Research Article

Keep it accurate and robust: An enhanced nuclei analysis framework

Wenhua Zhang [a], Sen Yang [b], Meiwei Luo [c], Chuan He [d], Yuchen Li [b,*], Jun Zhang [c,*], Xiyue Wang [b,*], Fang Wang [e,*]

[a] *Institute of Artificial Intelligence, Shanghai University, Shanghai 200444, China*
[b] *Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA 94305 USA*
[c] *Tencent AI Lab, Shenzhen 518057, China*
[d] *Shanghai Aitrox Technology Corporation Limited, Shanghai, 200444, China*
[e] *Department of Pathology, The Affiliated Yantai Yuhuangding Hospital of Qingdao University, Yantai, 264000, China*

ARTICLE INFO

ABSTRACT

Accurate segmentation and classification of nuclei in histology images is critical but challenging due to nuclei heterogeneity, staining variations, and tissue complexity. Existing methods often struggle with limited dataset variability, with patches extracted from similar whole slide images (WSI), making models prone to falling into local optima. Here we propose a new framework to address this limitation and enable robust nuclear analysis. Our method leverages dual-level ensemble modeling to overcome issues stemming from limited dataset variation. Intra-ensembling applies diverse transformations to individual samples, while inter-ensembling combines networks of different scales. We also introduce enhancements to the HoVer-Net architecture, including updated encoders, nested dense decoding and model regularization strategy. We achieve state-of-the-art results on public benchmarks, including 1st place for nuclear composition prediction and 3rd place for segmentation/classification in the 2022 Colon Nuclei Identification and Counting (CoNIC) Challenge. This success validates our approach for accurate histological nuclei analysis. Extensive experiments and ablation studies provide insights into optimal network design choices and training techniques. In conclusion, this work proposes an improved framework advancing the state-of-the-art in nuclei analysis. We will release our code and models to serve as a toolkit for the community.

1. Introduction

Advancing the analysis of pathology images remains crucial for improving cancer diagnosis and prognosis [66]. Pathologists heavily rely on information extracted from pathology images to determine tumor grades [18], predict patient survival rates [2], and anticipate responses to various therapies [55]. Among the many analytical steps involved, nuclear segmentation and classification, as well as cellular composition prediction, are particularly fundamental, as they enable the extraction of cell-based features useful for numerous downstream tasks [2,46,54,55]. While manual analysis is vulnerable to intra- and inter-observer variability [15], automatic methods can enable consistent and rapid analysis of these tasks.

Researchers have leveraged image processing techniques since the 1990s to analyze morphological cell features [8,26,12]. The recent advent of deep learning has further accelerated the field, with neu-

ral networks becoming a popular choice for solving biomedical challenges [36,65,44,50,51,9]. This has led to the development of several deep learning approaches for nuclear segmentation and classification [25,13,53,49], which have demonstrated substantial improvements over traditional techniques.

Instance segmentation in digital pathology is challenging due to the high degree of overlap between adjacent nuclei [34]. Recently, various methods have been proposed to tackle this challenge. These methods can be broadly categorized into two types: segmentation-only approaches and simultaneous segmentation and classification approaches. Segmentation-only approaches focus on segmenting individual instances without considering their classification information [1,27]. On the other hand, simultaneous segmentation and classification approaches aim to segment and classify instances simultaneously. For example, methods based on Mask R-CNN [3,48] propose regions of interest before segmenting and classifying objects within. However, inaccurate region proposals

can degrade segmentation and classification. An alternative approach is to use U-Net architectures [25,14,49], which directly predict pixel-level maps and apply post-processing for instance segmentation and classification. These methods are particularly well-suited for highly overlapping scenarios.

Cellular composition prediction can provide valuable insights into tumor types and cell populations [30,64]. As manual counting is tedious and subjective, automated approaches typically employ either regression to directly predict cell counts [37,62,13] or object detection to first identify cells and then count them [61,11]. While regression methods are simple and effective, detection provides both segmentation and counting within a single framework.

In this work, we present a novel deep-learning framework for joint nuclear segmentation, classification, and cellular composition prediction in digital pathology images. The framework is designed to improve robustness through ensemble modeling, fusing predictions from diverse improved base models.

Our framework consists of specialized base models, each trained on different data folds with varying encoder backbones. To enhance computational efficiency, we replace the heavy decoder branches of prior arts [25]. The predictions from each base model are then fused via averaging and post-processing to harness diversity. For cellular composition, detected nuclei are counted by class, establishing interdependency between nuclear identification and cellular composition analysis.

To evaluate the performance of our framework, we participated in the 2022 CONIC Challenge and achieved top rankings in cellular composition prediction and 3rd in nuclear segmentation/classification. Additional experiments on external datasets PanNuke [20,21] and MoNuSAC [56] also demonstrate superior performance over the state-of-the-art.

We list our key contributions as follows:

- Proposing a dual-ensemble system leveraging intra-model and inter-model ensembling to avoid overfitting by a single model that may fall into local optima.
- Developing an enhanced architecture for nuclear instance segmentation and classification by upgrading HoVer-Net with advanced encoders, transformed decoders, and regularization techniques.
- Demonstrating state-of-the-art performance of the enhanced framework on publically benchmarks for nuclear instance segmentation, classification, and composition prediction.

2. Related work

2.1. Nuclear segmentation and classification

Nuclear segmentation and classification have gained significant attention in recent years within computational pathology research [34,25, 63,3,48,1,27,49]. The field can be broadly categorized into two types: segmentation-only approaches and simultaneous segmentation and classification approaches.

Segmentation-only approaches focus on segmenting individual instances without considering their classification information. For instance, Abdel-Nasser et al. proposed a staining-invariant encoder and a weighted hybrid dilated convolution block to efficiently segment nuclear instances without color normalization [1]. Similarly, Hassan et al. proposed a learnable aggregation network to ensemble a set of individual nuclear segmentation models [27].

Simultaneous segmentation and classification approaches aim to segment and classify instances simultaneously. For example, methods based on Mark R-CNN propose candidate regions and then attempt to segment nuclear instances within each region proposal before classifying the detected nuclei [3,48]. However, the prevalence of overlapping nuclei introduces significant noise during region proposal, which cascades into errors in downstream segmentation and classification. To address this, some researchers have explored alternative frameworks based on U-Net architectures. A two-stage approach is often taken, where nu-

clei coordinates are first detected and then patches are cropped around these coordinates to feed into a separate classification model [53]. Other methods opt for a single model that elegantly handles both tasks of segmentation and classification simultaneously [25,63,49]. Notably, the advent of HoVer-Net [25] has a significant contribution in this area, which inspired several innovative variants. One such variant integrates tissue segmentation [58] to provide a contextual framework that aids in the precise identification of nuclei. Another variant introduces a multiple filter unit [57], designed to capture a wider range of features. In parallel, the CellViT [31] represents a shift towards utilizing vision transformers as encoders, taking advantage of the rich feature extraction capabilities of large-scale pretrained models like the Segment Anything Model [40]. Motivated by these advancements, our method echoes the unified model approach of [25], but with specific enhancements tailored for our research objectives.

Inter-image variability inherent to histology data is another major difficulty in nuclear recognition. H&E stained specimens sourced from different tissue types, fixation methods, or even hospitals can exhibit dramatic visual differences. Many publicly available datasets such as PanNuke [20,21] and CoNSeP [25] comprise H&E patches with relatively homogeneous styles since they draw data from only one or two sites. Models trained exclusively on such narrow datasets tend to overfit, rendering their ability to generalize to more diverse input lacking. Therefore, we opt to train our model on the CoNIC data [24], the largest pathology dataset to date spanning diverse sources from 5 different institutions. By learning from such heterogeneous data, we aim to develop a segmentation and classification model with broader applicability across visual domains.

### 2.2. Cellular composition prediction

Cellular composition prediction in histology slides plays a crucial role in computational pathology, enabling analyses for various cancer types. For instance, accurately quantifying mitotic figures provides a critical parameter for tumor grade in breast cancer [16,34]. Additionally, the percentage of tumor nuclei within the tumorbed area indicates the effect of neoadjuvant therapy in various cancers [4]. Computational composition analysis also sheds light on the tumor microenvironment (TME) [19], revealing the balance of malignant, immune, stromal, and healthy cell populations coexisting within the tumor ecology.

The manual process of counting cells slide-by-slide is laborious and prone to subjective errors and variability across pathologists. The advent of digital pathology has enabled the development of automated computational approaches for cellular composition estimation aimed at supplementing pathologists' analyses.

We can categorize existing research into two approaches: direct regression methods that output overall cell type counts without explicitly localizing individual nuclear instances [37,62,13], and methods that first detect and segment all nuclei before tallying counts per category to derive composition [61,11]. Since our goal is to develop an integrated model capable of both nuclear recognition and downstream composition prediction, we opt for the detection-based approach of first identifying nuclei via segmentation and classification, after which we can readily infer overall composition by counting nuclei in each predicted category.

2.3. Ensemble modeling

Model ensembling has emerged as an effective technique for improving the performance of deep learning systems. The approach involves combining multiple models to create an ensemble system that leverages the complementary strengths of the individual models [42]. A key advantage of ensembling is reducing the generalization error and variability of predictions [42]. The base models comprising the ensemble are often trained separately, using distinct architectures, algorithms, or training data splits [41]. Though containing multiple components,
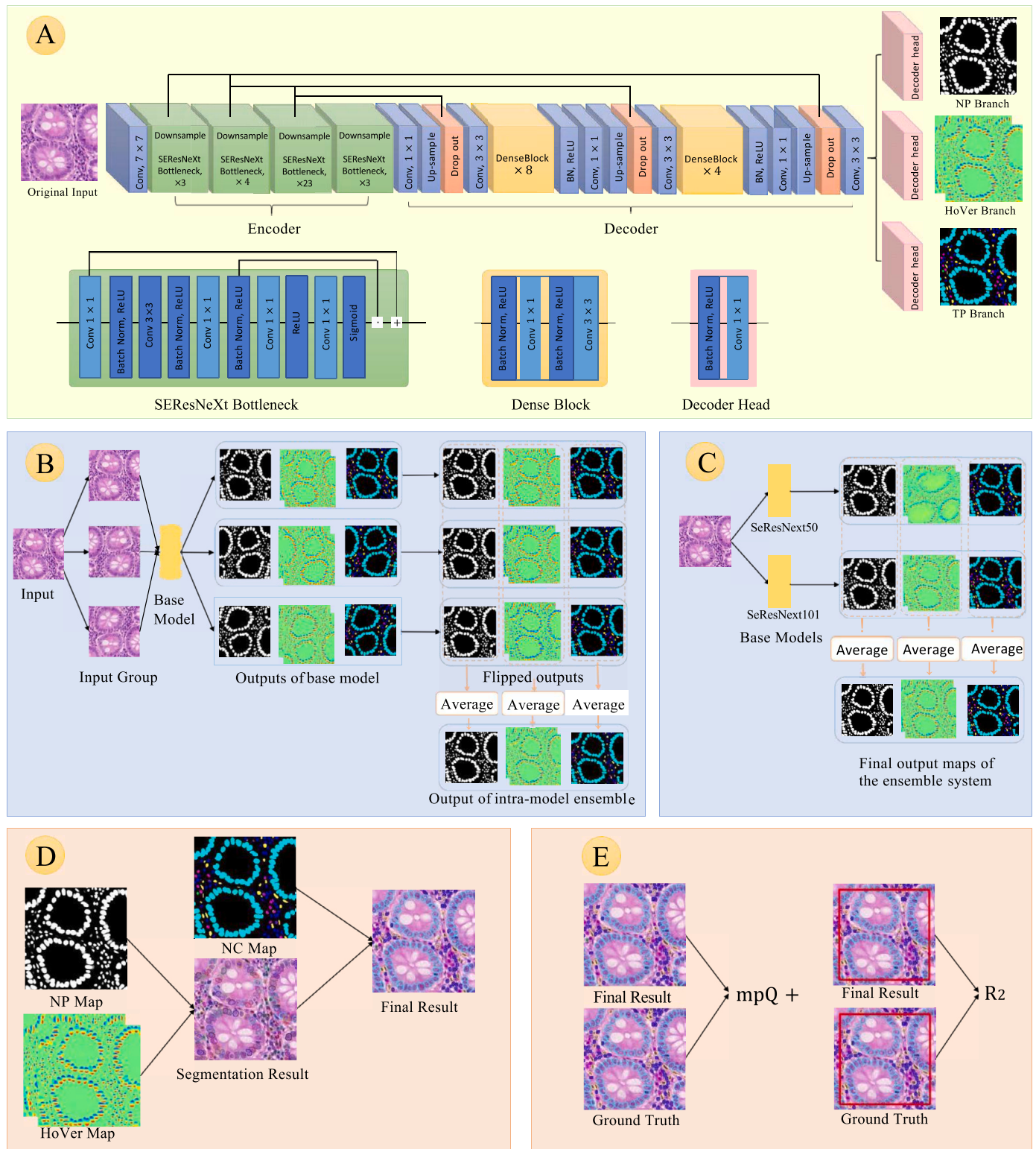
Fig. 1. Overview of the proposed method. A) The base model architecture utilizes an encoder-decoder structure adapted from HoVer-Net [25]. To improve compactness and effectiveness, the three decoder branches have been consolidated into one. Heavy dropout layers are also incorporated in the decoder to regularize training. While any encoder with a similar structure can be used, this example implements a SEResNeXt50 backbone [33]. B) The intra-model ensemble approach augments the input image with horizontal and vertical flips. Each base model makes predictions on the original and flipped inputs, which are averaged together after flipping the outputs back to the original orientation. This allows the network to leverage multiple views of the input during inference. C) The inter-model ensemble averages the output maps from base models with different encoder backbones to improve robustness. D) Post-processing utilizes the output maps for final instance recognition. Instances are first segmented using the nuclear presence (NP) and HoVer maps. These segmentation results are then grouped with the nucleus classification (NC) map for instance classification. E) Evaluation uses the multi-class panoptic quality ($mPQ+$) metric applied to the full input patch and multi-class coefficient of determination (R^2) applied to the 224×224 pixel center region.

the ensemble model essentially functions as a single unified model with lower aggregate error [42].

Model ensembling has been successfully applied in many medical imaging applications. For example, [10] averaged the predictions of several networks to reduce variance in segmenting neuronal structures, winning the 2012 ISBI EM Segmentation Challenge. [43] also utilized an ensemble of fully convolutional networks to detect white matter hyperintensities, achieving state-of-the-art performance in the 2017 MICCAI WMH Segmentation Challenge. Such results demonstrate the potential of ensemble techniques to boost performance.

Common ensembling approaches include max voting, which selects the majority vote [47,5], and averaging, which computes the mean predicted probabilities [5]. We implement an averaging ensemble, as it is widely adopted. Through designing ensemble algorithms tailored for our application, we aim to improve overall performance by synergistically combining multiple specialized base models.

3. Methods

Our proposed framework incorporates three vital components: 1) the base models for nucleus detection, 2) an ensemble system to combine base model outputs, and 3) post-processing techniques to refine ensemble predictions. At last, we describe the evaluation metrics and experimental setup.

3.1. Base models

Our approach is built upon diverse base models, each employing a different encoder architecture as the backbone. Inspired by the HoVer-Net framework [25], which has achieved state-of-the-art performance on joint nuclear segmentation and classification tasks, our base models share a common architecture, as illustrated in Fig. 1. Each base model outputs three maps: 1) a nuclear pixel (NP) map predicting nuclei segmentation, 2) a horizontal/vertical distance (HoVer) map encoding the coordinates of nuclear centers, and 3) a nucleus classification (NC) map categorizing each nucleus.

The NP map indicates whether each pixel belongs to a nuclear foreground region or the background. The HoVer map provides normalized horizontal and vertical distances to the centers of mass for clustered nuclei. By combining these two outputs, the model can separate touching and overlapping instances using a watershed transform during post-processing. Finally, the NC map classifies each segmented nucleus by performing pixel-wise voting within the predicted region. The category receiving the most votes defines the final call for that nucleus.

In our design, all three output maps share parameters within a common decoder pathway. By training these complementary maps jointly, we provide regularization for the decoder to learn robust features. We employ a single convolutional layer as the head for each map, minimizing added parameters. Our experiments validate that this simplified yet unified architecture achieves strong performance on both segmentation and classification. Moreover, by reducing computations, it allows us to use large batch sizes that accelerate training.

We utilize two state-of-the-art models: SEResNeXt50 and SEResNeXt101 [32], as encoder backbones. We train a specialized base model for each encoder variant. Next, we detail our ensemble approach to synergistically combine the strengths of these diverse base models.

3.1.1. Model regularization

To further enhance model generalization, we employ regularization strategies. As illustrated in Fig. 1A, we incorporate dropout layers periodically within the decoder after each upsampling layer. Dropout randomly omits or 'drops out' a subset of units during training, preventing the model from relying too heavily on particular features or cues that may vary across domains. This encourages the model to synthesize more holistic representations that do not overfit to nuances of the training distribution. Futhermore, our experiments demonstrate that the inclusion of dropout improves model performance on unseen test data.

3.1.2. Loss function

Following HoVer-Net [25], we employ a composite multi-task loss to optimize the model:

$$L = w_{NP}L_{NP} + w_{HoVer}L_{HoVer} + w_{NC}L_{NC}, \tag{1}$$

where w_{NP}, w_{HoVer}, and w_{NC} are the weights for the different loss sets, and L_{NP}, L_{HoVer}, and L_{NC} represent the loss terms for the nuclear pixel (NP) segmentation map, horizontal/vertical (HoVer) distance map, and nucleus classification (NC) map outputs respectively.

For the nuclear pixel (NP) branch, we apply a weighted combination of binary cross-entropy and dice losses. The binary cross-entropy loss provides pixel-level supervision for accurate nuclei vs. background classification. Meanwhile, the dice loss helps counter class imbalance between the typically smaller nuclear regions and the larger background area. Similarly, the nuclear classification (NC) branch is trained via analogous multi-task versions of cross-entropy and dice losses. Finally, for the horizontal/vertical distance (HoVer) branch, we utilize mean squared error and mean squared gradient error losses. These losses provide tailored supervision to precisely predict both the distance maps to nuclear centers themselves, as well as the distance gradient maps pointing to the centers. We refer the readers to the HoVer-Net [25] paper for more details of the loss function design.

3.2. Model ensemble system

To harness the diversity of our trained base models, we propose a two-level model ensemble system that aggregates their predictions. The system consists of two levels: intra-model ensembling and inter-model ensembling. Intra-model ensembing combines predictions from differently augmented inputs with the same based model, while inter-model ensembling combines predictions from different base models. This approach is illustrated in Fig. 1B and Fig. 1C.

3.2.1. Intra-model ensembling

Intra-model ensembling, as illustated in Fig. 1B, combines the predictions from multiple augmented versions of the same base model. We generate augmented inputs by horizontally and vertically flipping the original input patch and feed these flipped variants through the base model to produce multiple output maps for the same image. We flip the nuclear pixel (NP) and nuclear classification (NC) maps back to their original orientation and average them to smooth predictions. For the horizontal/vertical (HoVer) distance maps, we additionally inverse the sign of values based on flip direction before averaging. This intra-model ensembling reduces errors and noise by aggregating consistent predictions.

3.2.2. Inter-model ensembling

Inter-model ensembling, as illustated in Fig. 1C, combines the predictions from multiple base models with distinct encoder backbones. Each base model undergoes intra-model ensembling to produce averaged output maps, which are then averaged across base models to generate the final prediction. By combining the complementary strengths of each base model, inter-model ensembling creates a more robust ensemble model that leverages the diversity of the individual models.

3.3. Post-processing

We employ post-processing techniques to get the final nuclear instance segmentations and classifications. The process involves two primary steps: watershed segmentation and pixel-wise voting. The HoVer and NP maps serve as inputs to the watershed algorithm, where the HoVer map defines an energy landscape with nuclear centers as catchment basins, and the NP map removes background areas by zeroing their HoVer distances. Following segmentation, each nucleus is classified via pixel-wise voting within the nuclear instance region, assigning the most frequent class. By leveraging all three output maps in concert, the model

Table 1
Detailed information of the datasets. We provide a detailed overview of the three public datasets, including the number of patches, nuclei, and classes, to illustrate the composition of the datasets.

	No. Patches	No. Nuclei	Classes
MoNuSAC [56]	310	≈46,000	epithelial, lymphocyte, neutrophil, macrophage
PanNuke [20,21]	7901	≈200,000	neoplastic, inflammatory, connective, dead, epithelial
Lizard [23]	4981	≈500,000	epithelial, lymphocyte, plasma, eosinophil, neutrophil, connective tissue

is able to perform end-to-end detection, segmentation, and classification.

The workflow is illustrated in Fig. 1D, where the HoVer and the NP maps are utilized for instance segmentation, and the NC map is used for classification. The resulting segmented and classified instances are then used to count the number of nuclei of each category, enabling the prediction of nuclear composition within the central 224×224 region for image patches as shown in Fig. 1E.

3.4. Datasets

The quality and diversity of training data are crucial for developing robust computational pathology models. Histopathology images exhibit high variability due to factors such as tissue type, fixation method, and source institution. To mitigate dataset bias and train more generalizable models, we leverage multi-site datasets with heterogeneous visual styles for pre-training. We opt to train on the large-scale Lizard repository, which contains annotated pathology images aggregated from 6 unique domains, with iterative pathologist input to refine labels. We also validate our method on two other datasets: PanNuke and MoNuSAC. Detailed information on these datasets is illustrated in Table 1.

3.4.1. MoNuSAC

The MoNuSAC dataset [56] contains over 46,000 nuclei across 37 hospitals, 71 patients, 4 organs, and 4 nuclear classes - epithelial, lymphocyte, neutrophil, and macrophage. Sourced from TCGA, the whole slide images are sampled at 40x magnification. Annotations were performed by engineering graduate students and quality was checked by domain experts. Considering that the MoNuSAC dataset [56] does not provide additional data splits in its original publication, we chose not to perform cross-validation on it. Instead, we present the mean and standard deviation of 10 experimental runs, ensuring a consistent approach that facilitates direct comparison of performance metrics across all datasets.

3.4.2. PanNuke

The PanNuke [20,21] dataset contains multi-organ data across 19 tissue types with nearly 200,000 annotated nuclei in 5 clinically relevant classes - neoplastic, inflammatory, connective, dead, and epithelial. It consists of 481 visual fields sampled at 40x magnification where nuclei are semi-automatically segmented and verified by experts. To facilitate robust model evaluation, the dataset is pre-split into three folds. We adhere to the original split of the PanNuke dataset to perform cross-validation, as detailed in Table 10. Note that all other results for PanNuke presented in this paper are based on Fold 0, with mean and standard deviation calculated from 10 independent experimental runs, maintaining uniformity in our dataset evaluations.

3.4.3. Lizard

The Lizard dataset [23] contains colon tissue samples sourced from 5 distinct repositories - Glas [52], CRAG [22], CoNSep [25], DigestPath, PanNuke [20,21]. It is used by the CONIC 2022 challenge, with the composition detailed in an excel file provided by the challenge orga-

(a) Neutrophil (b) Epithelial (c) Lymphocyte

(d) Plasma (e) Eosinophil (f) Connective
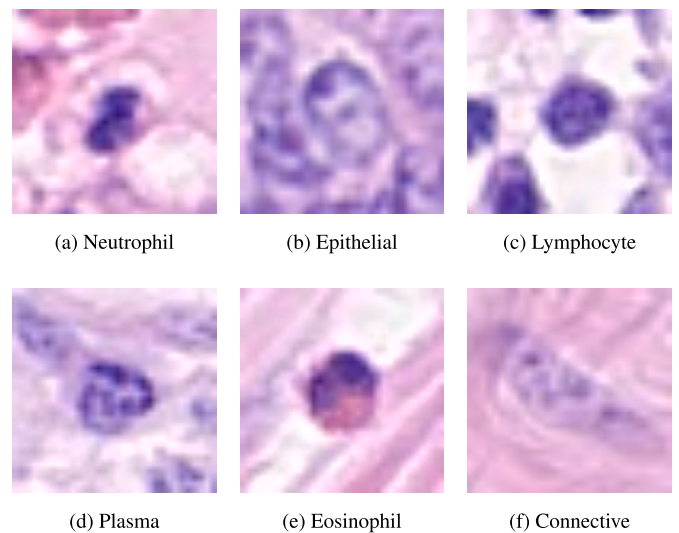
Fig. 2. Example nuclei from the 6 categories in the Lizard dataset, illustrating the challenges in segmentation and classification. Note the high degree of similarity among certain classes (e.g., Fig. 2(b) Epithelial, Fig. 2(c) Lymphocyte, and Fig. 2(d) Plasma), which hinders classification.

nizers [24,67].[1] To maintain the integrity of the original sources and to align with the cross-validation approach, we split the Lizard dataset into five folds corresponding to these sources to perform cross-validation, as shown in Table 10. Note that all other results for Lizard presented in this paper are performed on Fold 0, with mean and standard variations calculated from 10 times of experiments, as for MoNuSAC and PanNuke.

The dataset is cropped into 4981 256×256 patches containing around 500,000 annotated nuclei. Fig. 2 illustrates example patches from each nucleus category to provide intuition. While images of colon tissues from PanNuke are included in Lizard, we evaluated these two datasets independently without any overlap between training and test sets to avoid information leakage. Regions of interest are extracted from the whole slide images at 20x magnification and nuclei are annotated into 6 categories - epithelial, lymphocyte, plasma, eosinophil, neutrophil, and connective tissue. Example patches from Lizard are shown in Fig. 3.

4. Experiments

We conduct extensive experiments to evaluate the performance of our proposed framework. This section describes the evaluation metrics and experimental setup.

4.1. Evaluation metrics

We evaluate nuclei detection, segmentation, and composition estimation using two key metrics: $mPQ+$ for nuclear instance recognition and classification and R^2 for nuclear composition regression.

The $mPQ+$ metric evaluates combined nuclear segmentation and classification performance, adapted from the Panoptic Quality (PQ) metric [39] introduced in [25] for computational pathology tasks. Specifically, for each nuclear type t, PQ_t is defined as:

$$PQ_t = \frac{|TP_t|}{|TP_t| + \frac{1}{2}|FP_t| + \frac{1}{2}|FN_t|} \times \frac{\Sigma_{(x_t,y_t)\in TP_t} IoU(x_t, y_t)}{|TP_t|} \tag{2}$$

Here, x_t refers to a ground truth nucleus of type t, y_t refers to a predicted nucleus, and $IoU(x_t, y_t) > 0.5$ for matched nuclei pairs $(x_t, y_t) \in$

---

[1] https://conic-challenge.grand-challenge.org/.

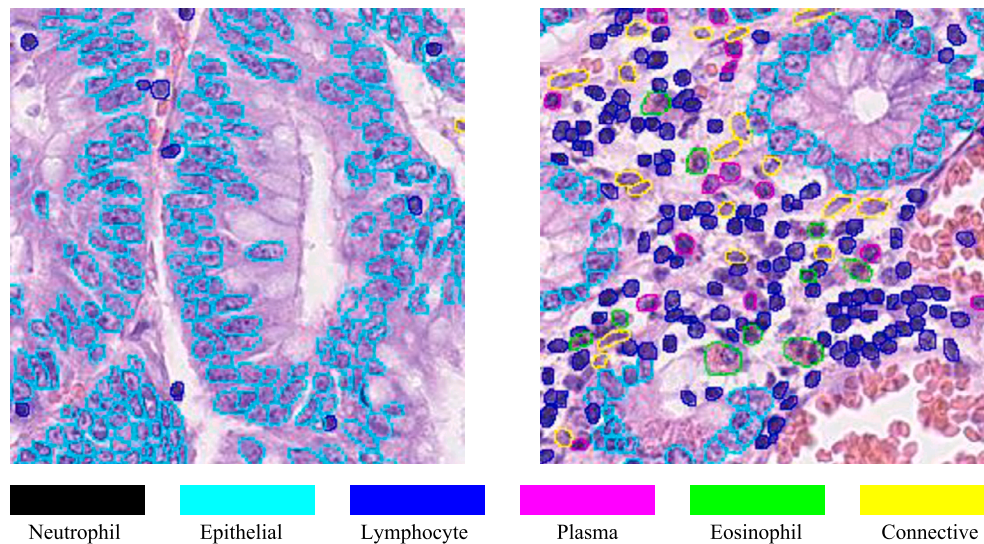| Neutrophil | Epithelial | Lymphocyte | Plasma | Eosinophil | Connective |

Fig. 3. Example images from the Lizard dataset. Note the high degree of nuclei overlapping, which poses a significant challenge for instance segmentation methods.

Table 2

Performance comparison of the proposed framework against other methods on the nuclear recognition and composition prediction tasks over multiple datasets. The proposed framework consistently outperforms the other methods by a substantial margin in most cases. Even in the rare instances where our approach does not surpass all others, it remains comparable to the best models. This demonstrates the effectiveness of our method for nuclear recognition and quantification. Further analysis of the results is provided in Section 5.1.

	$mPQ+$			R^2		
	Lizard	PanNuke	MoNuSAC	Lizard	PanNuke	MoNuSAC
Mask R-CNN [28]	0.4281±0.0110	0.4024±0.0246	0.5072±0.0069	0.4459±0.0233	0.6771±0.1016	0.6105±0.0187
Cascaded R-CNN [6]	0.4493±0.0077	0.4420±0.0125	0.5278±0.0027	0.4907±0.0214	**0.7640±0.0201**	0.6484±0.0272
QueryInst [17]	0.3549±0.0017	0.4405±0.0031	0.4938±0.0149	0.4552±0.0049	0.7134±0.0147	0.5813±0.2020
StarDist [60]	0.4194±0.0128	0.3855±0.0108	0.4097±0.0068	0.4795±0.0470	0.5662±0.0236	0.6774±0.1475
HoVer-Net [25]	0.4893±0.0217	0.3736±0.0111	0.4501±0.0154	0.7081±0.1148	0.5913±0.0237	0.5933±0.1419
Ours	**0.5599±0.0086**	**0.4738±0.0061**	**0.5567±0.0125**	**0.8437±0.0454**	0.7127±0.0553	**0.7968±0.0470**

TP_t. Unmatched ground truth and predicted nuclei are counted as false negatives (FN_t) and false positives (FP_t) respectively.

The $mPQ+$ metric evaluates segmentation and classification performance by averaging the PQ_t score across all nuclear types t. Critically, PQ_t is calculated over the full dataset rather than individual images, addressing potential issues that may arise from zero denominators when classes are absent from some images. For full details on the $mPQ+$ formulation and its advantages over standard mPQ, we direct readers to the work presenting this evaluation framework in the context of the CoNIC Challenge [24].

For quantitative composition estimation, we report the R^2 coefficient of determination. For each nuclear type t, R_t^2 is:

$$R_t^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,t} - \hat{y}_{i,t})}{\sum_{i=1}^{n}(y_{i,t} - \bar{y}_t)^2} \quad (3)$$

Here, $y_{i,t}$ is the ground truth count for nuclei of type t in image patch i, $\hat{y}_{i,t}$ is the predicted count, and $\bar{y}_t$ is the mean $y_{i,t}$ across patches. The overall multi-class R^2 score averages the per-type R_t^2 metrics.

As illustrated in Fig. 1E, $mPQ+$ is computed over the full image patch (256×256), while R^2 uses only the central 224×224 region to avoid edge effects.

4.2. Implementation details

In this work, we employ two powerful encoder backbones, SEResNeXt50 and SEResNeXt101 [32], to extract robust features in our multi-task segmentation framework. The SEResNeXt models are built upon the ResNeXt architecture and incorporate Squeeze-and-Excitation blocks to improve feature representation. During inference, we ensemble the

Table 3

Comparison of the top 10 teams on the CoNIC 2022 challenge leaderboard for the two evaluation metrics. Our proposed approach achieved 3rd place for multi-class panoptic quality ($mPQ+$) and 1st place for composition regression (R^2). The blind test set indicates the superiority of our method on unseen custom data.

Team	$mPQ+$	Team	R^2
EPFL |StarDist	0.50132	**Pathology AI**	**0.76413**
MDC Berlin |IFP Bern	0.47616	AI_medical	0.76250
Pathology AI	**0.46310**	EPFL |StarDist	0.75498
LSL000UD	0.46278	CIA Group	0.71902
AI_medical	0.45759	Softsensor_Group	0.71589
Arontier	0.45707	LSL000UD	0.70325
CIA Group	0.45092	Arontier	0.69145
MAIIA	0.43674	MBZUAI_CoNiC	0.67440
ciscNet	0.42947	MDC Berlin |IFP Bern	0.66432
MBZUAI_CoNiC	0.42078	Denominator	0.65498

predictions of both models to obtain the final segmentation result. By combining diverse models, we can leverage their strengths and improve robustness and accuracy.

Our training is performed on a high-performance workstation with 4 NVIDIA 3090 GPUs using distributed parallel training. This allows us to train with a large batch size of 16 for faster convergence. Models are trained for 50 epochs each, using the Adam optimizer [38] with an initial learning rate of 3e-4. The learning rate is decayed by a factor of 0.1 every 10 epochs after the first 30 epochs to improve convergence at the end of training. To prevent overfitting, we employ an early stopping strategy based on the validation loss in the challenge. Training stops once the loss stabilizes and stops improving, and we use the model from several epochs prior.

Table 4

Performance comparison between our proposed framework and the original HoVer-Net [25] on the Lizard dataset [23]. Our approach consistently and significantly outperforms HoVer-Net on these categories on both evaluation metrics, demonstrating the advantages of our method.

	$mPQ+$						R^2					
	Neutrophil	Epithelial	Lymphocyte	Plasma	Eosinophil	Connective	Neutrophil	Epithelial	Lymphocyte	Plasma	Eosinophil	Connective
HoVer-Net	0.2987±0.0399	0.6216±0.0213	0.7108±0.0196	0.4611±0.0177	0.3085±0.0524	0.5353±0.0084	0.8980±0.0436	0.9470±0.0244	0.5768±0.2246	0.7331±0.1057	0.6553±0.1481	0.4387±0.3953
Ours	0.3941±0.0227	0.6684±0.0028	0.7503±0.0047	0.5180±0.0114	0.3980±0.0218	0.6305±0.0084	0.9586±0.0126	0.9847±0.0048	0.7531±0.0495	0.7147±0.2430	0.7881±0.0450	0.8628±0.0393

Table 5

Performance comparison between our method and CellViT. The results indicate that our method, which does not leverage large-scale pretrained models, demonstrates performance that is highly comparable to CellViT.

	$CellViT_{256}$	CellViT-SAM-H	Ours
mPQ	0.4846±0.0503	0.4980±0.0413	0.5060±0.0022
bPQ	0.6696±0.0340	0.6793±0.0318	0.6353±0.0065

To enhance robustness and prevent overfitting, we apply extensive data augmentation techniques to our training set. These include random horizontal/vertical flips, random 90-degree rotations, random transposing, color jittering of brightness, contrast, saturation and hue, Gaussian blurring, median blurring, and motion blurring of the images. We only apply one blurring operation each time to avoid corrupting the image too much.

We use a weighted combination of loss components from each branch. Through a grid search, we found that w_{NP}, w_{HoVer}, and $w_{NC}=1$ work best. The NP branch uses $w_{CE}=2$ and $w_{Dice}=2$ for its cross entropy and Dice loss components. The hover branch uses $w_{mCE}=3$ and $w_{mDice}=1$ to weight its modified cross entropy and Dice losses. The keypoint loss uses $w_{mse}=2$ and $w_{msge}=2$ for the MSE and MSGE losses. Carefully balancing the loss components prevents one task from dominating the others during training.

5. Results

In this section, we present experimental results comparing our proposed multi-task segmentation framework against current state-of-the-art methods on two key tasks: 1) Nuclear instance segmentation and classification, and 2) Nuclear composition regression, as defined in the CoNIC 2022 Challenge [24]. We also use Dice to calculate the model's performance on segmentation. Ablation studies are conducted to investigate the contribution of each component of our framework.

5.1. Comparison with other state-of-the-art methods

We compare our approach to several state-of-the-art methods on nuclear instance segmentation and classification, including Mask R-CNN [28], Cascaded R-CNN [6], and QueryInst [17]. We also compare against HoVer-Net [25] and star-dist [49,60,59], two specifically designed methods for nuclear instance segmentation and classification. We show the quantitative results in Table 2, reporting standard evaluation metrics such as segmentation and classification ($mPQ+$), as well as nuclear composition regression (R^2 score). Our framework demonstrates superior performance across the majority of datasets and evaluation metrics, showcasing its robustness and versatility. In instances where our model does not lead in a particular metric, it remains highly competitive, exhibiting performance on par with the state-of-the-art methods. A case in point is the R^2 score for the PanNuke dataset, where our model, despite not achieving the foremost position, closely matches the performance of the leading cascaded R-CNN approach.

We have also conducted a comparative analysis of our method with the composite work CellViT [31] (see Table 5), utilizing the metrics employed in its original publication: mPQ and bPQ, which are variants of our enhanced metric, $mPQ+$. For a comprehensive understanding of these metrics, we direct the readers to the CellViT paper for detailed explanations.

We showcase visual results on the Lizard dataset in Fig. 4, highlighting cleaner and more accurate segmentations and classifications compared to other methods. Our multi-task approach appears to learn a robust feature representation that captures distinguishing characteristics between classes. Furthermore, the ensemble ability of our model allows it to incorporate information that may be overlooked in single models. Additionally, we break down results by each cell category in the Lizard dataset in Table 4, showing top scores in every category for both $mPQ+$ and R^2 metrics compared to the HoVer-Net baseline. These consistent gains demonstrate the broad improvements from our proposed approach and validate the utility of our architecture modifications.

In summary, both quantitative metrics and qualitative results confirm the effectiveness of our multi-task framework relative to state-of-the-art methods for nuclear instance segmentation, classification, and composition regression. The joint training provides complementary information that boosts performance across related tasks.

5.2. Comparison to top teams in CoNIC 2022 challenge

We participated in the 2022 Colon Nuclei Identification and Counting Challenge (CoNIC 2022 Challenge) organized by [24], a competition aimed at advancing research on automatic nuclear instance characterization for computational pathology applications. The challenge involved segmenting, classifying, and counting 6 nuclear types in provided multi-organ datasets. Blinded test data, including samples from new domains not seen during training, posed a significant challenge. To generalize well to diverse data, high-performing methods were needed.

Submissions were evaluated based on mean panoptic quality ($mPQ+$) for segmentation and classification, and R^2 score for nuclear composition regression. Table 3 shows the final scores of the top 10 ranked teams on the test set. The top methods achieved very close scores, indicating highly competitive solutions were developed by leading research groups. Our proposed multi-task framework achieved **1st place** for the nuclear composition R^2 metric with a score of 0.76413, and 3rd place based on the $mPQ+$ metric with a score of 0.46310.

This demonstrates our method's ability to generalize successfully to new datasets while maintaining top-tier performance relative to state-of-the-art techniques. The joint multi-task learning provides our model with robust representations that transfer well to unseen domains. Our strong ranking in this competitive challenge confirms the effectiveness of the proposed approach for real-world biomedical image analysis tasks requiring generalization.

5.3. Ablation study

To investigate the importance of individual components in our proposed framework, we conduct an ablation study. We analyze the contributions of four key components: encoder backbones, data augmentation strategies, class imbalance handling methods, and model ensemble techniques. Understanding the individual contributions of these components is essential for further improving the model.

5.3.1. Encoder backbones

The performance of the encoder is a crucial factor in determining the overall success of the model. We experiment with various encoder architectures, as listed in Table 6, including ResNet50, ResNet101 [29], SEResNeXt50, SEResNeXt101 [33], and swin-transformer [7,45]. The SEResNeXt models incorporate ResNeXts with squeeze-and-excitation
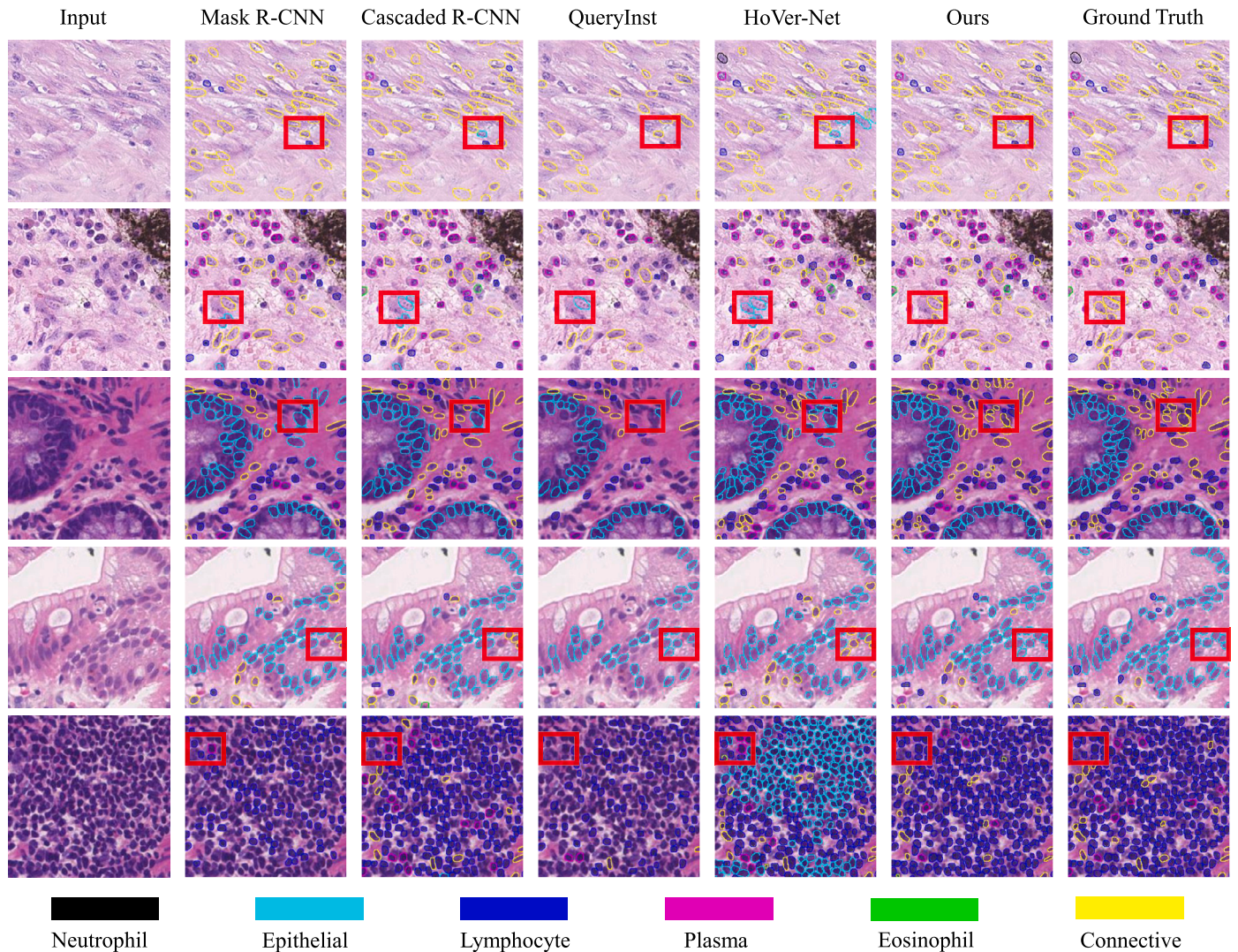
Fig. 4. Qualitative results comparing the proposed method against other state-of-the-art approaches on example images from the Lizard dataset [23]. The overlays visualize the output instance segmentation results, with different colors indicating the predicted nuclear categories. Examples of nuclei from each category are provided in Fig. 2. The proposed method produces more accurate predictions than other state-of-the-art techniques, as observed by examining the overlays. Our approach correctly identifies more instances, with fewer false positives and false negatives. Additionally, the nuclear category predictions match the ground truth more closely compared to other methods. This qualitative analysis highlights the improved performance of the proposed framework on this challenging nuclear recognition task.

Table 6

Evaluation of base models utilizing different encoder backbones. As shown, larger encoder models tend to achieve better or at least comparable performance to smaller models. Consistent with the findings in nnUNet-Revisited [35], we also observe that Transformer-based architectures do not surpass the performance of CNNs in our task. Thus, we employ SEResNeXt50 [33] and SEResNeXt101 [33] encoders in our framework.

	$mPQ+$			R^2		
	Lizard	PanNuke	MoNuSAC	Lizard	PanNuke	MoNuSAC
ResNet50	0.4869±0.0234	0.3437±0.0372	0.4569±0.0197	0.6921±0.1254	0.5560±0.0613	0.6356±0.1199
ResNet101	0.4980±0.0279	0.3902±0.0194	0.4686±0.0233	0.7313±0.1087	0.6378±0.0516	0.5487±0.1744
SEResNeXt50	0.5087±0.0327	0.4322±0.0216	0.4712±0.0156	0.7529±0.1143	0.6658±0.0622	**0.6356±0.0906**
SEResNeXt101	**0.5286±0.0373**	**0.4489±0.0196**	**0.4815±0.0142**	**0.7700±0.1667**	**0.6942±0.0479**	0.6288±0.1013
Swin-Transformer	0.4544±0.0683	0.3030±0.0203	0.4480±0.0062	0.6547±0.0642	0.6176±0.0957	0.5839±0.0631

Table 7
Impact on the performance of different data augmentation techniques applied during training using the SEResNeXt50 [33] encoder. Models trained with most augmentations outperform no augmentation, except distortion which is unstable. Thus, distortion is excluded from our final training pipeline.

	mPQ+	R^2
No Augmentation	0.5300±0.0128	0.7758±0.0318
Flip	0.5339±0.0126	0.7739±0.0288
Color Jitter	0.5382±0.0102	0.7962±0.0295
Blur	0.5356±0.0091	0.7915±0.0273
Distort	0.5198±0.0084	0.7628±0.0275

Table 8
Evaluation of dropout as a model regularization strategy. Dropout layers are included to prevent the model from developing an over-reliance on particular features, which could lead to increased sensitivity to variations in unseen test data.

	mPQ+	R^2
Without dropout	0.5504±0.0125	0.8394±0.0305
With dropout	0.5599±0.0086	0.8437±0.0454

Table 9
Evaluation of imbalanced class techniques. We explored various techniques to address class imbalance but found that they did not significantly improve performance. Therefore, we did not incorporate them into our method.

	mPQ+	R^2
Original	0.5599±0.0086	0.8437±0.0454
Focal Loss	0.5431±0.0112	0.8159±0.0306
Weighted Cross-Entropy	0.5520±0.0082	0.8374±0.0255

Table 10
Performance comparison across folds for Lizard and PanNuke datasets. The table details the performance of our model across the five folds of the Lizard dataset and the three folds of the PanNuke dataset. The Lizard dataset's composition, stemming from five distinct sources, justifies its division into separate folds for a comprehensive evaluation. For the PanNuke dataset, we adhered to the original data split as described in its publication. We excluded MoNuSAC from this analysis due to its singular test dataset, aligning with the original split for consistency with other studies.

Dataset	Split	Dice	mPQ+	R^2
Lizard	Fold 0	0.6557±0.0259	0.5599±0.0086	0.8437±0.0454
	Fold 1	0.6270±0.0185	0.5172±0.0096	0.7211±0.1257
	Fold 2	0.4779±0.0204	0.4768±0.0114	0.5982±0.0654
	Fold 3	0.6362±0.0223	0.5456±0.0096	0.7998±0.0941
	Fold 4	0.6483±0.0088	0.5603±0.0076	0.8366±0.0402
PanNuke	Fold 0	0.6355±0.0058	0.4738±0.0061	0.7127±0.0553
	Fold 1	0.6428±0.0102	0.4875±0.0099	0.7600±0.0658
	Fold 2	0.6442±0.0155	0.4892±0.0117	0.8041±0.0292

blocks to enhance representation learning, while the swin-transformer leverages attention mechanisms to emphasize important features.

Our results indicate that larger models, such as SEResNeXt101, tend to perform better due to their increased capacity to model complex patterns. However, this trend is not absolute, as evident from the slight outperformance of ResNet101 on the R^2 metric. Furthermore, very large models can overfit on the limited training data. Notably, our findings align with those in nnUNet-Revisited [35], suggesting that Transformer-based architectures do not surpass the performance of CNNs in our task. Based on these findings, we select SEResNeXt50 and SEResNeXt101 as our encoder backbones, as they achieve top performance on most datasets.

5.4. Data augmentation

Data augmentation is a powerful tool for improving the generalization of our model. We carefully select augmentations suitable for H&E histology images to avoid distorting important color and texture cues. As shown in Table 7, most techniques such as 1) random flips with a probability of 0.5, 2) random rotation with a probability of 0.5, 3) color jittering with brightness jitter of 0.2, contrast jitter of 0.25, saturation jitter of 0.2, and hue jitter of 0.5, 4) Gaussian blur with a blur limit of 3, 5) Median blur with a blur limit of 3, improve performance of our model by introducing more variation. However, we also observe that distortion augmentation can sometimes decrease performance metrics. This may due to over-augmentation. Thus, we exclude distortion and focus on other effective augmentation techniques to optimize our model's performance.

5.5. Class imbalance

Class imbalance is a common issue in nuclear instance segmentation, where the majority class (background) dominates the minority class (instances). To address this challenge, we experimented with focal loss and weighted cross-entropy loss on the Lizard dataset with our method. The results, as shown in Table 9, demonstrate that these techniques can indeed improve performance. However, we also acknowledge that these techniques often require dataset-spefic tuning, which can be time-consuming and labor-intensive. Nevertheless, our method still achieved good performance without the need for these techniques.

5.6. Cross validation

As shown in Table 10, our cross-validation experiments consistently show that our model performs robustly across different folds, indicating its strong generalization capabilities. The slight variability in results could possibly be attributed to the differences between the training and test datasets in each fold. Fold 2 of Lizard dataset stands out with results that are notably different from the rest, possibly due to unique characteristics or biases present in that specific subset. Despite this outlier, the overall consistency across the other folds confirms the reliability and generalization ability of our model.

5.7. Model regularization

By avoiding overfitting to specific, possibly spurious correlations within the training set, the model with dropout demonstrates improved robustness and accuracy when confronted with the variability inherent in real-world data. Our experimental results, as evidenced in Table 8, illustrate the improvement in model performance when dropout is employed as a regularization technique.

5.8. Model ensembling

Ensembling combines multiple diverse models to improve robustness. We use two ensemble strategies: 1) *intra-model*: averaging predictions from flips and rotations of the same input, and 2) *inter-model*: averaging predictions from different encoder backbones. Table 11 shows both provide consistent gains. The *intra-ensemble* leverages model smoothness over input variations, while the *inter-ensemble* benefits from complementary information in different backbones. Since real-world data varies, ensembling makes our framework adaptable.

In summary, our ablation analyses provide insight into optimal model architecture choices and training strategies for robust histology image analysis. The findings guide the design of our proposed multi-task framework.

Table 11

Evaluation of intra-model and inter-model ensemble techniques. "Intra-model" employs solely intra-model ensembling, and "Inter-Model" utilizes only inter-model ensembling. "Ours" combines both intra-model and inter-model ensembling strategies. The data demonstrates that ensemble methodologies consistently enhance models' performance.

	$mPQ+$			R^2		
	Lizard	PanNuke	MoNuSAC	Lizard	PanNuke	MoNuSAC
SEResNeXt50	0.5087±0.0327	0.4322±0.0216	0.4712±0.0156	0.7529±0.1143	0.6658±0.0622	0.6356±0.0906
SEResNeXt50 (Intra-model)	0.5488±0.0063	0.4326±0.0218	0.4916±0.0211	0.8455±0.0270	0.6659±0.0630	0.6988±0.0885
SEResNeXt101	0.5286±0.0373	0.4489±0.0196	0.4815±0.0142	0.7700±0.1667	0.6942±0.0479	0.6288±0.1013
SEResNeXt101 (Intra-model)	0.5482±0.0264	0.4506±0.0185	0.5011±0.0228	0.8009±0.1796	0.6955±0.0481	0.6904±0.1020
Inter-model	0.5352±0.0159	0.4686±0.0082	0.5101±0.0255	0.8185±0.0361	**0.7260±0.0528**	0.7134±0.1027
Ours (ensemble all)	**0.5599±0.0086**	**0.4738±0.0061**	**0.5567±0.0125**	**0.8437±0.0454**	0.7127±0.0553	**0.7968±0.0470**

6. Discussion

6.1. Practical usage

Our proposed method has the potential to be used in various practical applications in digital pathology. One potential application is in estimating tumor burden by counting the number of tumor nuclei and dividing the number by the total number of nuclei. This could be a valuable tool for clinicians, as it could aid in prognostic estimation and patient treatment planning for neo-adjuvant therapy. Our model could be embedded in a software that enables pathologists to view whole-slide images with our segmentation overlay, similar to Fig. 4, and our cellular composition prediction results displayed alongside. This could potentially aid doctors in their diagnostic workflow and ultimately improve patient care.

6.2. Limitations

A notable constraint of our approach is the computational complexity in our method's ensembling strategy. The necessity to process data across various base models in multiple iterations can lead to increased processing times, which may be considered a drawback for applications demanding real-time analytical capabilities. The computational intensity is primarily due to the extensive parameterization of our model, with a total of 88,739,206 parameters. This parameter count consists of the two encoder models: 33,671,843 parameters for the SEResNeXt50-based model and 55,067,363 parameters for the SEResNeXt101-based model. In comparison, the original HoVerNet model comprises a more modest 37,639,691 parameters.

Furthermore, the inference time of our method, measured on a NVidia 3090 GPU, stands at 0.1421 seconds per image. This is notably higher than the 0.0262 seconds required by HoVerNet under identical conditions. The increased inference time can be attributed to the tripled model inference demands for each intra-model ensemble, compounded by the inter-model ensemble, which itself requires two distinct intra-model ensembles. While this imposes certain limitations on the speed of our method, it is a reasonable trade-off for the enhanced accuracy and robustness that our ensembling strategy provides.

7. Conclusion

In this study, we present an enhanced framework for nuclear instance segmentation, classification, and composition prediction in histology images. Our approach builds upon the popular HoVer-Net architecture and incorporates several enhancements to improve its performance. Specifically, we employed advanced SEResNeXt encoders to extract richer features, modified the decoder for improved efficiency, and added dropout for regularization. Furthermore, we implemented a two-level model ensembling strategy to enhance generalization capabilities. Extensive experiments demonstrate the effectiveness of these contributions, resulting in significant complementary performance improvements that enable our framework to surpass state-of-the-art methods on public benchmarks. Our approach can be effortlessly extended to nuclear counting and composition regression tasks, achieving top results on the task. Ablation studies provide insights into optimal architectures

and training strategies for robust multi-task histology image analysis. The proposed framework has the potential to facilitate accurate cellular characterization and analysis in various pathology applications. Future work involves expanding the framework to additional modalities and pathology applications requiring accurate cellular characterization. To facilitate further research in this direction, we have released our code and models. Overall, this study demonstrates the effectiveness of our enhanced framework for nuclear instance segmentation, classification, and composition prediction in histology images, and its potential applications in computational pathology.

CRediT authorship contribution statement

Wenhua Zhang: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Sen Yang:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Meiwei Luo:** Writing – original draft, Formal analysis. **Chuan He:** Writing – review & editing, Resources, Formal analysis. **Yuchen Li:** Writing – review & editing, Supervision, Project administration, Formal analysis. **Jun Zhang:** Writing – review & editing, Supervision, Resources. **Xiyue Wang:** Writing – review & editing, Supervision, Methodology, Investigation. **Fang Wang:** Writing – review & editing, Supervision, Resources.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used Claude-instant in order to polish the language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors confirm that there are no conflicts

References

[1] Abdel-Nasser M, Singh VK, Mohamed EM. Efficient staining-invariant nuclei segmentation approach using self-supervised deep contrastive network. Diagnostics 2022;12:3024.

[2] Alsubaie N, Sirinukunwattana K, Raza SEA, Snead D, Rajpoot N. A bottom-up approach for tumour differentiation in whole slide images of lung adenocarcinoma. In: Medical imaging 2018: digital pathology, international society for optics and photonics; 2018. p. 105810E.

[3] Bancher B, Mahbod A, Ellinger I, Ecker R, Dorffner G. Improving mask r-cnn for nuclei instance segmentation in hematoxylin & eosin-stained histological images. In: MICCAI workshop on computational pathology, PMLR; 2021. p. 20–35.

[4] Bersanelli M. Tumour mutational burden as a driver for treatment choice in resistant tumours (and beyond). Lancet Oncol 2020;21:1255–7.

[5] Bühlmann P. Bagging, boosting and ensemble methods. In: Handbook of computational statistics. Springer; 2012. p. 985–1022.

[6] Cai Z, Vasconcelos N. Cascade r-cnn: high quality object detection and instance segmentation. IEEE Trans Pattern Anal Mach Intell 2019;1. https://doi.org/10.1109/tpami.2019.2956516.

[7] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: unet-like pure transformer for medical image segmentation. In: Proceedings of the European conference on computer vision workshops (ECCVW); 2022.

[8] Chan SW, Leung KS, Wong WF. An expert system for the detection of cervical cancer cells using knowledge-based image analyzer. Artif Intell Med 1996;8:67–90.

[9] Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst Appl 2011;38:9014–22.

[10] Ciresan D, Giusti A, Gambardella L, Schmidhuber J. Deep neural networks segment neuronal membranes in electron microscopy images. Adv Neural Inf Process Syst 2012;25.

[11] Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. In: International conference on medical image computing and computer-assisted intervention. Springer; 2013. p. 411–8.

[12] Comaniciu D, Meer P, Foran DJ. Image-guided decision support system for pathology. Mach Vis Appl 1999;11:213–24.

[13] Dawood M, Branson K, Rajpoot NM, Minhas F. Albrt: cellular composition prediction in routine histology images. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 664–73.

[14] Dogar GM, Shahzad M, Fraz MM. Attention augmented distance regression and classification network for nuclei instance segmentation and type classification in histology images. Biomed Signal Process Control 2023;79:104199.

[15] Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson AN, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA 2015;313:1122–32.

[16] Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 1991;19:403–10.

[17] Fang Y, Yang S, Wang X, Li Y, Fang C, Shan Y, et al. Instances as queries. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV); 2021. p. 6910–9.

[18] Fleming M, Ravula S, Tatishchev SF, Wang HL. Colorectal carcinoma: pathologic aspects. J Gastroint Oncol 2012;3:153.

[19] Galli F, Aguilera JV, Palermo B, Markovic SN, Nisticò P, Signore A. Relevance of immune cell and tumor microenvironment imaging in the new era of immunotherapy. J Exp Clin Cancer Res 2020;39:1–21.

[20] Gamper J, Koohbanani NA, Benet K, Khuram A, Rajpoot N. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: European congress on digital pathology. Springer; 2019. p. 11–9.

[21] Gamper J, Koohbanani NA, Graham S, Jahanifar M, Khurram SA, Azam A, et al. Pannuke dataset extension, insights and baselines. arXiv preprint. arXiv:2003.10778, 2020.

[22] Graham S, Chen H, Gamper J, Dou Q, Heng PA, Snead D, et al. Mild-net: minimal information loss dilated network for gland instance segmentation in colon histology images. Med Image Anal 2019;52:199–211.

[23] Graham S, Jahanifar M, Azam A, Nimir M, Tsang YW, Dodd K, et al. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 684–93.

[24] Graham S, Jahanifar M, Vu QD, Hadjigeorghiou G, Leech T, Snead D, et al. Conic: Colon nuclei identification and counting challenge 2022. arXiv:2111.14485, 2021.

[25] Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal 2019;101563.

[26] Haroske G, Dimmer V, Friedrich K, Meyer W, Thieme B, Theissig F, et al. Nuclear image analysis of immunohistochemically stained cells in breast carcinomas. Histochem Cell Biol 1996;105:479–85.

[27] Hassan L, Saleh A, Abdel-Nasser M, Omer OA, Puig D. Efficient multi-organ multi-center cell nuclei segmentation method based on deep learnable aggregation network. Trait Signal 2021;38.

[28] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2961–9.

[29] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.

[30] He S, Minn KT, Solnica-Krezel L, Anastasio MA, Li H. Deeply-supervised density regression for automatic cell counting in microscopy images. Med Image Anal 2021;68:101892.

[31] Hörst F, Rempe M, Heine L, Seibold C, Keyl J, Baldini G, et al. Cellvit: vision transformers for precise cell segmentation and classification. Med Image Anal 2024;94:103143.

[32] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. arXiv:1709.01507, 2019.

[33] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7132–41.

[34] Irshad H, Veillard A, Roux L, Racoceanu D. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. IEEE Rev Biomed Eng 2013;7:97–114.

[35] Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, et al. nnu-net revisited: a call for rigorous validation in 3d medical image segmentation. arXiv preprint. arXiv:2404.09556, 2024.

[36] Jia J, Zhu F, Ma X, Cao ZW, Li YX, Chen YZ. Mechanisms of drug combinations: interaction and network perspectives. Nat Rev Drug Discov 2009;8:111–28.

[37] Khan A, Gould S, Salzmann M. Deep convolutional neural networks for human embryonic cell counting. In: European conference on computer vision. Springer; 2016. p. 339–48.

[38] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint. arXiv:1412.6980, 2014.

[39] Kirillov A, He K, Girshick R, Rother C, Dollár P. Panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 9404–13.

[40] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision; 2023. p. 4015–26.

[41] Kotu V, Deshpande B. Data mining process. Predict Anal Data Min 2015;1:17–36.

[42] Kotu V, Deshpande B. Data science: concepts and practice. Morgan Kaufmann; 2018.

[43] Li H, Jiang G, Zhang J, Wang R, Wang Z, Zheng WS, et al. Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. NeuroImage 2018;183:650–65.

[44] Lin C, Chen W, Qiu C, Wu Y, Krishnan S, Zou Q. Libd3c: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing 2014;123:424–35.

[45] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV); 2021.

[46] Lu C, Romo-Bucheli D, Wang X, Janowczyk A, Ganesan S, Gilmore H, et al. Nuclear shape and orientation features from h&e images predict survival in early-stage estrogen receptor-positive breast cancers. Lab Invest 2018;98:1438–48.

[47] Polikar R. Ensemble based systems in decision making. IEEE Circuits Syst Mag 2006;6:21–45.

[48] Prabhu AV, Bishnoi J, Minu R. Segmentation of neoplastic cell nuclei for assisted cell labelling using mask r-cnn.

[49] Schmidt U, Weigert M, Broaddus C, Myers G. Cell detection with star-convex polygons. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Proceedings, part II; 2018. p. 265–73.

[50] Sedaghat-Hamedani F, Haas J, Zhu F, Geier C, Kayvanpour E, Liss M, et al. Clinical genetics and outcome of left ventricular non-compaction cardiomyopathy. Eur Heart J 2017;38:3449–60.

[51] Shen L, Chen H, Yu Z, Kang W, Zhang B, Li H, et al. Evolving support vector machines using fruit fly optimization for medical data classification. Knowl-Based Syst 2016;96:61–75.

[52] Sirinukunwattana K, Pluim JP, Chen H, Qi X, Heng PA, Guo YB, et al. Gland segmentation in colon histology images: the glas challenge contest. Med Image Anal 2017;35:489–502.

[53] Sirinukunwattana K, Raza SEA, Tsang YW, Snead DR, Cree IA, Rajpoot NM. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging 2016;35:1196–206.

[54] Sirinukunwattana K, Snead D, Epstein D, Aftab Z, Mujeeb I, Tsang YW, et al. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Sci Rep 2018;8:1–13.

[55] Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. Med Image Anal 2020:101813.

[56] Verma R, Kumar N, Patil A, Kurian NC, Rane S, Graham S, et al. Monusac2020: a multi-organ nuclei segmentation and classification challenge. IEEE Trans Med Imaging 2021;40:3413–23.

[57] Vo VTT, Kim SH. Mulvernet: nucleus segmentation and classification of pathology images using the hover-net and multiple filter units. Electronics 2023;12:355.

[58] Wang J, Qin L, Chen D, Wang J, Han BW, Zhu Z, et al. An improved hover-net for nuclear segmentation and classification in histopathology images. Neural Comput Appl 2023;35:14403–17.

[59] Weigert M, Schmidt U. Nuclei instance segmentation and classification in histopathology images with stardist. In: The IEEE international symposium on biomedical imaging challenges (ISBIC); 2022.

[60] Weigert M, Schmidt U, Haase R, Sugawara K, Myers G. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In: The IEEE winter conference on applications of computer vision (WACV); 2020.

[61] Xing F, Su H, Neltner J, Yang L. Automatic ki-67 counting using robust cell detection and online dictionary learning. IEEE Trans Biomed Eng 2013;61:859–70.

[62] Xue Y, Ray N, Hugh J, Bigras G. Cell counting by regression using convolutional neural network. In: European conference on computer vision. Springer; 2016. p. 274–90.

[63] Yao K, Huang K, Sun J, Hussain A, Jude C. Pointnu-net: simultaneous multi-tissue histology nuclei segmentation and classification in the clinical wild. arXiv preprint. arXiv:2111.01557, 2021.

[64] Zhu Y, Chen Z, Zheng Y, Zhang Q, Wang X. Real-time cell counting in unlabeled microscopy images. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 694–703.

[65] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet 2018;9:515.

[66] Xiyue W, Junhan Z, Marostica E, Yuan W, Jietian J, Jiayu Z, Ruijiang L, Hongping T, Kanran W, Yu L, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature Publishing Group UK London 2024:1–9.

[67] Simon G, Quoc Dang V, Mostafa J, Martin W, Uwe S, Wenhua Z, Jun Z, Sen Y, Jinxi X, Xiyue W, et al. CoNIC Challenge: pushing the frontiers of nuclear detection, segmentation, classification and counting. Med Image Anal 2024;92:103047.