Affine-Consistent Transformer for Multi-Class Cell Nuclei Detection

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

Abstract

Multi-class cell nuclei detection is a fundamental prerequisite in the diagnosis of histopathology. It is critical to efficiently locate and identify cells with diverse morphology and distributions in digital pathological images. Most existing methods take complex intermediate representations as learning targets and rely on inflexible post-refinements while paying less attention to various cell density and fields of view. In this paper, we propose a novel Affine-Consistent Transformer (AC-Former), which directly yields a sequence of nucleus positions and is trained collaboratively through two sub-networks, a global and a local network. The local branch learns to infer distorted input images of smaller scales while the global network outputs the large-scale predictions as extra supervision signals. We further introduce an Adaptive Affine Transformer (AAT) module, which can automatically learn the key spatial transformations to warp original images for local network training. The AAT module works by learning to capture the transformed image regions that are more valuable for training the model. Experimental results demonstrate that the proposed method significantly outperforms existing state-of-the-art algorithms on various benchmarks.

1. Introduction

A major task of pathologists is to make a diagnosis with digital pathological images, which are obtained by scanning tissue slides with a whole-slide scanner [33, 42]. In this process, a pathologist is required to provide the grading of tumors and to classify benign and malignant diseases [31, 15], by locating and identifying certain histological structures such as lymphocytes, cancer nuclei, and glands. In some applications, instead of locating pixels on each nucleus boundary, it could be useful to quantify the different categories of cells. For example, the counts of tumor cells and lymphocytes have been utilized as an effective prognostic marker [12]. Thus, in this paper, we do not focus on predicting the nucleus sizes or boundaries, but only aim at inferring the types and rough locations of cell nuclei in digital slide images, following the previous work [1].

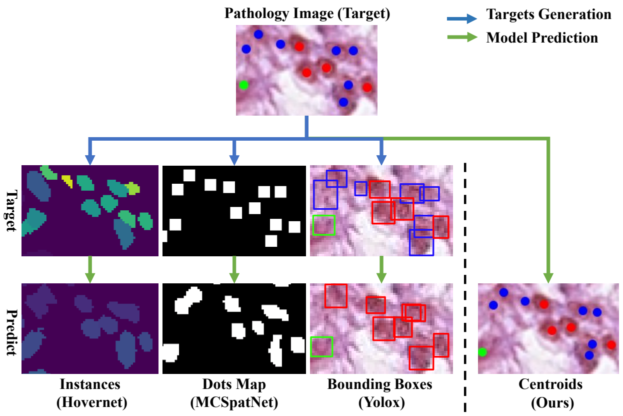

Figure 1. The visual comparison of predictions and training targets between existing methods and ours. Different types of nuclei are marked by red, green and blue boxes or centroids. Our method can predict a sequence of position coordinates and categorical labels of nuclei directly from an input pathological image.

In the early stage, automatic nucleus detection and classification have been achieved by handcrafted features based methods[35, 3, 49, 4]. These methods lack sufficient accuracy and generalization, while deep learning (DL) models can tackle these issues via learning robust representations.

For nuclei detection, existing DL methods can be divided into three groups, according to the different forms of prediction targets. As Figure 1 shows, the first group is to outline the contour or to locate the region of each single nucleus via pixel-wise prediction [14, 9, 16, 41, 29, 30]. These methods rely on the high-quality boundary annotation of nuclei that are expensive and time-consuming. The second group [17, 1, 38] is to pixel-wisely predict centroids or dilated centroids ('Dots Map' in Figure 1), by convert-

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

ing the detection into a binary segmentation task. Due to the diversity of cell density, the boundaries between adjacent nuclei are often confused, which makes these methods fail to segment adherent nuclei and results in missed detection. The third group is to predict the bounding boxes of nuclei [39, 23, 56] based on the anchors of pixels, but the performance of these methods are affected by anchor parameters and post-processing. Adjacent nuclei with unclear boundaries increase the difficulty of detecting bounding boxes. Thus, we propose to convert the nuclei detection into a task of directly predicting a set of cell positions and categories.

Besides, the diversity of image scale and nuclei density causes more difficulties in the detection and classification tasks. Higher magnification levels or scaling factors could lead to a smaller field of view and more sparse distribution of nuclei. We claim that it is essential to develop a robust model for different image scales. Some existing works [53, 50] employ multi-scale deep learning architectures or simply unify the scale by dividing patches, which does not take the prediction consistency among multiple scales into consideration.

To avoid the synthesis of indirect learning targets, we consider formulating the nucleus localization and classification problem as a sequence generation task. A transformer-based framework is adopted to decode a list of position coordinates and category labels of nuclei in a direct manner. To adapt to diverse scales, we further split the transformer framework into two network branches, a local network and a global one, which aim to infer global-scale images and their local-scale views, respectively. The local network is not only supervised by the ground-truth annotations but also guided by the global network that captures broader contextual information from the large-scale input. Thus, the well-trained networks could accommodate to diverse fields of view. To compute the training losses, a matching algorithm is utilized to assign each target nucleus to a nucleus proposal in the predicted sequence. Importantly, we claim that not all local image regions are equal for training a scale-consistent nuclei detection model. Therefore, we propose a novel adaptive affine transformer that predicts a series of affine transformation parameters to harvest the key local-scale inputs for improving the global-local training. Since our proposed framework is trained to deal with various fields of view and distributions of cells, it has the potential to well separate the densely distributed nuclei from each other and to reduce the missed detection rate.

In short, our major contributions are summarized as three folds:

- We introduce a novel Adaptive Affine Transformer that automatically learns to augment effective multi-scale samples for training.
- We propose an Affine-Consistent Transformer framework for nuclei detection. Its local branch learns to output a set of nucleus-level predictions with small-scale inputs, guided by the global branch with a large-scale input.
- We conduct extensive experiments and demonstrate the state-of-the-art performance of our method on three widely-used benchmarks.

2. Related Work

Nuclei Detection and Classification Many methods have been developed to locate and identify cell nuclei. According to the different representations of prediction targets, they can be divided into three types: instance based, dots map based and bounding box based methods. The instance based methods [14, 9, 16, 52] first use neural networks to output pixel-level predictions such as semantic segmentation maps and distance maps, and then obtain the mask of each single nucleus via some post-processing methods like watershed algorithm [9]. Some works [54, 27, 34] detect nuclei with the generic instance segmentation methods proposed for natural images. However, these instance-based approaches require expensive pixel-level annotations of each nucleus boundary, while our method only needs lower-cost annotations of nucleus position for the detection task.

For the dots map based methods, they either regress a pixel-wise density map to locate the nuclei at the peak [48, 17, 11, 1, 45, 46], or classify image patches with sliding windows [37, 50]. Abousamra *et al.* [1] formulate the nuclei detection problem as a binary segmentation task of dilated nucleus centroids, while Wang *et al.* [45] extract local features and performs adversarial alignment for domain adaptive nuclei detection. Although compelling models have been proposed, these dots map based methods could fail to separate two adjacent nuclei when dealing with intensively distributed cells. Wu *et al.* [46] detect 3D centroids by estimating the intensity peaks of voting regions, which is different from our method that updates and outputs the centroid coordinates directly.

Some other methods [51, 39, 22, 44, 23, 47] utilize the bounding boxes of cells as training targets. Sun *et al.* [39] compute discriminative features based on similarity learning to boost the classification performance, while Liang *et al.* [23] propose a GA-RPN module integrating the guided anchoring (GA) into a region proposal network (RPN) to generate more suitable object proposals. Wu *et al.* [47] detect nuclei centroids with an RCNN-SliceNet that still depends on the pipeline of producing and suppressing region proposals. These methods usually take a large number of anchor boxes as candidates and adopt the non-maximum suppression (NMS) algorithm as post-processing to screen out the highly overlapping boxes. In this paper, we avoid the

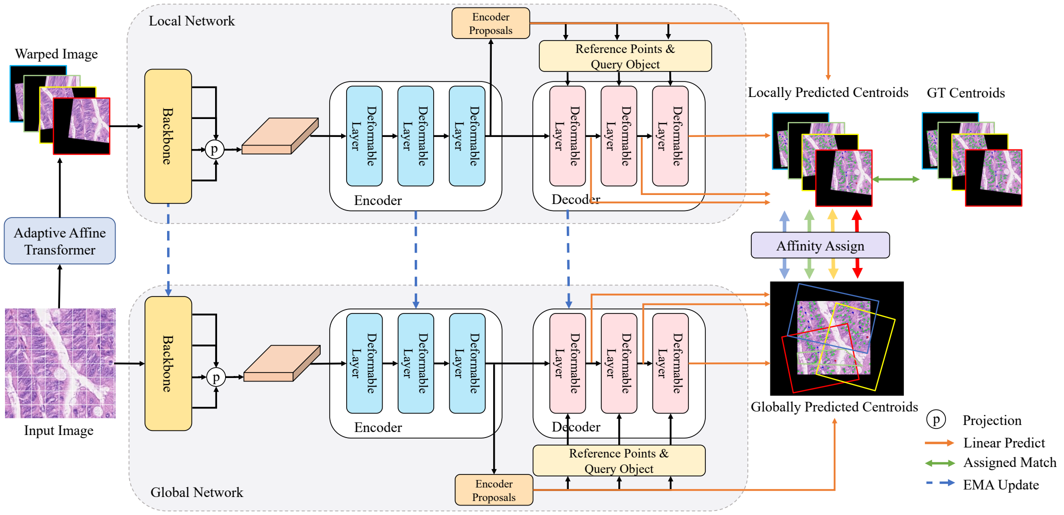

Figure 2. The overall framework of Affine-Consistent Transformer. An input training image is sent through an Adaptive Affine Transformer module to generate a series of affine transformed images. The original image and the transformed images are fed into two associated networks respectively to produce two relevant point sets with categorical scores. After one-to-one matching these two point sets, the Hungarian loss is calculated to update the local network. The output of the global network is used to coordinate the scale consistency of the local network. The global network is updated via the exponential mean average (EMA) manner.

tedious inference process of existing methods, and exploit transformers to directly decode the positions and category scores of nuclei.

Transformer-based Object Detection Object detection aims to predict the bounding boxes and category labels of objects in an image. Transformer-based methods [5, 40, 25, 55, 21, 18] view object detection as a set prediction problem, using transformer modules [43] to directly output a final set of object-level predictions without further post-processing. Carion *et al.* first propose a fully end-to-end object detector DETR [5] but it suffers from slow convergence and limited spatial resolution of features. In follow-up works, Zhu *et al.* propose Deformable DETR [57] attending to a small set of key sampling points instead of all possible pixels. Different from existing transformer-based detection models, we develop a new transformer framework that not only produces affine transformation matrices for learnable augmentation, but also adapts to nuclei detection via predicting the nucleus centroids as a sequence of points.

3. Methodology

In this paper, we propose an Affine-Consistent Transformer (AC-Former). The workflow of the proposed AC-Former is shown in Figure 2. During the training, a local and a global networks cooperate with each other. The local network is trained by both the nucleus centroids from the warped images and the predicted centroids from the global network to ensure the scale consistency. The global network is continuously updated by the local network via the exponential moving average (EMA) strategy. We first introduce the proposed AC-Former framework and then describe the essential designs: an adaptive affine transformer and the local-global network architecture.

3.1. Affine-Consistent Transformer

The proposed method learns to directly produce the centroid locations of nuclei with the corresponding confidence scores. Given a pathological image of size $H \times W \times 3$, we use the proposed adaptive affine transformer to generate warped images $I_i, i \in \{1, \cdots, M\}$. And then the local network locates the nucleus positions to generate the coordinates and category scores from the encoder and decoder. In the training stage, for the i^{th} warped image, I_i the local network outputs $D+1$ sets: $y^i = \{y^{ij} | j \in \{0, \cdots, D\}\}$, where D is the number of layers in the decoder and y^{ij} is a set of predicted centroids with coordinates and categorical scores. y^{i0} is the encoder output while y^{ij} with $j > 0$ is the j^{th} decoder output. Meanwhile, the global network infers the whole input image to obtain the globally predicted results $\bar{y}$. Thus, the overall loss is defined as:

$$\mathcal{L}(y, \hat{y}, \bar{y}) = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=0}^{D} (\mathcal{L}_m(y^{ij}, \hat{y}^i) + \alpha \mathcal{L}_m(y^{ij}, \bar{y}^{ij})), \tag{1}$$

where $\mathcal{L}_m$ denotes the Hungarian loss [5] between one-to-one aligned nuclei, $\hat{y}$ is the ground truth of warped centroids. $\hat{y}^i$ represents the result of aligning the ground truth with the i^{th} locally warped image, while $\bar{y}^{ij}$ is obtained by aligning the j^{th} globally predicted set with the i^{th} distorted image. α is a weight term to balance the global branch loss

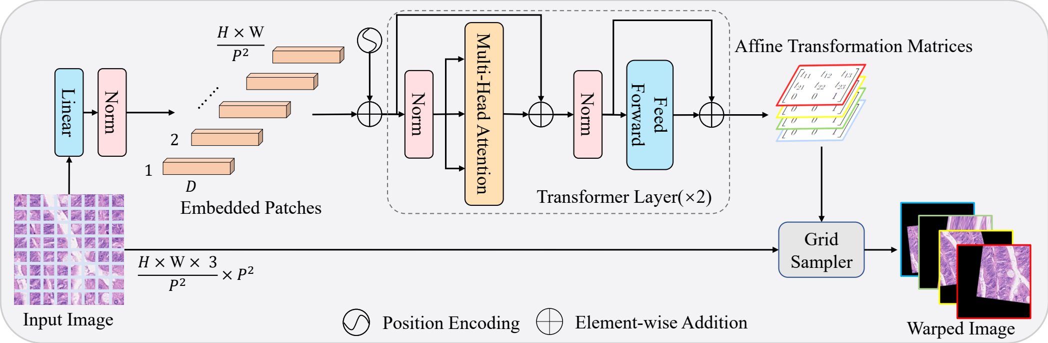

Figure 3. The detailed architecture of the Adaptive Affine Transformer. The affine transformation matrices are generated by self-attention modules, and will be adopted to perform affine warping on input images with a grid sampler. The overall structure is derivable and it will learn to automatically augment image patches that benefit the model training.

and the local loss based on ground-truths.

We introduce the prediction of the global network for supervision to enhance the spatial scale consistency of the trained model. In the centroids sets predicted by the global network, since the redundancy can not be eliminated by the non-maximum suppression (NMS) [32], we use the maximum category probability as the evaluation scores of centroids to select candidates with a threshold t. Only the centroids whose maximum category probability is not smaller than t are reserved. Empirically we set $t = 0.3$. During the inference stage, we simply employ the global network to infer a testing image, and produce $D + 1$ sets of nucleus centroids. Only the last set is adopted as the final prediction of the overall method.

3.2. Adaptive Affine Transformer

Since the distribution of nuclei is not uniform, local views obtained by random transformations are not equally beneficial for the training. Thus, we propose to learn to synthesize the local views with a trainable model. To embrace long-range contexts, we develop a transformer-based model, Adaptive Affine Transformer (AAT) to analyze the input image and automatically generate the parameters of an affine transformation.

As shown in Figure 3, the proposed adaptive affine transformer divides an input image into $\frac{HW}{P^2}$ patches, feeds the flattened patches to a linear projection layer, and obtains the embedded patches of size $\frac{HW}{P^2} \times E$, where E denotes the number of embedding dimensions and P is the size of each image patch. After that, the embedded patches are added with their sinusoidal position embeddings, and then passed through a transformer [43] encoder with 2 layers. These transformer layers compute the global dependency between image patches with the multi-head attention mechanism. The output is linearly projected into M affine transformation matrices $\{A_1, ..., A_M\}$. We do not adopt perspective transformations but only affine transformations. The point-wise transformation process can be formulated as :

$$\sigma(u, v, A) = A \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix}, \tag{2}$$

where u and v are the coordinates of a sampled point in the input image. u' and v' are the target coordinates after the affine transformation with A.

To avoid generating unidentifiable images of low-quality, we constrain the matrix parameters in a moderate range. For scaling factors in the matrix, t_{11}, t_{22}, we clamp them into [0.2, 2]; for translation, t_{13}, t_{23} are clamped into [-0.5, 0.5]; for rotation & shear, t_{12}, t_{21} are clamped into [-1, 1]. Then we feed these M matrices with an input image into a differentiable grid sampler [19] to yield M distorted images. Let $V(i, j)$ return the pixel value at the position (i, j) of the input image (in Figure 3). $I(u', v')$ denotes the pixel value at the position (u', v') of the warped image, and can be computed by bilinear interpolation:

$$I(u', v') = \begin{bmatrix} 1-u & u \end{bmatrix} \begin{bmatrix} V(\lfloor u \rfloor, \lfloor v \rfloor) & V(\lfloor u \rfloor, \lceil v \rceil) \\ V(\lceil u \rceil, \lfloor v \rfloor) & V(\lceil u \rceil, \lceil v \rceil) \end{bmatrix} \begin{bmatrix} 1-v \\ v \end{bmatrix} \tag{3}$$

To produce the pixel values of M warped images, each integer coordinate of the i^{th} warped image is set to (u', v') and its corresponding (u, v) can be solved via Eq. 2, given the i^{th} affine transformation matrix A_i. Then the pixel value at the position (u', v') of a warped image can be calculated via Eq. 3. To yield $\hat{y}^i$ in Eq. 1, that is to say, to align $\hat{y}$ with i^{th} distorted image, we set each centroid coordinate in $\hat{y}$ to (u, v), and compute the corresponding (u', v') via Eq. 2 with A_i. The resulted coordinates (u', v') out of the range $([0, H], [0, W])$ are removed from $\hat{y}^i$. Computing $\bar{y}^{ij}$, namely, aligning the j^{th} predicted set of the global network with the i^{th} warped image, is performed in a similar manner. The process can be formulated as: $\bar{y}^{ij} = \{\sigma(\bar{y}^j_q, A_i) | q \in \{1, \cdots, |\bar{y}^j|\}\}$ where $\bar{y}^j_q$ denotes the q^{th} centroid coordinate of the globally predicted set $\bar{y}^j$.

The proposed adaptive affine transformer module enables the network to adaptively learn enhanced features that are robust for spatial transformations.

3.3. Global-Local Network Architecture

Backbone As illustrated in Figure 2, the global or the local network consists of 3 components: a backbone, an encoder and a decoder. We adopt ConvNeXt-B [26] as the backbone that acts as a deep feature extractor. The backbone yields a list of feature maps of different scales.

Encoder The encoder has 3 deformable attention [57] layers and 2 fully-connected (FC) layers. The input of the encoder is a stack of multi-scale feature maps output by the backbone. For each attention layer in the encoder, its input and output have the same shape, and its query elements are set to all pixel-level feature vectors in its multi-scale input feature maps. After the attention layers, each feature vector is separately sent into the FC layers, which produce the categorical scores and the coordinate offsets relative to the feature position. Only n feature vectors with the highest confidence are preserved as object query embeddings and their coordinates are recorded as the reference points.

Decoder The decoder has 3 layers and each layer contains a deformable attention module and 2 FC layers. Different from the encoder, the deformable attention module in the decoder takes the object queries from the encoder as query elements. After the attention enhances the object query embeddings, the two FC layers predict a 2D offset and categorical scores for each object query, respectively. The 2D offset is added to and updates the corresponding reference point. For deep supervision [36], the 3 decoder layers provide 3 sets of predictions, respectively.

The loss is calculated between each predicted set of centroid proposals and a set of target centroids that could be ground truths or predicted by the global network. The number of centroid proposals C is far more than that of target centroids T. Let $y_c = \{(u_c, v_c), c_c\}$ denote a centroid proposal, and $y_t = \{(u_t, v_t), c_t\}$ denote a centroid target. A centroid proposal or centroid target consists of the coordinates and category scores. A pair-wise cost matrix $\mathcal{E}$ is computed by measuring the cost between each proposal and each target:

$$\mathcal{E}(y_c, y_t) = \beta_1 ||(u_c, v_c) - (u_t, v_t)||_2^2 + \beta_2 l_{focal}(c_c, c_t), \quad (4)$$

where $l_{focal}(\cdot)$ denotes the Focal Loss [24]. β_1 and β_2 provide a balance between the position regression and classification. The Focal Loss is used to mitigate the class imbalance of the nuclei classification task. It is defined as:

$$l_{focal}(c_c, c_t) = \frac{1}{K} \sum_{k=1}^{K} -\lambda_1 (1 - c_{ck} \cdot c_{tk})^{\lambda_2} c_{tk} log(c_{ck}), \quad (5)$$

where λ_1 and λ_2 are the balanced factors and the focusing parameter, K denotes the number of classes. Then we conduct the association with the Hungarian algorithm [20] based on the $C \times T$ cost matrix $\mathcal{E}$, and obtain T matching positives and $C - T$ negatives. Our goal is to narrow the coordinate and categorical difference between the positive proposals and their corresponding target, and to amplify the categorical difference between the negatives and the positives. For all positive and negative samples in a warped image I_i, the loss is formulated as:

$$\mathcal{L}_m(y^{ij}, \hat{y}^i) = \frac{1}{T}(\omega_1 \sum_{p=1}^{T} ||(u_p^{ij}, v_p^{ij}) - (\hat{u}_p^i, \hat{v}_p^i)||_2^2 + \omega_2 l_{focal}(c_p^{ij}, \hat{c}_p^i) + \omega_3 \sum_{n=T+1}^{C} l_{focal}(c_n^{ij}, \hat{c}_n^i)), \quad (6)$$

where ω_1, ω_2 and ω_3 are weight terms. $\{(u_p^{ij}, v_p^{ij}), c_p^{ij}\}$ denotes the p^{th} matching positive centroid of the j^{th} predicted set in the i^{th} warped image and $\{(\hat{u}_p^i, \hat{v}_p^i), \hat{c}_p^i\}$ is the corresponding target in the ground truth $\hat{y}$ after the i^{th} affine transformation. For the negative proposals, we define their classification target $\hat{c}_n^i$ as a new empty category. The loss between the local and the global predictions can be calculated as $\mathcal{L}_m(y^{ij}, \bar{y}^{ij})$ in Eq. 3, in a way similar to Eq. 6.

In the training stage, the encoder and each of the 3 decoder layers predict a set of centroids separately, resulting in 4 (1+3) sets of centroids. which are all used to compute the loss. During the inference, we take the output of the last decoder layer as the final prediction. The hyper-parameters $\alpha, \beta 1, \beta 2, \lambda 1, \lambda 2, \omega 1, \omega 2, \omega 3$ are set based on MMDetection [6] and Deformable DETR [57], and do not need any complicated tuning.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate the proposed approach on three publicly available datasets, CoNSeP [14], BRCA-M2C [1] and Lizard [13]. CoNSeP is a colorectal nuclear dataset, consisting of 41 H&E stained image tiles from 16 colorectal adenocarcinoma whole-slide images (WSIs). BRCA-M2C is a breast cancer dataset and consists of 120 image tiles from 113 patients. Both CoNSeP and BRCA-M2C contain three types of cells: inflammatory, epithelial, or stromal. Lizard has 291 histology images of colon tissue from six different dataset sources, containing nearly half a million labeled nuclei in H&E stained colon tissue. Lizard provides nucleus-level class labels for epithelial cells, connective tissue cells, lymphocytes, plasma cells and neutrophils. CoNSeP and Lizard contain the annotated contour masks of nuclei while BRCA-M2C only has the labels of centroids. To run instance based and bounding box based methods on

Table 1. Results on three benchmarks, 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-socre for the inflammatory, epithelial, stromal, neutrophils, lymphocytes, plasma, Eosinophil and connective tissue cells, respectively. For each row, the best method is in bold type and the second best method is underlined.

	F-score↑	Hovernet [14]	DDOD [7]	TOOD [10]	MCSpatNet [1]	SONNET [9]	DAB-DETR [25]	UperNet-ConvNeXt[26]	AC-Former (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	**0.635**
	$F_c^{Epi.}$	0.604	0.436	0.616	0.608	0.502	0.440	0.625	**0.635**
	$F_c^{Stro.}$	0.391	0.429	0.382	0.527	0.366	0.443	0.542	**0.568**
	$\overline{F_c}$	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	**0.739**
BRCA-M2C	$F_c^{Infl.}$	0.454	0.394	0.400	0.424	0.343	0.437	0.423	**0.474**
	$F_c^{Epi.}$	0.577	0.544	0.559	0.627	0.411	0.634	0.636	**0.637**
	$F_c^{Stro.}$	0.339	0.373	0.315	**0.387**	0.281	0.380	0.353	0.344
	$\overline{F_c}$	0.457	0.437	0.425	0.479	0.345	0.484	0.471	**0.485**
	F_d	0.74	0.659	0.662	0.794	0.653	0.705	0.785	**0.796**
Lizard	$F_c^{Neu.}$	0.210	0.025	0.029	0.105	0.09	0.142	0.205	**0.270**
	$F_c^{Epi.}$	0.665	0.584	0.615	0.601	0.599	0.653	0.714	**0.788**
	$F_c^{Lym.}$	0.472	0.342	0.404	0.457	0.538	0.544	0.611	**0.690**
	$F_c^{Pla.}$	0.376	0.130	0.152	0.228	0.370	0.356	0.333	**0.475**
	$F_c^{Eos.}$	0.367	0.124	0.157	0.220	0.365	0.295	0.403	**0.450**
	$F_c^{Con.}$	0.492	0.347	0.383	0.484	0.143	0.559	0.578	**0.671**
	$\overline{F_c}$	0.430	0.259	0.290	0.349	0.351	0.425	0.474	**0.557**
	F_d	0.729	0.561	0.606	0.713	0.682	0.656	0.764	**0.782**

BRCA-M2C, we follow the work [1] to apply the SLIC [2] algorithm to generate superpixels as instances. We split the fully labeled samples into training, validation, and test sets, following the official partition [14, 1, 13].

Evaluation metrics. We follow the work [14] and use F-score to evaluate the detection and classification tasks. A higher F-score means better performance. For the detection, we compute the Euclidean distance between each predicted centroid and GT to yield a cost matrix. Then we run the Hungarian algorithm [20] with the cost matrix to obtain the paired results, and set the pairs beyond 6 pixels to unpaired samples. The predicted centroids belonging to some pair are correctly detected centroids (TP_d, d for detection) while the rests are overdetected predicted centroids (FP_d). The GT centroids without matched predictions are called misdetected GT (FN_d). The detection F-score is calculated with the size of the above sets of nuclei: $F_d = \frac{2TP_d}{2TP_d+FP_d+FN_d}$.

For the classification task with K classes, TP_d are further split into the following subsets: correctly classified centroids of Type k (TP_c^k), incorrectly classified centroids of Type k (FP_c^k) and incorrectly classified centroids of types other than Type k (FN_c^k). The classification F-score is defined as: $F_c^k = \frac{2TP_c^k}{2(TP_c^k+FP_c^k+FN_c^k)+FP_d+FN_d}$.

Implementation details. Our implementation is based on MMDetection [6]. We use AdamW [28] optimizer with a learning rate of 2^{-4} to train models. We load the ImageNet [8] pre-trained weights for initializing the ConvNeXt-B backbone and the embedding dimension E is set to 128. For training and inference, we remove the centroid proposal whose maximum category score is smaller than a threshold of 0.5 and no more than n proposals are preserved. n is usually set to 1000. During the training, the network is trained with only the local branch loss in early steps and then the overall loss function (Eq. 1) is optimized with $\alpha = 0.1$. During the inference, we adopt the global network for prediction with sliding windows. We apply the multi-scale training with sizes between 600 and 800, and infer the image with a size of 800×800. More details are listed in the supplementary material.

4.2. Comparison with State-of-the-arts

As shown in Table 1, we compare our proposed method with the state-of-the-art approaches which can jointly segment/detect and classify cell nuclei. These approaches include the instanced based methods (Hovernet [14], SONNET [9]), the bounding box based methods (DAB-DETR [25], TOOD [10], DDOD [7]), and the dots map based methods (UperNet with ConvNeXt [26] backbone, MCSpatNet [1]). In Table 1, the proposed method achieves the highest mean F-score in both detection and classification tasks on the benchmarks CoNSeP, BRCA-M2C and Lizard.

For the Lizard dataset, our proposed method demon-

Table 2. The results of ablation study. The results are obtained on the CoNSeP and Lizard datasets. 'BL' means training the single-branch baseline in our method, with original pathological images. 'RC', 'RR' and 'RT' denote the random crop, random rotation and random affine transformation strategies for synthesizing local views. 'AAT' is the Adaptive Affine Transformer and 'GL' is the supervised loss based on the global branch.

Methods	CoNSeP					Lizard							
	$F_c^{Infl.}$	$F_c^{Epi.}$	$F_c^{Stro.}$	$\overline{F_c}$	F_d	$F_c^{Neu.}$	$F_c^{Epi.}$	$F_c^{Lym.}$	$F_c^{Pla.}$	$F_c^{Eos.}$	$F_c^{Con.}$	$\bar{F}_c$	F_d
BL	0.571	0.627	0.538	0.579	0.696	0.042	0.740	0.629	0.395	0.348	0.526	0.447	0.715
BL+RC	0.604	0.625	0.560	0.596	0.713	0.011	0.707	0.625	0.365	0.335	0.629	0.445	0.725
BL+RR	0.595	0.619	0.545	0.584	0.704	0.010	0.743	0.612	0.389	0.347	0.633	0.456	0.732
BL+RT	0.617	0.565	0.541	0.574	0.711	0.070	0.741	0.62	0.414	0.393	0.660	0.474	0.749
BL+AAT	0.603	0.627	0.553	0.594	0.721	0.234	0.774	0.659	0.440	0.428	0.615	0.525	0.752
BL+RC+GL	0.606	**0.648**	0.551	0.602	0.730	0.187	0.758	0.677	0.439	0.430	0.662	0.526	0.769
BL+RR+GL	0.626	0.638	0.555	0.606	0.726	0.140	0.758	0.643	0.411	0.404	0.657	0.502	0.747
BL+RT+GL	0.**642**	0.562	0.543	0.582	0.725	0.174	0.768	0.664	0.453	**0.454**	**0.679**	0.532	0.775
BL+AAT+GL (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**

strates 1.8% F-socre in detection and 8.3% F-score in classification higher than the second best model MCSpatNet, respectively. Some existing models show inferior results. It may be that Lizard is a newly released and challenging benchmark that has the class imbalance problem. Figure 4 presents a qualitative comparison between the state-of-the-art algorithms and the proposed network. More results are presented in the supplemental materials.

Comparison using the Same Backbone. Consider that a large-size of high-capacity backbone may improve or degrade the performance due to over-fitting. To fairly reproduce the existing methods in Table 1, their backbones are set following their original paper. All the backbones are pre-trained on the ImageNet. Note that in Table 1 even though DAB-DETR and UperNet use the same backbone ConvNeXt as our method, the proposed model significantly exceeds them by 2.4%-12% in F_d on the CoNSep dataset.

Comparison with Bounding Box based Methods. Bounding box methods provide more information like cell sizes, but their labels are more expensive than the centroid labels used by our method. We only aim at locating more cells with correct labels and reducing the missing rate, which can be used to compute the counts of cells as prognostic markers [12]. In Table 1, two competitive bounding-box models DDOD and DAB-DETR are compared with ours. The proposed method surpasses the two models by more than 9% F-score in detection on the BRCA-M2C dataset. Such a performance gap unveils those centroid-based methods are superior to bounding-box based ones for nuclei detection.

4.3. Ablation Study

Effectiveness of the proposed AAT module. In Table 2, 'BL' denotes the baseline that is the global branch in our method (Figure 2). 'BL+AAT' is a dual-branch model using the AAT to warp images and the EMA strategy to update the global branch. Comparing 'BL+AAT' with 'BL' shows that the AAT module improves the baseline by 7.8% in $\bar{F}_c$ and 3.7% in F_d on the Lizard dataset, which is significant.

Comparison with Non-learnable Data Augmentation. Since the AAT can learn to transform input image patches for training, we compare the AAT with other traditional non-learnable data augmentations. As Table 2 shows, we implement 3 kinds of non-learnable augmentation: Random Cropping, Random Rotation, and Random Affine Transformation, which are denoted as '+RC', '+RR', and '+RT'.

'BL+AAT+GL' denotes our overall proposed method, while 'GL' means the Global Loss (the supervision from the global branch, see the lower half of Table 2). By replacing AAT with RC/RR/RT, we obtain the results of 'BL+RC+GL', 'BL+RR+GL', 'BL+RT+GL'. Our proposed method outperforms the three models by 2.5%-5.5% F-score in classification and 0.7%-3.5% F-score in detection, on the Lizard dataset. The results suggest that the proposed AAT is superior to the 3 kinds of data augmentation strategies. Interestingly, the AAT brings more significant improvements in the sub-task of cell classification, which indicates that the proposed module does synthesize useful input samples for learning more robust semantic features.

Effectiveness of the Global-Local Architecture. We investigate the strengths of the global-local framework. In Table 2, we analyze the results of different image transformation strategies, and find that the use of Global Loss $\mathcal{L}(y, \bar{y})$ from the global network can achieve stable improvements. Specifically, on Lizard, 'BL+AAT+GL (Ours)' surpasses 'BL+AAT' by 3% and 3.2% F-scores in detection and classification, respectively. The statistics suggest that using sub-network learning from large-scale inputs can help train another sub-network to adapt to various fields of view.

Efficiency Analysis. Our AAT and dual-branch design are used only in the training stage. Thus, our method enjoys the same inference efficiency as the single-branch baseline. As Table 3 shows, our method avoids post-processing and

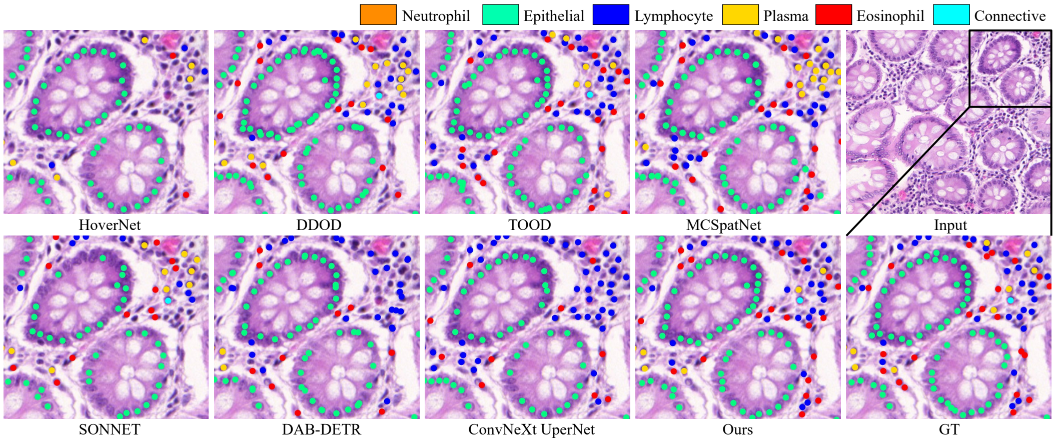

Figure 4. Qualitative comparison on the Lizard dataset. The four leftmost columns visualize the predictions of existing methods and ours. The rightmost column displays the input image and the ground truth annotations. Five types of cell are marked with dilated nucleus centroids in five different colors. As the results show, the category distribution of our method is the closest to that of the ground truths.

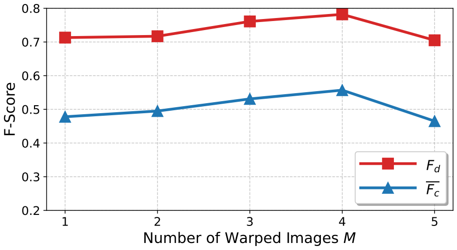

Figure 5. The effects of the number of warped images generated by the Adaptive Affine Transformer on Lizard dataset. The F-scores of each nucleus category are in the supplemental materials.

takes less than 0.1 s for inferring an image, while the post-processing based methods are 3 times slower (> 0.33 s).

Amount of Warped Images. Figure 5 shows the results of investigations about how the number of warped images M for training a local network affects the testing F-score. The results show that setting M to 4 performs the best. Setting M from 2 to 4 can achieve consistent improvements in comparison to the model with $M = 1$. A large number of distorted images (*e.g.*, $M = 5$) would degrade the results.

Table 3. Efficiency Analysis of the state-of-the-art methods and ours with a Tesla v100 (32GB), Intel Xeon Gold 6248 on CoNSeP.

	Time (s)	Memory (GB)	
Methods	Inference+Post-process	Inference	Training
Baseline (BL)	0.097	11.191	12.298
HoverNet	0.021+0.376	11.380	12.892
MCSpatNet	0.058+0.287	2.679	6.228
SONNET	0.081+0.250	21.031	30.493
Ours	0.097	11.191	24.206

5. Conclusion

In the paper, we propose a novel affine-consistent transformer framework that directly predicts a list of locations and categories for multi-class nuclei detection without complex post-refinements. We first introduce an Adaptive Affine Transformer module, which can automatically learn argumentation strategies to warp training input images, and enhance the model adaptability and accuracy. Next, we propose two associated networks that adapt to local-scale image views under the guidance of global-scale predictions, to boost the consistency and robustness of the model. Extensive experiments on three benchmarks have demonstrated the strengths of our overall framework AC-Former and the proposed AAT module on nuclei detection and classification tasks.

The limitation is that our model could fail to locate incomplete nuclei that are split at image boundaries due to the lack of contextual information. To solve the issue, we may try to crop highly-overlapping image patches and stitch their results better in future work.

Acknowledgements This work was supported in part by the Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), in part by the Guangdong Basic and Applied Basic Research Foundation (2023A1515011464, 2020B1515020048), in part by the National Natural Science Foundation of China (No. 62102267, No. 61976250), 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.

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] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11):2274–2282, 2012.

[3] Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2009.

[4] Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Learning to detect cells using non-overlapping extremal regions. In *MICCAI*, pages 348–356, 2012.

[5] 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.

[6] 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.

[7] 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.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[9] 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.

[10] 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.

[11] Zunlei Feng, Zhonghua Wang, Xinchao Wang, Yining Mao, Thomas Li, Jie Lei, Yuexuan Wang, and Mingli Song. Mutual-complementing framework for nuclei detection and segmentation in pathology image. In *ICCV*, pages 4036–4045, 2021.

[12] Wolf Herman Fridman, Franck Pagès, Catherine Sautès-Fridman, and Jérôme Galon. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, 12(4):298–306, 2012.

[13] 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.

[14] 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.

[15] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[16] Hongliang He, Zhongyi Huang, Yao Ding, Guoli Song, Lin Wang, Qian Ren, Pengxu Wei, Zhiqiang Gao, and Jie Chen. CDNet: Centripetal direction network for nuclear instance segmentation. In *ICCV*, pages 4026–4035, 2021.

[17] Henning Höfener, André Homeyer, Nick Weiss, Jesper Molin, Claes F Lundström, and Horst K Hahn. Deep learning nuclei detection: A simple approach can deliver state-of-the-art results. *Computerized Medical Imaging and Graphics*, 70:43–52, 2018.

[18] Junjia Huang, Haofeng Li, Weijun Sun, Xiang Wan, and Guanbin Li. Prompt-based grouping transformer for nucleus detection and classification. In *MICCAI*, 2023.

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *NeurIPS*, 28, 2015.

[20] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[21] Haofeng Li, Junjia Huang, Guanbin Li, Zhou Liu, Yihong Zhong, Yingying Chen, Yunfei Wang, and Xiang Wan. View-disentangled transformer for brain lesion detection. In *19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[22] Xia Li, Zhenhao Xu, Xi Shen, Yongxia Zhou, Binggang Xiao, and Tie-Qiang Li. Detection of cervical cancer cells in whole slide images using deformable and global context aware faster RCNN-FPN. *Current Oncology*, 28(5):3585–3601, 2021.

[23] Hao Liang, Zhiming Cheng, Haiqin Zhong, Aiping Qu, and Lingna Chen. A region-based convolutional network for nuclei detection and segmentation in microscopy images. *Biomedical Signal Processing and Control*, 71:103276, 2022.

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

[25] 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.

[26] 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.

[27] De Rong Loh, Wen Xin Yong, Jullian Yapeter, Karupppasamy Subburaj, and Rajesh Chandramohanadas. A deep learning approach to the screening of malaria infection: Automated and rapid cell counting, object detection and instance segmentation using mask r-cnn. *Computerized Medical Imaging and Graphics*, 88:101845, 2021.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[29] Wei Lou, Haofeng Li, Guanbin Li, Xiaoguang Han, and Xiang Wan. Which pixel to annotate: a label-efficient nuclei segmentation framework. *IEEE Transactions on Medical Imaging*, 42(4):947–958, 2022.

[30] Wei Lou, Xinyi Yu, Chenyu Liu, Xiang Wan, Guanbin Li, Siqi Liu, and Haofeng Li. Multi-stream cell segmentation with low-level cues for multi-modality images. In *Competitions in Neural Information Processing Systems*, pages 1–10. PMLR, 2023.

[31] Sidra Nawaz and Yinyin Yuan. Computational pathology: Exploring the spatial dimension of tumor ecology. *Cancer Letters*, 380(1):296–303, 2016.

[32] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *ICPR*, volume 3, pages 850–855, 2006.

[33] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The Lancet Oncology*, 20(5):e253–e261, 2019.

[34] May Phu Paing, Adna Sento, Toan Huy Bui, and Chuchart Pintavirooj. Instance segmentation of multiple myeloma cells using deep-wise data augmentation and mask r-cnn. *Entropy*, 24(1):134, 2022.

[35] Marina E Plissiti, Christophoros Nikou, and Antonia Charchanti. Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):233–241, 2010.

[36] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Dsod: Learning deeply supervised object detectors from scratch. In *ICCV*, pages 1919–1927, 2017.

[37] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016.

[38] Tatsuhiko Sugimoto, Hiroaki Ito, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Multi-class cell detection using modified self-attention. In *CVPRW*, pages 1855–1863, 2022.

[39] Yibao Sun, Xingru Huang, Huiyu Zhou, and Qianni Zhang. SRPN: similarity-based region proposal networks for nuclei and cells detection in histology images. *Medical Image Analysis*, 72:102142, 2021.

[40] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *ICCV*, pages 3611–3620, 2021.

[41] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *MICCAI*, pages 36–46, 2021.

[42] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature Medicine*, 27(5):775–784, 2021.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[44] Ching-Wei Wang, Sheng-Chuan Huang, Yu-Ching Lee, Yu-Jie Shen, Shwu-Ing Meng, and Jeff L Gaol. Deep learning for bone marrow cell detection and classification on whole-slide images. *Medical Image Analysis*, 75:102270, 2022.

[45] Zhi Wang, Xiaoya Zhu, Ao Li, Yuan Wang, Gang Meng, and Minghui Wang. Global and local attentional feature alignment for domain adaptive nuclei detection in histopathology images. *Artificial Intelligence in Medicine*, 132:102341, 2022.

[46] Liming Wu, Alain Chen, Paul Salama, Kenneth W Dunn, and Edward J Delp. 3d centroidnet: nuclei centroid detection with vector flow voting. In *ICIP*, pages 651–655. IEEE, 2022.

[47] Liming Wu, Shuo Han, Alain Chen, Paul Salama, Kenneth W Dunn, and Edward J Delp. Rcnn-slicenet: A slice and cluster approach for nuclei centroid detection in three-dimensional fluorescence microscopy images. In *CVPR*, pages 3755–3765, 2021.

[48] Yuanpu Xie, Fuyong Xing, Xiangfei Kong, Hai Su, and Lin Yang. Beyond classification: structured regression for robust cell detection using convolutional neural network. In *MICCAI*, pages 358–365, 2015.

[49] Hongming Xu, Cheng Lu, Richard Berendt, Naresh Jha, and Mrinal Mandal. Automatic nuclei detection based on generalized laplacian of gaussian filters. *IEEE Journal of Biomedical and Health Informatics*, 21(3):826–837, 2016.

[50] Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *IEEE Transactions on Medical Imaging*, 35(1):119–130, 2015.

[51] Safoora Yousefi and Yao Nie. Transfer learning from nucleus detection to classification in histopathology images. In *International Symposium on Biomedical Imaging*, pages 957–960, 2019.

[52] Xinyi Yu, Guanbin Li, Wei Lou, Siqi Liu, Xiang Wan, Yan Chen, and Haofeng Li. Diffusion-based data augmentation for nuclei image segmentation. In *MICCAI*, 2023.

[53] Zitao Zeng, Weihao Xie, Yunzhe Zhang, and Yao Lu. RIC-Unet: An improved neural network based on unet for nuclei segmentation in histology images. *IEEE Access*, 7:21420–21428, 2019.

[54] Donghao Zhang, Yang Song, Dongnan Liu, Haozhe Jia, Siqi Liu, Yong Xia, Heng Huang, and Weidong Cai. Panoptic segmentation with an end-to-end cell r-cnn for pathology image analysis. In *MICCAI*, pages 237–244, 2018.

[55] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *ICLR*, 2023.

[56] Hong-Yu Zhou, Chengdi Wang, Haofeng Li, Gang Wang, Shu Zhang, Weimin Li, and Yizhou Yu. Ssmd: Semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation. *Medical Image Analysis*, 72:102117, 2021.

[57] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.