2.64% and 5.42% over the second-best method, respectively (Table 2).

Table 2. Comparison of the final segmentation results between our method and the methods in previous years.

Dataset	LUAD-HistoSeg					BCSS-WSSS				
Method	TE	NEC	LYM	TAS	mIoU	TUM	STR	LYM	NEC	mIoU
HistoSegNet [5]	45.59	36.30	58.28	50.82	47.75	33.14	46.46	29.05	1.91	27.64
TransWS (seg) [18]	57.04	49.98	59.46	58.59	56.27	44.71	36.49	41.72	38.08	40.25
OEEM [11]	73.81	70.49	71.89	69.48	71.42	74.86	64.68	48.91	61.03	62.37
MLPS [7]	73.90	77.48	73.61	69.53	73.63	74.54	64.45	52.54	58.67	62.55
TPRO (Ours)	**75.80**	**80.56**	**78.14**	**72.69**	**76.80**	**77.95**	**65.10**	**54.55**	**64.96**	**65.64**

Comparison on Segmentation Results. To further evaluate our proposed method, we trained a segmentation model using the extracted pseudo-labels and compared its performance with previous methods. Due to its heavy reliance on dataset-specific post-processing steps, HistoSegNet [5] failed to produce the desired results on our datasets. As we have previously analyzed since the datasets we used are all multi-label images, it was challenging for the segmentation branch of TransWS [18] to perform well, and it failed to provide an overall benefit to the model. Experimental results also indicate that the IoU scores of its segmentation

Table 3. Comparison the effectiveness of label text(LT), knowledge text(KT), and deep supervision(DS).

LT	DS	KT	TE	NEC	LYM	TAS	mIoU
			68.11	75.24	64.95	66.57	68.72
✓			72.39	72.44	71.37	68.67	71.22
✓	✓		72.41	72.11	74.21	70.07	72.20
✓	✓	✓	**74.82**	**77.55**	**76.40**	**70.98**	**74.94**

Table 4. Comparison of pseudo labels extracted from the single stage and our fused version.

	TE	NEC	LYM	TAS	mIoU
stage2	67.16	65.28	67.38	55.09	63.73
stage3	72.13	70.83	73.47	69.46	71.47
stage4	72.69	77.57	76.06	69.81	74.03
fusion	**74.82**	**77.55**	**76.40**	**70.98**	**74.94**

branch were even lower than the pseudo-labels of the classification branch. By training the segmentation model of OEEM [11] using the pseudo-labels extracted by CAM [20] in Table 1, we can observe a significant improvement in the final segmentation results. The final segmentation results of MLPS [7] showed some improvement compared to its pseudo-labels, indicating the effectiveness of the Multi-layer Pseudo Supervision and Classification Gate Mechanism strategy proposed by MLPS [7]. Our segmentation performance surpassed all previous methods. Specifically, our mIoU scores exceeded the second-best method by 3.17% and 3.09% on LUAD-HistoSeg and BCSS-WSSS datasets, respectively. Additionally, it is worth noting that we did not use any strategies specifically designed for the segmentation stage.

3.4 Ablation Study

The results of our ablation experiments are presented in Table 3. We set the baseline as the framework shown in Fig. 2 with all text information and deep supervision strategy removed. It is evident that the addition of textual information increases our pseudo-label mIoU by 2.50%. Furthermore, including the deep supervision strategy and knowledge attention module improves our pseudo-label by 0.98% and 2.74%, respectively. These findings demonstrate the significant contribution of each proposed module to the overall improvement of the results.

In order to demonstrate the effectiveness of fusing pseudo-labels from the last three stages, we have presented in Table 4 the IoU scores for each stage's pseudo-labels as well as the fused pseudo-labels. It can be observed that after fusing the pseudo-labels, not only have the IoU scores for each class substantially increased, but the mIoU score has also increased by 0.91% compared to the fourth stage.

4 Conclusion

In this paper, we propose the TPRO to address the limitation of weakly supervised semantic segmentation on histopathology images by incorporating text supervision and external knowledge. We argue that image-level labels alone cannot provide sufficient information and that text supervision and knowledge attention can provide additional guidance to the model. The proposed method achieves the best results on two public datasets, LUAD-HistoSeg and BCSS-WSSS, demonstrating the superiority of our method.

Acknowledgment. This work was supported in part by the Natural Science Foundation of Ningbo City, China, under Grant 2021J052, in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYK-FZD06), in part by the National Natural Science Foundation of China under Grant 62171377, in part by the Key Technologies Research and Development Program under Grant 2022YFC2009903/2022YFC2009900, in part by the Key Research and Development Program of Shaanxi Province, China, under Grant 2022GY-084, and in part by the Science and Technology Innovation Committee of Shenzhen Municipality, China, under Grants JCYJ20220530161616036.

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4981–4990 (2018)
2. Alsentzer, E., et al.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
3. Amgad, M., et al.: Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics **35**(18), 3461–3467 (2019)
4. Chan, L., Hosseini, M.S., Plataniotis, K.N.: A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. Int. J. Comput. Vision **129**, 361–384 (2021)
5. Chan, L., Hosseini, M.S., Rowsell, C., Plataniotis, K.N., Damaskinos, S.: Histosegnet: semantic segmentation of histological tissue type in whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10662–10671 (2019)
6. Chen, H., Qi, X., Yu, L., Heng, P.A.: Dcan: deep contour-aware networks for accurate gland segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2487–2496 (2016)
7. Han, C., et al.: Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Med. Image Anal. **80**, 102487 (2022)
8. Johnson, A.E., et al.: Mimic-iii, a freely accessible critical care database. Sci. Data **3**(1), 1–9 (2016)
9. Lee, J., et al.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
10. Lee, S., Lee, M., Lee, J., Shim, H.: Railroad is not a train: saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5495–5505 (2021)
11. Li, Y., Yu, Y., Zou, Y., Xiang, T., Li, X.: Online easy example mining for weakly-supervised gland segmentation from histology images. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV. pp. 578–587. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_55

12. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Stoyanov, D., et al. (eds.) LABELS/CVII/STENT -2018. LNCS, vol. 11043, pp. 180–189. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01364-6_20
13. Qaiser, T., et al.: Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. Med. Image Anal. **55**, 1–14 (2019)
14. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
16. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: revisiting the resnet model for visual recognition. Pattern Recogn. **90**, 119–133 (2019)
17. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. Adv. Neural. Inf. Process. Syst. **34**, 12077–12090 (2021)
18. Zhang, S., Zhang, J., Xia, Y.: Transws: Transformer-based weakly supervised histology image segmentation. In: Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings. pp. 367–376. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-21014-3_38
19. Zhao, B., et al.: Triple u-net: hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. Med. Image Anal. **65**, 101786 (2020)
20. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)