Affine-Consistent Transformer for Multi-Class Cell Nuclei Detection
Supplementary Material

Junjia Huang[1,2†] Haofeng Li[2†] Xiang Wan[2] 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
[2]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
huangjj77@mail2.sysu.edu.cn, {lhaof,wanxiang}@sribd.cn, liguanbin@mail.sysu.edu.cn

Table 1. The detailed architecture of ConvNeXt-B backbone.

Stem	Stage 1	Stage 2	Stage 3	Stage 4
4×4, stride 4	$\begin{bmatrix} d7\times7, 96 \\ 1\times1, 384 \\ 1\times1, 96 \end{bmatrix}\times3$	$\begin{bmatrix} d7\times7, 96 \\ 1\times1, 384 \\ 1\times1, 96 \end{bmatrix}\times3$	$\begin{bmatrix} d7\times7, 96 \\ 1\times1, 384 \\ 1\times1, 96 \end{bmatrix}\times3$	$\begin{bmatrix} d7\times7, 96 \\ 1\times1, 384 \\ 1\times1, 96 \end{bmatrix}\times3$
$H/4\times W/4$	$H/4\times W/4$	$H/8\times W/8$	$H/16\times W/16$	$H/32\times W/32$

A. Complete Implementation Details

Backbone Network Architecture. We use ConvNeXt-B [12] as the backbone to extract image features. The details of the ConvNeXt-B are shown in Table 1. The 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 obtain another feature map of a smaller size $H/64 \times W/64$. We extract four feature maps in total. The channel number of each feature map is aligned via a convolutional layer and a normalization layer to 256. For fair comparison, we implement the existing methods and use the same backbone as their own paper/code. Their backbones are reported in Table 2. All the backbones are pretrained on ImageNet. Meanwhile, we calculate F_d of each testing image as sample data, and conduct t-test to obtain p-values on CoNSeP dataset, as shown in Table 2. Node that the mean detection F1 of nuclei in each test images $\overline{F_d} = \sum_{n=1}^{N} \frac{2TP_d^n}{2TP_d^n+FP_d^n+FN_d^n}/N$ is different from the F_d reported in the draft, which is formulated as $F_d = \frac{2\sum_{n=1}^{N} TP_d^n}{2\sum_{n=1}^{N} TP_d^n+\sum_{n=1}^{N} FP_d^n+\sum_{n=1}^{N} FN_d^n}$.

Training Details and Hyper-parameters. We build our model with the MMDetection framework [3]. During the training, for each training image, we apply data augmentation to obtain multiple samples. We feed an augmented sample into two branches. In the local branch, the sample image first passes through an Adaptive Affine Transformer (AAT) module, generating M affine transformation matrices. The grid sampler in AAT generates warped images of the same size as the input, according to the affine transformation matrices. At the early stage of the training, the local network is supervised by the warped GTs, and updates the global network via the exponential moving average (EMA) strategy. After γ steps, we introduce the prediction of the global network as additional supervision with a weight α. Using the global loss at the start will introduce excessive noise and cause the non-convergence of the model training. We provide the notation, description and values of the hyper-parameters for three datasets in Table 3. Most of these hyper-parameters do not need complicated tuning, but are set following the MMDetection framework [3].

Data augmentation. Before applying data augmentation, each pathology image is divided into non-overlapping image patches. Then we randomly flip and scale each sample into multiple sizes from 672 to 800. This step is mainly to improve the number and diversity of the data. During inference, we use sliding windows with non-overlapping image patches. Each sliding window of a testing image is sent into the global network for prediction.

B. Ablation Analysis

Varying thresholds of maximum category probability. Table 4 shows the results of the varying threshold of maximum category probability on the *Lizard* dataset [8]. The threshold, denoted as t, is used to select candidate points from the prediction of the global sub-network, and then these candidate points act as supervision signals to train the local sub-network. We experiment with the thresholds from 0.1 to 0.7. The results show that only using high-

[†] Junjia Huang and Haofeng Li contribute equally to this work.
[*] Guanbin Li is the corresponding author.

Table 2. The statistical significance test between the existing methods and ours on CoNSeP dataset. $\overline{F_d}$ denotes the mean of detection F-scores of all testing images. * means p-value≤0.05. ** means p-value≤0.01.

F-score↑	Hovernet [9]	DDOD [4]	TOOD [6]	MCSpatNet [1]	SONNET [5]	DAT-DETR [11]	ConvNeXt [12]	AC-Former(Ours)
$\overline{F_d}$	0.615	0.545	0.625	0.706	0.582	0.615	0.698	**0.726**
p-value	0.002*	0.000**	0.002*	0.032*	0.000**	0.002*	0.036*	-
Backbone	ResNet50	ConvNeXt	ResNet101	Vgg16	EfficientB0	ConvNeXt	ConvNeXt	ConvNeXt

Table 3. Descriptions and values of the hyper-parameters used in the *CoNSeP*, *BRCA* and *Lizard* benchmarks. Most of the hyper-parameters are set following the default values in the MMDetection framework [3], which do not require complicated tuning.

Hyper-parameter	Description	*CoNSeP*	*BRCA*	*Lizard*
M	Number of warper images	4	4	4
t	Threshold of maximum category probability	0.3	0.3	0.3
α	Global loss weight	0.1	0.1	0.1
β_1	Hungarian cost matrix weight for regression	5	5	5
β_2	Hungarian cost matrix weight for classification	2	2	2
λ_1	Focal Loss balanced factor	0.25	0.25	0.25
λ_2	Focal Loss focusing parameter	2	2	2
ω_1	Positive samples weight for regression	5	5	5
ω_2	Positive samples weight for classification	2	2	2
ω_3	Negative samples weight for classification	2	2	2
e	EMA rate	0.999	0.999	0.999
γ	Training steps with only local loss	1600	1600	3200
n	Number of query object	1000	1000	1000

Table 4. Effect of the threshold of maximum category probability t on Lizard dataset.

t	0.1	0.2	0.3	0.4	0.5	0.6	0.7
$F_c^{Neu.}$	0.121	0.123	**0.270**	0.204	0.240	0.069	0.009
$F_c^{Epi.}$	0.748	0.753	**0.788**	0.785	0.760	0.739	0.552
$F_c^{Lym.}$	0.641	0.653	0.690	**0.693**	0.654	0.501	0.139
$F_c^{Pla.}$	0.415	0.437	**0.475**	0.471	0.449	0.342	0.148
$F_c^{Eos.}$	0.412	0.399	**0.450**	0.415	416	0.293	0.012
$F_c^{Con.}$	0.610	0.630	**0.671**	0.656	0.634	0.396	0.035
$\overline{F_c}$	0.491	0.498	**0.557**	0.537	0.526	0.392	0.149
F_d	0.732	0.744	**0.782**	0.775	0.753	0.621	0.356

Table 5. Effect of the amount of warped images M on Lizard dataset.

M	$F_c^{Neu.}$	$F_c^{Epi.}$	$F_c^{Lym.}$	$F_c^{Pla.}$	$F_c^{Eos.}$	$F_c^{Con.}$	$\overline{F_c}$	F_d
1	0.119	0.714	0.637	0.402	0.375	0.619	0.478	0.713
2	0.163	0.713	0.633	0.427	0.396	0.640	0.495	0.717
3	0.199	0.755	0.670	0.465	0.438	0.661	0.531	0.761
4	**0.270**	**0.788**	**0.690**	**0.475**	**0.450**	**0.671**	**0.557**	**0.782**
5	0.064	0.707	0.627	0.417	0.353	0.620	0.465	0.705

Table 6. Effect of the training steps γ with only the local loss on Lizard dataset.

γ	$F_c^{Neu.}$	$F_c^{Epi.}$	$F_c^{Lym.}$	$F_c^{Pla.}$	$F_c^{E}os.$	$F_c^{Con.}$	$\overline{F_c}$	F_d
8k	0.209	0.752	0.635	0.461	0.451	0.647	0.526	0.755
16k	0.238	0.767	0.669	0.458	**0.470**	0.652	0.546	0.769
32k	**0.270**	**0.788**	**0.690**	0.475	0.450	0.671	**0.557**	**0.782**
48k	0.255	0.750	0.670	**0.478**	0.466	0.654	0.546	0.765

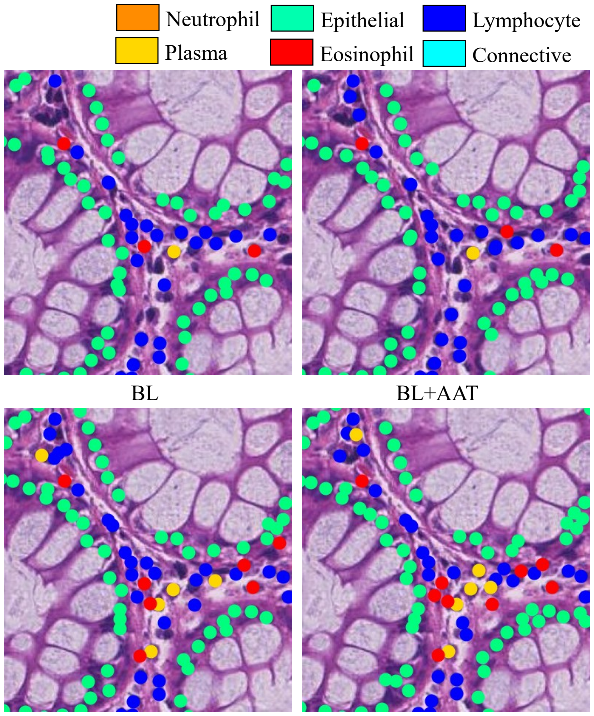
score candidate points as supervision will reduce performance. When $t \leq 0.2$, the number of candidate points is likely to exceed the number of the GT centroids, and then the local network could pay too much attention to the noisy prediction of the global network. It could result in a drop of performance as shown in Table 4. The best results are achieved by setting t to 0.3.

Amount of warped images. Table 5 shows the detailed F-scores for each nucleus category on *Lizard* when using different numbers of warped images. Setting M to 4 obtains the highest $\overline{F_c}$ 0.557. When setting M to a small value (< 4), the warped image samples could be insufficient, which results in a drop of 2%-8% F-scores in classification. Setting M to a large value (=5) could make the training process too slow to converge, which causes a drop of 9% in classification F-score in comparison to the model with $M = 4$.

The training steps with local loss only. During training, we first train the network with the local branch loss only. After γ steps, the overall loss (including the local and the global losses) is optimized with $\alpha = 0.1$. Table 6 shows the results of varying training steps with local loss only on *Lizard*. When γ is set to 8000 or 16000, the local network has not well trained and its corresponding global network could output noisy results. That could make the global loss unpromising, and lead to a drop of 1%-3% classification F-score in comparison to setting $\gamma = 32000$. Setting γ to a large value (48000) shows a slight decrease of 1.1% $\overline{F_c}$. It may be due to that the model over-fits the GT to some degree.

Table 7. Effect of focal loss and local network branch. The results are obtained on the *CoNSeP* and *Lizard* datasets. '*Local Network*' is our proposed method using the local branch for inference while '*Global (Ours)*' using the global branch for prediction. Both of them utilize a focal Hungarian loss. '*DETR Loss*' denotes the results of training our proposed dual-branch model with the loss in DETR [2] instead of a focal loss. The best result in each column is in bold type.

Methods	CoNSeP					Lizard							
	$F_c^{Infl.}$	$F_c^{Epi.}$	$F_c^{Stro.}$	$\overline{F_c}$	F_d	$F_c^{Neu.}$	$F_c^{Epi.}$	$F_c^{Lym.}$	$F_c^{Pla.}$	$F_c^{Eos.}$	$F_c^{Con.}$	$\overline{F_c}$	F_d
DETR Loss	0.632	0.611	0.533	0.592	0.735	0.115	0.704	0.595	0.371	0.251	0.546	0.430	0.729
Local Network	0.623	0.634	0.564	0.607	0.739	0.191	0.776	0.688	0.448	0.417	0.657	0.529	0.774
Global (Ours)	**0.635**	**0.635**	**0.568**	**0.613**	**0.739**	**0.270**	**0.788**	**0.690**	**0.475**	**0.450**	**0.671**	**0.557**	**0.782**



BL

BL+AAT

BL+AAT+GL

GT

Figure 1. Qualitative comparison of the ablation study on *Lizard*. Six types of cells are marked with dilated nucleus centroids in six different colors. '*BL*' is our baseline of a single network branch. '*AAT*' denotes our proposed module Adaptive Affine Transformer. '*BL+AAT*' is a dual-branch model that uses the AAT to warp images, and the EMA strategy to update the global branch. '*GL*' means the global loss that is computed between the local and global predictions. '*BL+AAT+GL*' denotes our proposed method.

The performance of local network. After training our proposed dual-branch model, any one of the two branches can be used to infer testing images. In Table 7, '*Local Network*' and '*Global (Ours)*' denote utilizing the local branch and the global branch of our trained dual-branch model to predict results, respectively. As shown in Table 7, '*Global (Ours)*' outperforms the '*Local Network*' by 0.8% and 2.8% F-scores in detection and classification tasks on *Lizard*. Thus, our proposed method adopts the global sub-network for inference.

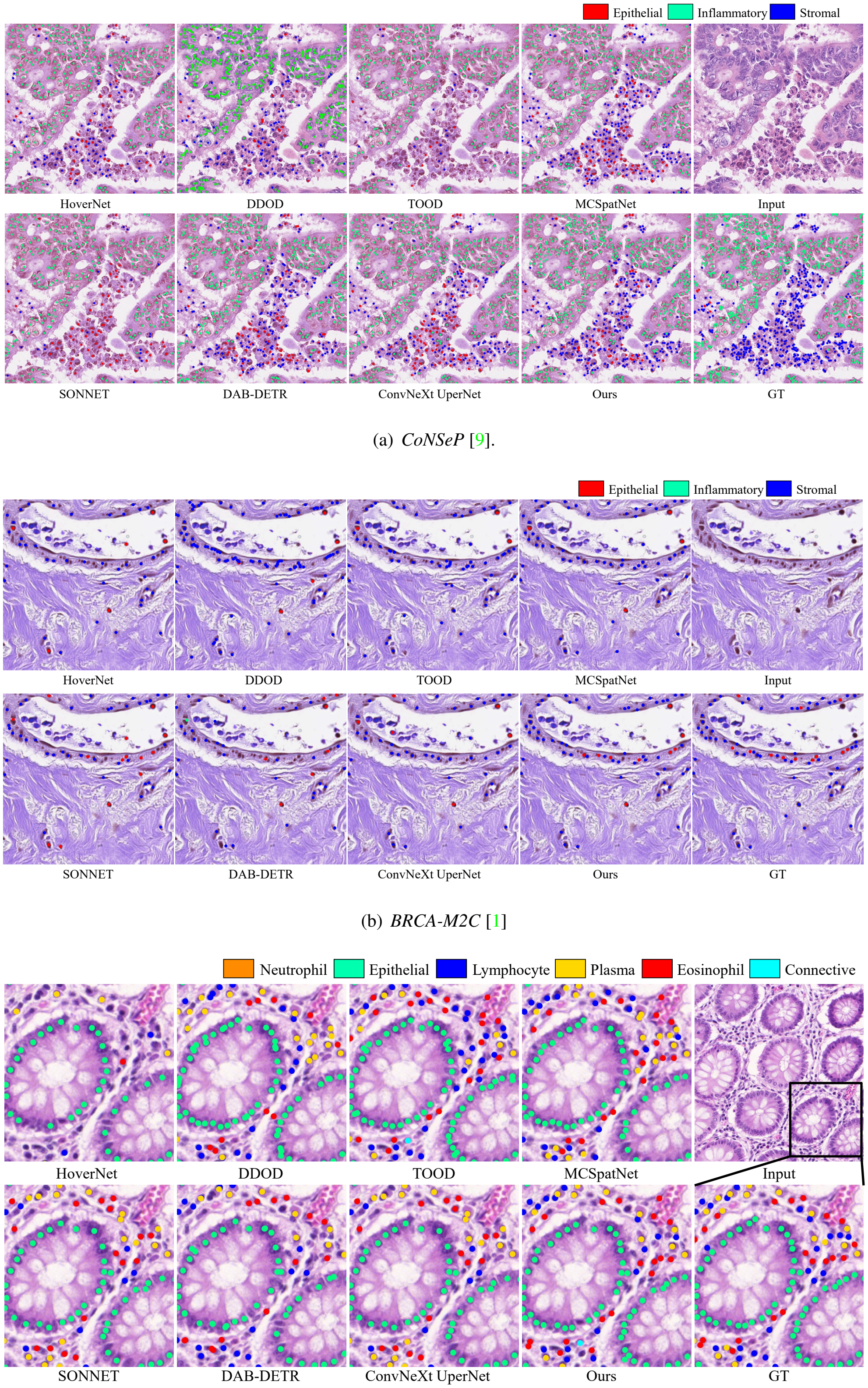
The effectiveness of focal loss. The Focal Loss [10] can effectively address the problem of unbalanced categories and improve the classification performance. In Table 7, '*DETR Loss*' means training our proposed dual-branch model with the naive Hungarian loss [2] instead of a focal loss, while '*Global (Ours)*' means training the dual-branch model with a focal Hungarian loss as described in our method section. Both the '*DETR Loss*' and '*Global (Ours)*' use the global branch for inference.

Comparing '*Global (Ours)*' with '*DETR Loss*' in Table 7, we find that the conventional cross-entropy loss without focal loss is difficult to deal with unbalanced classes. Especially on *Lizard* [8], the classification performance of neutrophils is much worse than those of other categories. On *CoNSeP* [9], even with the same-level F-score of detection, our method '*Global (Ours)*' still exceeds the model using '*DETR-Loss*' by 2.1% F-score in the classification task. These results suggest that the focal loss in our proposed method '*Global (Ours)*' can effectively ease the issue of class imbalance.

C. Qualitative Comparison

Figure 1 shows the qualitative comparison among '*BL*', '*BL+AAT*' and '*BL+AAT+GL*' methods. '*BL*' is our baseline that contains a single network branch and is trained with original pathological images. '*AAT*' denotes our proposed Adaptive Affine Transformer. '*BL+AAT*' is a dual-branch model that uses the AAT to yield warped images for the local branch, and the EMA strategy to update the global branch. '*BL+AAT+GL*' denotes our finally proposed method. GL means global loss that is computed between the local and global predictions. The results indicate that using the AAT module and the global loss can make the model more robust to identify the hard samples in a dense distribution of nuclei. For example, the Eosinophil are scarce in Figure 1 and are surrounded by numerous nuclei of other types. Note that our proposed method '*BL+AAT+GL*' can locate and recognize most of these hard samples in Figure 1.

Figure 2 displays the qualitative comparison on three dataset. Our method performs better than existing approaches in the cases of both densely distributed cells (Figure 2(a)) and sparse cells (Figure 2(b)).



(a) *CoNSeP* [9].

(b) *BRCA-M2C* [1]

(c) *Lizard* [8].

Figure 2. Qualitative comparison on the *CoNSeP*, *BRCA* and *Lizard* datasets. We compare the proposed method with existing state-of-the-art approaches, including Hovernet [9], SONNET [5]), DAB-DETR [11], TOOD [6], Yolox [7], UperNet [13] with ConvNeXt [12] backbone and MCSpatNet [1].

References

[1] Shahira Abousamra, David Belinsky, John Van Arnam, Felicia Allard, Eric Yee, Rajarsi Gupta, Tahsin Kurc, Dimitris Samaras, Joel Saltz, and Chao Chen. Multi-class cell detection using spatial context representation. In *ICCV*, pages 4005–4014, 2021. 2, 4

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 3

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1, 2

[4] Zehui Chen, Chenhongyi Yang, Qiaofei Li, Feng Zhao, Zheng-Jun Zha, and Feng Wu. Disentangle your dense object detector. In *ACM MM*, pages 4939–4948, 2021. 2

[5] Tan NN Doan, Boram Song, Trinh TL Vuong, Kyungeun Kim, and Jin T Kwak. SONNET: A self-guided ordinal regression neural network for segmentation and classification of nuclei in large-scale multi-tissue histology images. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3218–3228, 2022. 2, 4

[6] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. TOOD: Task-aligned one-stage object detection. In *ICCV*, pages 3490–3499, 2021. 2, 4

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 4

[8] Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In *ICCVW*, pages 684–693, 2021. 1, 3, 4

[9] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. 2, 3, 4

[10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 3

[11] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *ICLR*, 2022. 2, 4

[12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 1, 2, 4

[13] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 4