Prompt-Based Grouping Transformer for Nucleus Detection and Classification

Junjia Huang[1,2], Haofeng Li[2,3], Weijun Sun[4], Xiang Wan[2], and Guanbin Li[1(✉)]

[1] School of Computer Science and Engineering, Research Institute of Sun Yat-sen University in Shenzhen, Sun Yat-sen University, Guangzhou, China
liguanbin@mail.sysu.edu.cn
[2] Shenzhen Research Institute of Big Data, Shenzhen, China
[3] The Chinese University of Hong Kong, Shenzhen, China
[4] Guangdong University of Technology, Guangzhou, China

Abstract. Automatic nuclei detection and classification can produce effective information for disease diagnosis. Most existing methods classify nuclei independently or do not make full use of the semantic similarity between nuclei and their grouping features. In this paper, we propose a novel end-to-end nuclei detection and classification framework based on a grouping transformer-based classifier. The nuclei classifier learns and updates the representations of nuclei groups and categories via hierarchically grouping the nucleus embeddings. Then the cell types are predicted with the pairwise correlations between categorical embeddings and nucleus features. For the efficiency of the fully transformer-based framework, we take the nucleus group embeddings as the input prompts of backbone, which helps harvest grouping guided features by tuning only the prompts instead of the whole backbone. Experimental results show that the proposed method significantly outperforms the existing models on three datasets.

Keywords: Nuclei classification · Prompt tuning · Clustering · Transformer

1 Introduction

Nucleus classification is to identify the cell types from digital pathology image, assisting pathologists in cancer diagnosis and prognosis [3,30]. For example, the

This work was supported in part by the Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), in part by the National Natural Science Foundation of China (No. 62102267, NO. 61976250), in part by the Guangdong Basic and Applied Basic Research Foundation (2023A1515011464, 2020B1515020048), in part by the Shenzhen Science and Technology Program (JCYJ20220818103001002, JCYJ20220530141211024), and the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen.
J. Huang and H. Li—Contribute equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43993-3_55.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 569–579, 2023.
https://doi.org/10.1007/978-3-031-43993-3_55

involvement of tumor-infiltrating lymphocytes (TILs) is a critical prognostic variable for the evaluation of breast/lung cancer [4,29]. It is a challenge to infer the nucleus types due to the diversity and unbalanced distribution of nuclei. Thus, we aim to automatically classify cell nuclei in pathological images.

A number of methods [7,10,14,23–25,33,34] have been proposed for automatic nuclei segmentation and classification. Most of them use a U-shape model [28] for training to produce dense predictions with expensive pixel-level labels. In this paper, we aim to obtain the location and category of cells, which only needs affordable labels of centroids or bounding boxes. The task can be solved by generic object detector [17,26,27], but they are usually built for everyday objects whose positions and combinations are quite random. Differently, in pathological images, experts often identify nuclear communities via their relationships and spatial distribution. Some recent methods resort to the spatial contexts among nuclei. Abousamra *et al.* [1] adopt a spatial statistical function to model the local density of cells. Hassan *et al.* [11] build a location-based graph for nuclei classification. However, the semantics similarity and dissimilarity between nucleus instances as well as the category representations have not been fully exploited.

Based on these observations, we develop a learnable Grouping Transformer based Classifier (GTC) that leverages the similarity between nuclei and their cluster representations to infer their types. Specifically, we define a number of nucleus clusters with learnable initial embeddings, and assign nucleus instances to their most correlated clusters by computing the correlations between clusters and nuclei. Next, the cluster embeddings are updated with their affiliated instances, and are further grouped into the categorical representations. Then, the cell types can be well estimated using the correlations between the nuclei and the categorical embeddings. We propose a novel fully transformer-based framework for nuclei detection and classification, by integrating a backbone, a centroid detector, and the grouping-based classifier. However, the transformer framework has a relatively large number of parameters, which could cause high costs in fine-tuning the whole model on large datasets. On the other hand, there exist domain gaps in the pathological images of different organs, staining, and institutions, which makes it necessary to fine-tune models to new applications. Thus, it is of great significance to tune our proposed transformer framework efficiently.

Inspired by the prompt tuning methods [13,16,20] which train continuous prompts with frozen pretrained models for natural language processing tasks, we propose a grouping prompt based learning strategy for efficient tuning. We prepend the embeddings of nucleus clusters to the input space and freeze the entire pre-trained transformer backbone so that these group embeddings act as prompt information to help the backbone extract grouping-aware features. Our contributions are: (1) a prompt-based grouping transformer framework for end-to-end detection and classification of nuclei; (2) a novel grouping prompt learning mechanism that exploits nucleus clusters to guide feature learning with low tuning costs; (3) Experimental results show that our method achieves the state-of-the-art on three public benchmarks.

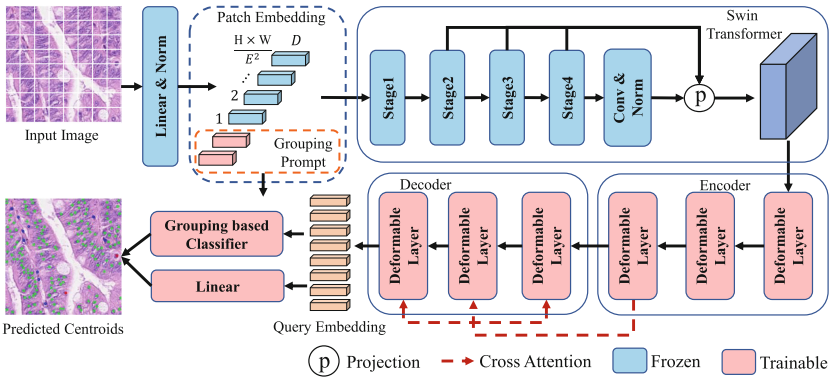

Fig. 1. The architecture of Prompt-based Grouping Transformer.

2 Methodology

As shown in Fig. 1, We propose a novel framework, Prompt-based Grouping Transformer (PGT), which directly outputs the coordinates of nuclei centroids and leverages *grouping prompts* for cell-type prediction. In the architecture, the detection and classification parts are interdependent and can be trained together. The proposed framework consists of a transformer-based nucleus detector, a grouping transformer-based classifier, and a grouping prompt learning strategy, which are presented in the following.

2.1 Transformer-Based Centroid Detector

Backbone. We adopt Swin Transformer [21] as the backbone to learn deep features. The pixel-level feature maps output from Stage 2 to Stage 4 of the backbone are extracted. Then the Stage-4 feature map is downsampled with a 3×3 convolution of stride 2 to yield another lower-resolution feature map. We obtain four feature maps in total. The channel number of each feature map is aligned via a 1×1 convolution layer and a group normalization operator.

Encoder and Decoder. The encoder and decoder have 3 deformable attention layers [35], respectively. The multi-scale feature maps output by the backbone are fed into the encoder in which the pixel-level feature vectors in all these feature maps are updated via deformable self-attention. After the attention layers, we send each feature vector into 2 fully connected (FC) layers separately to obtain the fine-grained categorical scores of each pixel. Only the Q feature vectors with the highest confidence are preserved as object embeddings and their position coordinates are recorded as reference points. Each decoder layer utilizes cross-attention to enhance the object embeddings by taking them as queries/values and the updated feature maps as keys. The enhanced query embeddings are fed into 2 FC layers to regress position offsets which are added to and refine the

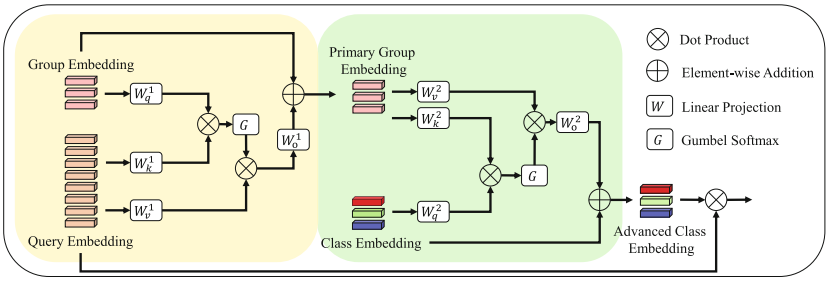

Fig. 2. The Grouping Transformer based Classifier.

reference points. The reference points output by the last decoder layer are the finally detected nucleus centroids. The last query embeddings from the decoder are sent to the proposed classifier for cell type prediction.

2.2 Grouping Transformer Based Classifier

In Fig. 2, we develop a Grouping Transformer based Classifier (GTC) that takes grouping prompts $g \in \mathbb{R}^{G\times D}$ and query embeddings $q \in \mathbb{R}^{Q\times D}$ as inputs, and yields categorical scores for each nucleus query. To divide the queries into primary groups, The similarity matrix $S \in \mathbb{R}^{G\times Q}$ between the query embeddings and the grouping prompts is built via inner product and Gumbel-Softmax [12] operation as Eq. (1):

$$S = softmax(W_q^1 g \cdot (W_k^1 q)^T + \gamma/\tau), \tag{1}$$

where W_q^1 and W_k^1 are the weights of learnable linear projections, $\gamma \in \mathbb{R}^{G\times Q}$ are i.i.d random samples drawn from the distribution $Gumbel(0, 1)$ and τ denotes the Softmax temperature. Then we utilize the hard assignment strategy [31,32] and assign the query embedding to different groups as Eq. (2):

$$\hat{S} = one\text{-}hot(argmax(S)) + S - sg(S), \tag{2}$$

where $argmax(S)$ returns a $1 \times Q$ vector, and $one\text{-}hot(\cdot)$ converts the vector to a binary $G \times Q$ matrix. sg is the stop gradient operator for better training of the one-hot function [31,32]. Then we merge the embeddings belonging to the same group into a primary group via Eq. (3):

$$g_p = g + W_o^1 \frac{\hat{S} \cdot W_v^1 q}{\sum_{i=1}^{G} \hat{S}_i} \tag{3}$$

where g_p denotes the embeddings of primary groups, W_v^1 and W_o^1 are learnable linear weights. To separate the primary groups into the cell categories, we measure the similar matrix between the primary groups g_p and learnable class embeddings $c_e \in \mathbb{R}^{C\times D}$ to yield advanced class embeddings $c_a \in \mathbb{R}^{C\times D}$, in the same way as Eq.(1)–(3). To classify each centroid query, we measure the similarity between each query embedding and the advanced class embeddings.

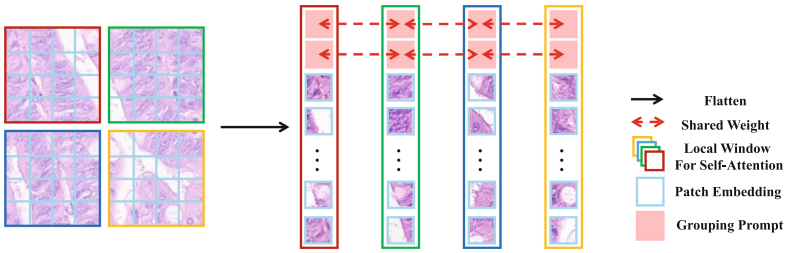

Fig. 3. The inputs with grouping prompts of the Shift-Window transformer backbone.

The category whose advanced embedding is most similar to a query, is assigned to the centroid query. The classification results $c \in \mathbb{R}^{C \times Q}$ are computed as: $c = c_a \cdot q^T$.

2.3 Loss Function

The proposed method outputs a set of centroid proposals $\{(x_q, y_q)|q \in \{1, \cdots, Q\}\}$ with a decoder layer, and their corresponding cell-type scores $\{c_q|q \in \{1, \cdots, Q\}\}$ with our proposed classifier. To compute the loss with detected centroids, we use the Hungarian algorithm [15] to assign K target centroids (ground truth) to proposal centroids and get P positive (matched) samples and $Q - P$ negative (unmatched) samples. The overall loss is defined as Eq. (4):

$$L(y, \hat{y}) = \frac{1}{P} \sum_{i=1}^{P} \left(\omega_1 ||(x_i, y_i) - (\hat{x}_i, \hat{y}_i)||_2^2 + \omega_2 FL(c_i, \hat{c}_i) \right) + \omega_3 \sum_{j=P+1}^{Q} FL(c_j, \hat{c}_j), \tag{4}$$

where $\omega_1, \omega_2, \omega_3$ are weight terms, (x_i, y_i) is the i^{th} matched centroid coordinates, $(\hat{x}_i, \hat{y}_i)$ is the target coordinates. c_i and c_j denote the categorical scores of matched and unmatched samples, respectively. As the target of unmatched samples, $\hat{c}_j$ is set to an empty category. $FL(\cdot)$ is the Focal Loss [18] for training the proposed classifier. We adopt the deep supervision strategy [35]. In the training, each decoder layer produces the side outputs of centroids and query embeddings that are fed into a GTC for classifying nuclei. For the 3 decoder layers, they yield 3 sets of detection and classification results for the loss in Eq. (4).

2.4 Grouping Prompts Based Tuning

To avoid the inefficient fine-tuning of the backbone, we propose a new and simple learning strategy based on grouping prompts, as shown in Fig. 1. We inject a set of prompt embeddings as extra input of the Swin-Transformer [21], and only tune the prompts instead of the backbone. To learn group-aware representations, we further propose to share the embeddings of prompts with those of initial groups in the proposed GTC. Such prompt embeddings are define as *Grouping Prompts*.

For a typical Swin-Transformer backbone, an input pathological image $I \in \mathbb{R}^{H \times W \times 3}$ is divided into $\frac{HW}{E^2}$ image patches of size $E \times E$. We first embed each

image patch into a D-dimensional latent space via a linear projection. Then we randomly initialize the grouping prompts $g \in \mathbb{R}^{G \times D}$ as learnable parameters, and concatenate them with the patch embeddings as input. Note that in the backbone, input patch embeddings are separated into different local windows and the grouping prompts are also inserted into each window, as shown in Fig. 3. Our proposed grouping prompt based learning consists of two phases, pre-tuning and prompt-tuning. In the pre-tuning phase, we adopt the Swin-b backbone pre-trained on ImageNet, replace the GTC head in our model (Fig. 1) with 2 FC layers, and train the overall framework without prompts and GTC. In the prompt-tuning phase, grouping prompts are added to the input of the backbone and GTC, while the backbone parameters are frozen.

3 Experiments and Results

3.1 Datasets and Implementation Details

CoNSeP[1] [10] is a colorectal nuclear dataset with three types, consisting of 41 H&E stained image tiles from 16 colorectal adenocarcinoma whole-slide images (WSIs). The WSIs are at 20× magnification and the size of the slides is 500 × 500. We split them following the official partition [1,10].

BRCA-M2C[2] [1] is a breast cancer dataset with three types and consists of 120 image tiles from 113 patients. The WSIs are at 20× magnification and the size of the slides ranges from 465 × 465 to 504 × 504. We follow the work [1] to apply the SLIC [2] algorithm to generate superpixels as instances and split them into 80/10/30 slides for training/validation/testing.

Lizard[3] [9] has 291 histology images of colon tissue from six datasets, containing nearly half a million labeled nuclei in H&E stained colon tissue. The WSIs are at 20× magnification with an average size of 1,016 × 917 pixels.

Our implementation and the setting of hyper-parameters are based on MMDetection [5]. The number of grouping prompts G is 64. Random crop, flipping, and scaling are used for data augmentation. Our method is trained with PyTorch on a 48 GB GPU (NVIDIA A100) for 12–24 h (depending on the dataset size). More details are listed in the supplementary material.

3.2 Comparison with the State-of-the-Art

The proposed method is compared with the state-of-the-art models: the existing methods for detecting and classifying cells in pathological images, i.e., HoverNet [10], MCSpatNet [1], SONNET [7], and the sate-of-the-art methods for object detection in natural images, i.e., DDOD [6], TOOD [8], DAB-DETR [19] and UperNet with ConvNeXt backbone [22]. As shown in Table 1, our method exceeds all

[1] https://warwick.ac.uk/fac/cross_fac/tia/data/hovernet/.
[2] https://github.com/TopoXLab/Dataset-BRCA-M2C/.
[3] https://warwick.ac.uk/fac/cross_fac/tia/data/lizard/.

Table 1. Comparison with existing methods on CoNSeP, BRCA-M2C and Lizard. For each dataset, we report the F-score of each class (F_c^k), the mean F-score over all classes ($\overline{F_c}$) and the detection F-score (F_d). $F_c^{Infl.}$, $F_c^{Epi.}$, $F_c^{Stro.}$, $F_c^{Neu.}$, $F_c^{Lym.}$, $F_c^{Pla.}$, $F_c^{Eos.}$ and $F_c^{Con.}$ denote the F-score for the inflammatory, epithelial, stromal, neutrophils, lymphocytes, plasma, Eosinophil and connective tissue cells, respectively. For each row, the best result is in **bold** and the second best is underlined.

	F-score↑	Hovernet [10]	DDOD [6]	TOOD [8]	MCSpatNet [1]	SONNET [7]	DAB-DETR [19]	ConvNeXt [22]	(Ours)
		2019	2021	2021	2021	2022	2022	2022	-
CoNSeP	$F_c^{Infl.}$	0.514	0.516	0.622	0.583	0.563	0.531	0.618	**0.623**
	$F_c^{Epi.}$	0.604	0.436	0.616	0.608	0.502	0.440	0.625	**0.639**
	$F_c^{Stro.}$	0.391	0.429	0.382	0.527	0.366	0.443	0.542	**0.577**
	$\overline{F_c}$	0.503	0.494	0.540	0.573	0.477	0.471	0.595	**0.613**
	F_d	0.621	0.554	0.608	0.722	0.590	0.619	0.715	**0.738**
BRCA-M2C	$F_c^{Infl.}$	0.454	0.394	0.400	0.424	0.343	0.437	0.423	**0.473**
	$F_c^{Epi.}$	0.577	0.544	0.559	0.627	0.411	0.634	0.636	**0.686**
	$F_c^{Stro.}$	0.339	0.373	0.315	0.387	0.281	0.380	0.353	**0.409**
	$\overline{F_c}$	0.457	0.437	0.425	0.479	0.345	0.484	0.471	**0.523**
	F_d	0.74	0.659	0.662	0.794	0.653	0.705	0.785	**0.799**
Lizard	$F_c^{Neu.}$	0.210	0.025	0.029	0.105	0.09	0.142	0.205	**0.301**
	$F_c^{Epi.}$	0.665	0.584	0.615	0.601	0.599	0.653	0.714	**0.762**
	$F_c^{Lym.}$	0.472	0.342	0.404	0.457	0.538	0.544	0.611	**0.664**
	$F_c^{Pla.}$	0.376	0.130	0.152	0.228	0.370	0.356	0.333	**0.403**
	$F_c^{Eos.}$	0.367	0.124	0.157	0.220	0.365	0.295	0.403	**0.457**
	$F_c^{Con.}$	0.492	0.347	0.383	0.484	0.143	0.559	0.578	**0.644**
	$\overline{F_c}$	0.430	0.259	0.290	0.349	0.351	0.425	0.474	**0.538**
	F_d	0.729	0.561	0.606	0.713	0.682	0.656	0.764	**0.779**

the other methods on three benchmarks with both detection and classification metrics. Specifically, on the CoNSeP dataset, our approach achieves 1.6% higher F-score on the detection (F_d) and 1.8% higher F-score on the classification ($\overline{F_c}$) than the second best methods MCSpatNet [1] and UperNet [22]. On BRCA-M2C dataset, our method has 0.5% higher F_d and 3.9% higher $\overline{F_c}$, compared with the second best models MCSpatNet [1] and DAB-DETR [19]. Besides, on Lizard dataset, our method outperforms UperNet [22] by more than 1.5% and 6.4% on F_d and $\overline{F_c}$, respectively. Meanwhile, we conduct t-test on CoNSeP dataset for statistical significance test. The details are listed in the supplementary material. The visual comparisons are shown in Fig. 4. With the context information from surrounding nuclei, our method effectively reduces the misclassification rate of the lymphocytes and neutrophil categories (Blue and Red).

3.3 Ablation Analysis

The strengths of the grouping transformer based classifier and the grouping prompts are verified on CoNSeP dataset, as shown in Table 2. Prompt-based Grouping Transformer (PGT) is our proposed detection and classification architecture with grouping prompts and the GTC (in Fig. 1), while the

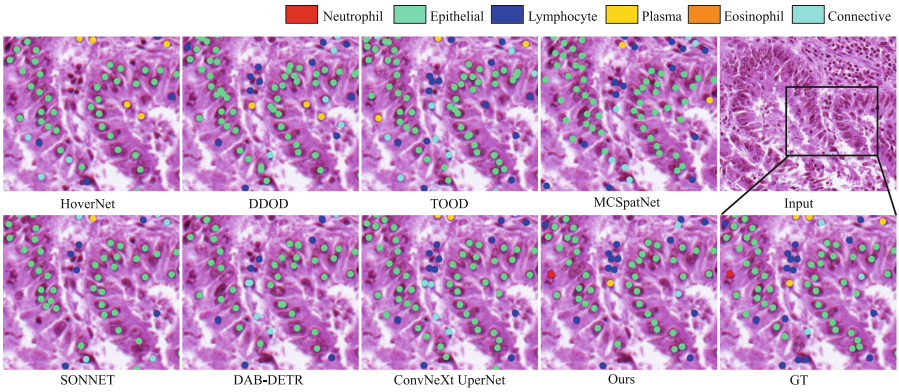

Fig. 4. The visualization results on CoNSeP dataset. (Color figure online)

Table 2. Ablation study on CoNSeP. PGT is the overall detection-classification framework. PT denotes training the network with Prompt Tuning. GTC means using the Grouping Transformer-based Classifier. * means freezing the weights of the backbone.

Methods	$F_c^{Infl.}$	$F_c^{Epi.}$	$F_c^{Stro.}$	$\overline{F_c}$	F_d	Tuned Params (M)
PGT (Full)	0.631	0.641	0.572	0.615	0.735	102.2
w/o GTC & PT (Baseline)	0.599	0.600	0.570	0.590	0.714	95.767
w/o PT*	0.602	0.604	0.558	0.588	0.713	15.321
w/o GTC*	0.615	0.604	0.564	0.594	0.724	8.895
w/ detached GTC & PT*	0.577	0.623	0.545	0.582	0.714	15.429
PGT* (Ours)	**0.623**	**0.639**	**0.577**	**0.613**	**0.738**	15.379

'Baseline' has no these two settings. PT means using naive prompt tuning. GTC means classifying nuclei with the grouping transformer. Our method achieves comparable results to the fully fine-tuning PGT with tuning only 15% parameters. Compared to the Baseline, our method yields 2.4% higher F_d and 2.3% higher $\overline{F_c}$, respectively, which shows the effective combination of the grouping classifier and prompts. 'detached GTC & PT' means that group features and prompts are independent. Our method surpasses the detached setting by 2.4% in F_d and 3.1% in $\overline{F_c}$, which suggests that sharing embeddings of groups and prompts is effective. With a frozen backbone, the performances of 'w/o PT' and 'w/o GTC' are both dropping, which verifies the strength of the prompt tuning and the GTC module, respectively.

Table 3 shows **the effect of different numbers of grouping prompts** on CoNSeP dataset. When the number of groups is small, the classification result is inferior. When the group number is large than 64, the groups may contain too few nuclei to capture their common patterns. It is suggested to set the group number to a moderate value such as 64.

Table 3. The effects of the number of grouping prompts G on CoNSeP.

F-score↑	8	16	32	64	128
F_d	0.727	0.724	0.726	**0.738**	0.723
$\overline{F_c}$	0.600	0.599	0.604	**0.613**	0.583

Table 4. $\overline{F_d}$ denotes the mean of detection F-scores of all testing images. * means p-value ≤0.05. ** means p-value ≤0.01.

F-score↑	Hovernet [10]	DDOD [6]	TOOD [8]	MCSpatNet [1]	SONNET [7]	DAT-DETR [19]	ConvNeXt -UperNet [22]	PGT* (Ours)
$\overline{F_d}$	0.615	0.545	0.625	0.706	0.582	0.615	0.698	**0.728**
p-value	0.001*	0.000**	0.000*	0.027*	0.000**	0.000*	0.012*	–

The Statistical Tests. As shown in Table 4, We calculate F_d of each testing image as sample data and conduct t-test to obtain p-values on the CoNSeP dataset. The p-values are computed between our method and the others.

4 Conclusion

We propose a new prompt-based grouping transformer framework that is fully transformer-based, and can achieve end-to-end nuclei detection and classification. In our framework, a grouping-based classifier groups nucleus features into cluster and category embeddings whose correlations with nuclei are used for identifying cell types. We further propose a novel learning scheme, which shares group embeddings with prompt tokens and extracts features guided by nuclei groups with less tuning costs. The results not only suggest that our method can obtain competitive performance on nuclei classification, but also indicate that the proposed prompt learning strategy can enhance the tuning efficiency.

References

1. Abousamra, S., et al.: Multi-class cell detection using spatial context representation. In: ICCV, pp. 4005–4014 (2021)
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
3. Aeffner, F., et al.: Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. J. Pathol. Inform. **10**(1), 9 (2019)
4. Bremnes, R.M., et al.: The role of tumor-infiltrating lymphocytes in development, progression, and prognosis of non-small cell lung cancer. J. Thorac. Oncol. **11**(6), 789–800 (2016)
5. Chen, K., et al.: MMDetection: open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)

6. Chen, Z., Yang, C., Li, Q., Zhao, F., Zha, Z.J., Wu, F.: Disentangle your dense object detector. In: ACM MM, pp. 4939–4948 (2021)
7. Doan, T.N., Song, B., Vuong, T.T., Kim, K., Kwak, J.T.: SONNET: a self-guided ordinal regression neural network for segmentation and classification of nuclei in large-scale multi-tissue histology images. JBHI **26**(7), 3218–3228 (2022)
8. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: TOOD: task-aligned one-stage object detection. In: ICCV, pp. 3490–3499 (2021)
9. Graham, S., et al.: Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In: ICCV Workshop, pp. 684–693 (2021)
10. Graham, S., et al.: HoVer-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med. Image Anal. **58**, 101563 (2019)
11. Hassan, T., Javed, S., Mahmood, A., Qaiser, T., Werghi, N., Rajpoot, N.: Nucleus classification in histology images using message passing network. Med. Image Anal. **79**, 102480 (2022)
12. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: ICLR (2017)
13. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13693, pp. 709–727. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_41
14. Kiran, I., Raza, B., Ijaz, A., Khan, M.A.: DenseRes-Unet: segmentation of overlapped/clustered nuclei from multi organ histopathology images. Comput. Biol. Med. **143**, 105267 (2022)
15. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Res. Logist. Q. **2**(1–2), 83–97 (1955)
16. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: EMNLP, Punta Cana, Dominican Republic, pp. 3045–3059. Association for Computational Linguistics (2021)
17. Li, X., Li, Q., et al.: Detection and classification of cervical exfoliated cells based on faster R-CNN. In: IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), pp. 52–57 (2019)
18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV, pp. 2980–2988 (2017)
19. Liu, S., et al.: DAB-DETR: dynamic anchor boxes are better queries for DETR. In: ICLR (2022)
20. Liu, X., et al.: P-tuning: prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Dublin, Ireland, pp. 61–68. Association for Computational Linguistics (2022)
21. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR, pp. 11976–11986 (2022)
23. Liu, Z., Wang, H., Zhang, S., Wang, G., Qi, J.: NAS-SCAM: neural architecture search-based spatial and channel joint attention module for nuclei semantic segmentation and classification. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 263–272. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_26
24. Lou, W., Li, H., Li, G., Han, X., Wan, X.: Which pixel to annotate: a label-efficient nuclei segmentation framework. IEEE TMI **42**(4), 947–958 (2022)

25. Lou, W., et al.: Multi-stream cell segmentation with low-level cues for multi-modality images. In: Competitions in Neural Information Processing Systems, pp. 1–10. PMLR (2023)
26. Nair, L.S., Prabhu, R., Sugathan, G., Gireesh, K.V., Nair, A.S.: Mitotic nuclei detection in breast histopathology images using YOLOv4. In: 12th International Conference on Computing Communication and Networking Technologies, pp. 1–5 (2021)
27. Obeid, A., Mahbub, T., Javed, S., Dias, J., Werghi, N.: NucDETR: end-to-end transformer for nucleus detection in histopathology images. In: Qin, W., Zaki, N., Zhang, F., Wu, J., Yang, F. (eds.) CMMCA 2022. LNCS, vol. 13574, pp. 47–57. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17266-3_5
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Salgado, R., et al.: The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Ann. Oncol. **26**(2), 259–271 (2015)
30. Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.W., Snead, D.R., Cree, I.A., Rajpoot, N.M.: Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE TMI **35**(5), 1196–1206 (2016)
31. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS, vol. 30 (2017)
32. Xu, J., et al.: GroupViT: semantic segmentation emerges from text supervision. In: CVPR, pp. 18134–18144 (2022)
33. Zeng, Z., Xie, W., Zhang, Y., Lu, Y.: RIC-Unet: an improved neural network based on Unet for nuclei segmentation in histology images. IEEE Access **7**, 21420–21428 (2019)
34. Zhou, H.Y., et al.: SSMD: semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation. Med. Image Anal. **72**, 102117 (2021)
35. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021)