RT-ISTR: A REAL-TIME INSTANCE SEGMENTATION TRANSFORMER NETWORK FOR MICROSCOPIC IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-time instance segmentation in microscopic imaging is hindered by the trade-off between pixel-wise accuracy and computational efficiency. Existing Transformer-based models deliver high precision but suffer from prohibitive parameter counts and latency, while CNN-based alternatives sacrifice boundary quality. We propose RT-ISTR, an end-to-end Transformer framework that unifies detection and segmentation with minimal overhead. Our lightweight Multi-scale Mask Enhancement Module fuses shallow spatial details from the backbone with deep semantic features. It performs this fusion via hierarchical upsampling and element-wise addition. As a result, the module produces rich prototype masks.The detection branch's query embeddings are reused to dynamically query these prototypes through dot-product, enabling pixel-level masks at negligible extra cost. On COCO, RT-ISTR-L achieves 44.6% AP^{mask} with 102 GFLOPs and 32M parameters, running at 64 FPS, outperforming YOLOv8, YOLO11, and YOLO26 in mask accuracy. RT-ISTR challenges the paradigm that high-accuracy segmentation demands heavy models, offering an efficient, deployable solution for real-time microscopic analysis and edge-based applications.

1 INTRODUCTION

Microscopic image analysis is a cornerstone of computational biology, enabling automated quantification and interpretation of subcellular and tissue architectures from optical and electron microscopy data Mo et al. (2024); Wei et al. (2025). The computational tasks underpinning this domain are broadly categorised into object detection, semantic segmentation, and instance segmentation Han et al. (2026); Long et al. (2015); Zhou et al. (2022). Among these, object detection localises targets using bounding boxes and assigns categorical labels, whereas instance segmentation further distinguishes individual objects by generating pixel-level masks. This fine-grained capability is particularly critical for accurately extracting morphological features in complex scenes He et al. (2017); Bolya et al. (2019); Li et al. (2023).

In recent years, rapid advancements in deep learning architectures for visual recognition, particularly convolutional neural networksHe et al. (2016); Liu et al. (2022) and vision transformersDosovitskiy (2020); Liu et al. (2021); Siméoni et al. (2025), have substantially elevated the performance of instance segmentation models. This progress has facilitated the application of instance segmentation to a variety of biomedical applications, including whole-slide images such as those in the CoNIC challenge Graham et al. (2024) and Ocelot 2023 Ryu et al. (2023), as well as cellular microscopic image analysis in SegPC-2021 Gupta et al. (2023) and MitoEM Franco-Barranco et al. (2023). Despite these successes, existing general-purpose instance segmentation models Chen et al. (2021); Wang et al. (2023); Roy et al. (2023); Zhou et al. (2020); Li et al. (2023); Cheng et al. (2022), except for the YOLO Jocher et al. (2023) model, still face substantial challenges in real-time detection. Moreover, the inherently complex backgrounds and blurred edges in microscopic cellular images often lead to degraded generalization of off-the-shelf models, necessitating the design of specialized, domain-adaptive segmentation strategies tailored to the characteristics of microscopic data.

Despite the proliferation of instance segmentation models, existing methodologies are broadly categorised into CNN-basedHe et al. (2017); Jocher et al. (2023) and Transformer-basedCarion et al. (2020); Hu et al. (2021) architectures, each presenting inherent trade-offs that limit their direct applicability to microscopic imagery. CNN-based two-stage methods, such as the Mask R-CNN

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

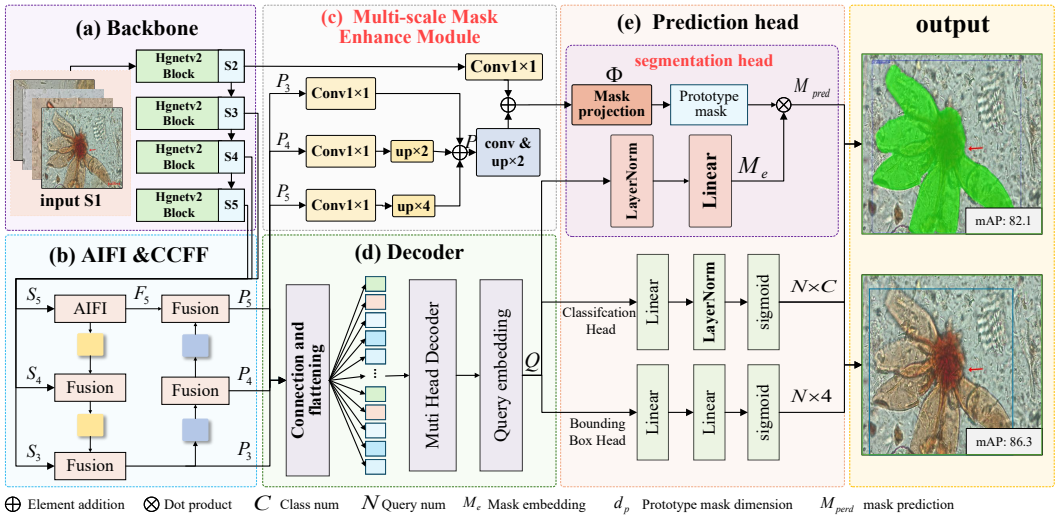

Figure 1: Overall structure of the model. The proposed RT-ISTR consists of five components: (a) a feature extraction network that collects multi-scale features (*i.e.,* a feature pyramid network / FPN); (b) the AIFI and CCFF modules from D-FINE are adopted for cross-scale information fusion; (c) a multi-scale mask enhancement module is used to fuse multi-scale features to generate a prototype mask; (d) a decoder that generates query embeddings required for both segmentation and detection; and (e) prediction heads, where a segmentation head is added to enable instance segmentation.

seriesHuang et al. (2019); Chen et al. (2019a); Cai & Vasconcelos (2019), achieve commendable mask quality via region proposal networks and RoIAlign operations, but their sequential pipeline incurs prohibitive computational overhead for real-time processing. Conversely, single-stage CNN variantsLong et al. (2015); Jocher et al. (2023) like YOLACT Bolya et al. (2019) pursue architectural simplicity by generating linear combinations of prototype masks, offering high inference speed at the cost of compromised boundary delineation in dense, overlapping cellular scenes. On the other hand, Transformer-based end-to-end frameworksCarion et al. (2020), including ISTRHu et al. (2021), Mask2FormerCheng et al. (2022), and Mask DINO Hu et al. (2021); Li et al. (2023), have reformed the paradigm through mask classification and cross-attention mechanisms, delivering superior flexibility and global context modelling. Nevertheless, their reliance on heavy query embeddings and pixel decoders results in excessive parameter counts (e.g., the lightest RF-DETRRobinson et al. (2026) variant still exceeds 30M parameters), rendering them impractical for deployment on resource-constrained edge devices commonly used in high-throughput microscopy.

Critically, the aforementioned generic models are predominantly tailored for natural imagery, overlooking the distinct characteristics of microscopic data. For example, the traditional Chinese medicine microscope has the characteristics of complex image background and unclear edge. When applied off-the-shelf, these models either sacrifice fine-grained mask fidelity to maintain tolerable frame rates or demand computational budgets far exceeding the latency constraints of real-time pathological screening and cellular phenotyping. This dilemma necessitates a dedicated, domain-adaptive solution that reconciles high-resolution spatial fidelity with deep semantic extraction, while remaining computationally tractable. To this end, this study pivots towards addressing two pivotal challenges: (i) How to efficiently integrate high-resolution spatial details with deep semantic features within a lightweight real-time detection backbone to recover pixel-accurate masks; and (ii) How to leverage the detection branch's query embeddings to dynamically parameterise the mask generation process, thereby boosting instance discrimination in densely packed microscopic environments.

To address the above issues, we propose a Real-Time Instance Segmentation Transformer Network, namely **RT-ISTR**. Specifically, we design a lightweight and efficient Multi-scale Mask Enhancement Module that deeply fuses the high-resolution pixel embedding maps from the backbone network with the features enhanced by the efficient encoder, thereby generating enhanced prototype masks. Furthermore, we extend a mask prediction branch in parallel based on D-FINEPeng et al. (2025), reusing its bounding-box query embeddings to dynamically query the content of the proto-

type masks, thereby yielding pixel-level predictions. Consequently, RT-ISTR achieves end-to-end instance segmentation and object detection based on a Transformer while maintaining low computational overhead.

Our main contributions can be summarised as follows:

- We propose a lightweight Multi-scale Mask Enhancement Module that deeply fuses high-resolution features from the backbone with encoder-enhanced features to generate enhanced prototype masks, thereby providing high-quality pixel-level representations for subsequent instance segmentation.
- We utilise the query embeddings of D-FINE to perform dot-product querying over the prototype masks for segmentation content, dynamically retrieving prototype mask content, thereby filling the gap in D-FINE's segmentation tasks.
- Leading real-time instance segmentation performance is achieved on the COCOLin et al. (2014) dataset, and accurate segmentation of irregular cell morphologies is realised on a microscopic cell dataset, providing effective technical support for microscopic identification.

2 METHODOLOGY

Notations. The input image is denoted as $\mathbf{S}_1 \in \mathbb{R}^{H\times W\times 3}$. The multi-scale features output by the backbone are represented as $\mathbf{S}_l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times d}$, where $l \in \{2,3,4,5\}$ corresponds to downsampling factors $4, 8, 16, 32$, respectively. After adaptive feature interaction and cross-scale fusion, the enhanced features are denoted as $\mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5$. The decoder produces N query embeddings $\mathbf{Q} \in \mathbb{R}^{N\times d}$, which are used to predict class labels $\mathbf{c}$, bounding boxes $\mathbf{b}$, and mask embeddings $\mathbf{M}_e \in \mathbb{R}^{N\times d_p}$. The prototype masks are denoted as $\mathbf{P}_m \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times d_p}$, and the final mask predictions are given by $\mathbf{M}_{\text{pred}} \in \mathbb{R}^{N\times\frac{H}{4}\times\frac{W}{4}}$, obtained via matrix multiplication between $\mathbf{M}_e$ and $\mathbf{P}_m$.

Problem Definition. Object detection predicts class labels $\mathbf{c}$ and bounding boxes $\mathbf{b}$; instance segmentation further predicts binary masks $\mathbf{m}_i$ for pixel-level localization. For efficiency, masks are typically downsampled and later upsampled. Extending D-FINE (originally designed for detection), our framework reuses the same backbone and encoder, adds prototype mask interaction via query embeddings, and shares decoder queries for both tasks. Mask embeddings derived from queries are multiplied with prototype masks $\mathbf{P}_m$ to produce final predictions. Thus, detection and segmentation share feature extraction and decoding, with segmentation only introducing prototype interaction to enable real-time pixel-level precision.

2.1 MOTIVATION

Microscopic image analysis requires both precise instance localisation and pixel-accurate segmentation in densely packed, morphologically complex scenes. Existing instance segmentation methods, however, fall short for real-time microscopy applications. CNN-based approaches like Mask R-CNN He et al. (2017) and HTC Chen et al. (2019a) achieve reasonable accuracy but rely on multi-stage pipelines that introduce prohibitive latency. Transformer-based frameworks such as Mask2Former Cheng et al. (2022) and Mask DINO Li et al. (2023) offer superior global context modelling. Yet, their heavy query decoders and pixel attention mechanisms incur excessive parameters and computational costs, making them impractical for edge devices.

Moreover, these generic models are designed for natural imagery and overlook the distinct characteristics of microscopy data, *i.e.,* cluttered backgrounds, ambiguous boundaries, and severe occlusions. Off-the-shelf deployment forces an undesirable trade-off, sacrificing mask fidelity for speed or exceeding real-time budgets for accuracy. This motivates our pursuit of two research questions: (i) How can high-resolution spatial details and deep semantics be fused efficiently for pixel-accurate masks? (ii) How can detection query embeddings be reused to dynamically generate masks without introducing heavy auxiliary modules?

To address these challenges, we propose RT-ISTR, a real-time instance segmentation transformer that introduces a lightweight Multi-scale Mask Enhancement Module and reuses detection queries

for mask prediction with minimal overhead. The overall framework is illustrated in Figure 1, with each component detailed in the following subsections.

2.2 Feature Extraction

Given an input image $\mathbf{S}_1 \in \mathbb{R}^{H\times W\times 3}$, we adopt an Hgnetv2 PaddlePaddle (2019) as the backbone network for multi-scale feature extraction. Through a hierarchical downsampling structure, this network produces four feature maps at different resolutions, which can be formally expressed as:

$$\{\mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4, \mathbf{S}_5\} = f_{\text{Hgnetv2}}(\mathbf{S}_1), \tag{1}$$

where $f_{\text{Hgnetv2}}(\cdot)$ denotes the forward propagation of the backbone network. Specifically, $\mathbf{S}_l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times d}$ $(l = 2, 3, 4, 5)$ corresponds to feature layers with downsampling factors of $4, 8, 16, 32$, respectively. Among these, $\mathbf{S}_3, \mathbf{S}_4, \mathbf{S}_5$ are employed for multi-scale modeling in the D-FINE object detection branch, whereas the high-resolution feature $\mathbf{S}_2$ preserves rich spatial details and provides fine-grained pixel-level features for subsequent prototype mask generation in the instance segmentation branch, thereby contributing to improved mask edge accuracy and small-object segmentation performance.

2.3 Attention-based Intra-scale Feature Interaction and CNN-based Cross-scale Feature Fusion

After obtaining the multi-scale features $\mathbf{S}_3, \mathbf{S}_4, \mathbf{S}_5$ from the backbone network, the Attention-based Intra-scale Feature Interaction (AIFI) and CNN-based Cross-scale Feature Fusion (CCFF) modules from D-FINE are introduced to enhance semantic representations and facilitate sufficient interaction among multi-scale information. Specifically, AIFI is first applied to the highest semantic-level feature map $\mathbf{S}_5$, where long-range dependencies within the image are modeled via a self-attention mechanism to capture global contextual information. This process can be formalized as:

$$\mathbf{F}_5 = f_{\text{AIFI}}(\mathbf{S}_5), \tag{2}$$

where $f_{\text{AIFI}}(\cdot)$ denotes the adaptive feature interaction operation based on an efficient encoder, and its output $\mathbf{F}_5 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times d}$ preserves the original resolution while enhancing semantic expression.

Subsequently, the CCFF module, which is specifically improved under the inspiration of the GLEAN architecture, performs cross-scale fusion on $\mathbf{S}_3$, $\mathbf{S}_4$, and the enhanced $\mathbf{F}_5$. Through a top-down pathway that progressively propagates high-level semantics, supplemented by learnable fusion weights, an enhanced feature pyramid with strong semantic information and moderate spatial details is generated:

$$\{\mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5\} = f_{\text{CCFF}}(\mathbf{S}_3, \mathbf{S}_4, \mathbf{F}_5), \tag{3}$$

where $\mathbf{P}_l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times d}$ $(l = 3, 4, 5)$ corresponds to the enhanced feature maps with downsampling factors of $8, 16, 32$, respectively. The specific implementation details of the above AIFI and CCFF modules can be found in the literature Peng et al. (2025). Through these modules, global attention modeling and multi-scale fusion can be efficiently performed on the last three feature layers of the backbone, thereby providing more robust feature representations for subsequent detection and segmentation tasks.

2.4 Multi-scale Mask Enhancement Module

The quality of instance masks relies critically on the simultaneous preservation of fine spatial details and robust semantic context. In our framework, the shallow backbone feature $\mathbf{S}_2$ retains high-resolution spatial cues such as object boundaries and edge contours, but lacks high-level semantics. Conversely, the encoder outputs $\{\mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5\}$ carry rich categorical and global contextual information but suffer from coarse spatial resolution. To bridge this gap without introducing prohibitive computational overhead, we design a lightweight Multi-scale Mask Enhancement module. It synergistically integrates the high-resolution spatial details from $\mathbf{S}_2$ with the deep semantics from $\mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5$, generating high-fidelity prototype masks for the subsequent segmentation head.

The MME module achieves efficient cross-scale integration through channel alignment, progressive upsampling, and element-wise addition. The inputs are the pyramid features $\{\mathbf{S}_2, \mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5\}$.

We adopt a two-stage progressive fusion strategy rather than a direct one-step upsampling. This progressive design allows deep semantic signals to be gradually refined as they merge with shallower layers, mitigating aliasing and semantic dilution that often occur when directly upsampling high-level features by a large factor.

We first aggregate the three encoder-output semantic features. Independent 1×1 convolutions are applied to $\mathbf{P}_3$, $\mathbf{P}_4$, and $\mathbf{P}_5$ for channel alignment. The transformed $\mathbf{P}_4$ and $\mathbf{P}_5$ are then bilinearly upsampled by factors of 2 and 4, respectively, to match the resolution of $\mathbf{P}_3$ (*i.e.,* $\frac{H}{8} \times \frac{W}{8}$). These aligned features are fused via element-wise addition to produce an intermediate semantically rich feature $\mathbf{P}$:

$$\begin{aligned}\mathbf{P} =& \text{Conv}_{1\times1}^{P_3}(\mathbf{P}_3) + Up_{2\times}\big(\text{Conv}_{1\times1}^{P_4}(\mathbf{P}_4)\big) + \\ & Up_{4\times}\big(\text{Conv}_{1\times1}^{P_5}(\mathbf{P}_5)\big),\end{aligned} \tag{4}$$

where $\mathbf{P} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times d}$. This top-down pathway propagates deep semantics to the mid-level, providing a semantically grounded foundation for the subsequent spatial enhancement.

To embed fine-grained spatial structures, we project the shallow high-resolution feature $\mathbf{S}_2$ via a separate 1×1 convolution, yielding $\mathbf{S}_2' \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times d}$. Meanwhile, the intermediate feature $\mathbf{P}$ is passed through another 1×1 convolution (denoted as $\text{Conv}_{1\times1}^{P}$) to adapt its feature distribution for better compatibility with the shallow space. The transformed $\mathbf{P}$ is bilinearly upsampled by a factor of 2 and element-wise added to $\mathbf{S}_2'$, producing the final fused feature map:

$$\mathbf{F} = \mathbf{S}_2' + \text{Upsample}_{2\times}\big(\text{Conv}_{1\times1}^{P}(\mathbf{P})\big), \tag{5}$$

where $\mathbf{F} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times d}$ preserves the high resolution of $\mathbf{S}_2$ while being enriched by deep semantics.

Given the fused feature $\mathbf{F}$, a mask projection head (consisting of a 1×1 convolution, normalisation, and non-linear activation) maps it to the prototype mask space:

$$\mathbf{P}_m = \Phi(\mathbf{F}), \tag{6}$$

where $\mathbf{P}_m \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times d_p}$, $\Phi(\cdot)$ denotes the projection operation and d_p is the prototype dimension. Conceptually, $\mathbf{P}_m$ can be interpreted as a set of d_p learnable basis masks. Each channel captures a distinct shape primitive, *e.g.,* horizontal edges, circular blobs, or corner patterns. The hyperparameter d_p controls the representational capacity of this basis.

Notably, the entire MME module is composed solely of bilinear upsampling and lightweight 1×1 convolutions. It avoids heavy transposed convolutions, multi-head self-attention, or iterative refinement, which are commonly used in mask decoders, *e.g.,* Mask2Former. Consequently, it introduces negligible overhead to the baseline detection framework, preserving the real-time inference capability required for edge-based microscopic analysis.

2.5 Detection and Segmentation Heads

In the baseline D-FINE detector, the decoder produces a set of query embeddings $\mathbf{Q} \in \mathbb{R}^{N\times d}$. These queries are fed into a classification head and a bounding box regression head, yielding the final class predictions $\mathbf{c} \in \mathbb{R}^{N\times C}$ and box coordinates $\mathbf{b} \in \mathbb{R}^{N\times 4}$, respectively.

A critical design choice in RT-ISTR is to reuse the same $\mathbf{Q}$ from the detection branch, for instance, mask generation, rather than introducing a separate set of learnable mask queries or a heavy pixel decoder *e.g.,* as in Mask2Former or Mask DINO. This raises a legitimate concern: detection queries are optimised primarily for localisation, i.e., bounding-box centre, scale, and extent, whereas mask prediction requires fine-grained, pixel-level discrimination both internally and at the edges. However, we argue that after being iteratively refined across multiple decoder layers, the query embeddings $\mathbf{Q}$ in D-FINE evolve beyond mere positional anchors. Through cross-attention with multi-scale encoder features, these queries gradually aggregate rich instance-level semantics, encompassing global shape, pose, and coarse appearance. Consequently, $\mathbf{Q}$ serves as a holistic instance representation that is readily transferable to mask prediction. A simple linear projection is sufficient to bridge the minor task gap, as it re-encodes the existing semantic content of $\mathbf{Q}$ into the mask embedding space, without requiring expensive task-specific query branches or dense pixel-level attention.

Specifically, we first pass $\mathbf{Q}$ through a single linear layer to predict mask embeddings $\mathbf{M_e} \in \mathbb{R}^{N \times d_p}$, *i.e.,*

$$\mathbf{M_e} = \mathbf{W}^{\mathrm{T}}\mathbf{Q} + \mathbf{b}, \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{d \times d_p}$ and $\mathbf{b} \in \mathbb{R}^{d_p}$ are learnable parameters, and d_p is the prototype dimension. This projection aligns the feature dimension of the instance-level queries with that of the prototype masks $\mathbf{P}_m$. The final instance masks $\mathbf{M}_{\mathrm{pred}}$ are then obtained by a simple dot-product between the mask embeddings and the prototype masks, *i.e.,*

$$\mathbf{M}_{\mathrm{pred}} = \mathbf{M_e} \otimes \mathbf{P}_m^{\mathrm{T}}, \tag{8}$$

where $\otimes$ denotes matrix multiplication, yielding $\mathbf{M}_{\mathrm{pred}} \in \mathbb{R}^{N \times \frac{H}{4} \times \frac{W}{4}}$. This operation dynamically weights each prototype basis channel according to the semantic content of each query, thereby activating the relevant spatial regions to produce the final pixel-wise mask for each instance.

This design is central to preserving RT-ISTR's real-time inference speed while still achieving competitive mask accuracy, making it particularly suitable for resource-constrained microscopic analysis scenarios.

2.6 Mask Optimization Objective

To supervise the training of the proposed RT-ISTR, we adopt a multi-task loss that combines the original detection losses from the D-FINE baseline with two newly introduced mask-specific losses. Following Peng et al. (2025), the detection branch is optimised using five loss terms, *i.e.,*

(i) $\mathcal{L}_{\text{bbox}}$, the standard ℓ_1 loss for bounding box regression; (ii) $\mathcal{L}_{\text{vfl}}$, the Varifocal Loss Zhang et al. (2021) for class prediction, which down-weights the contribution of negative samples to address class imbalance; (iii) $\mathcal{L}_{\text{giou}}$, the Generalized IoU Loss Yu et al. (2016) for improving box localization accuracy; (iv) $\mathcal{L}_{\text{ddf}}$, the Distribution Discrepancy Refinement Loss introduced in D-FINE, which refines the predicted bounding box distributions; and (v) $\mathcal{L}_{\text{fgl}}$, the Focal Loss adapted for foreground-background classification in the query matching process.

For instance segmentation, we introduce two complementary losses to optimise the predicted masks $\mathbf{M}_{\text{pred}}$, *i.e.,* (i) $\mathcal{L}_{\text{focal}}$ Lin et al. (2017) is applied per-pixel to address the extreme foreground-background imbalance inherent in binary mask prediction. (ii) $\mathcal{L}_{\text{dice}}$ Milletari et al. (2016) is the Dice loss, which maximises the overlap between the predicted mask and the ground-truth mask.

The total loss is formulated as a weighted sum of the detection and segmentation losses, *i.e.,*

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \lambda_1\mathcal{L}_{\text{bbox}} + \lambda_2\mathcal{L}_{\text{vfl}} + \lambda_3\mathcal{L}_{\text{giou}} \\ &+ \lambda_4\mathcal{L}_{\text{ddf}} + \lambda_5\mathcal{L}_{\text{fgl}} + \lambda_6\mathcal{L}_{\text{focal}} + \lambda_7\mathcal{L}_{\text{dice}}.\end{aligned} \tag{9}$$

where the values of λ_1, λ_2, λ_3, λ_4, and λ_5 follow the settings in D-FINEPeng et al. (2025). In addition, different values of λ_6 and λ_7 are set to systematically evaluate the impact of the newly added loss terms on the final mask test results.

3 Experiments

We conduct experiments on the COCO benchmark and a Traditional Chinese Medicine microscopic cell dataset (TCM) to validate RT-ISTR. Section 3.1 describes the experimental setup, Section 3.2 reports the main results, and Section 3.3 presents ablation studies.

3.1 Experiment setting

Built upon D-FINE, the MME module and the newly added segmentation head are employed to realize a complete instance segmentation method [1]. With AP^{box} and AP^{mask} adopted as the primary evaluation metrics, our method is compared against multiple state-of-the-art models. Detailed experimental parameters are provided in Appendix B. To evaluate the effectiveness of the proposed method, experiments were conducted on the COCO dataset Lin et al. (2014) and a self-built Traditional Chinese Medicine microscopic cell instance segmentation dataset (TCM dataset).

[1] Code will be publicly available after the paper is published.

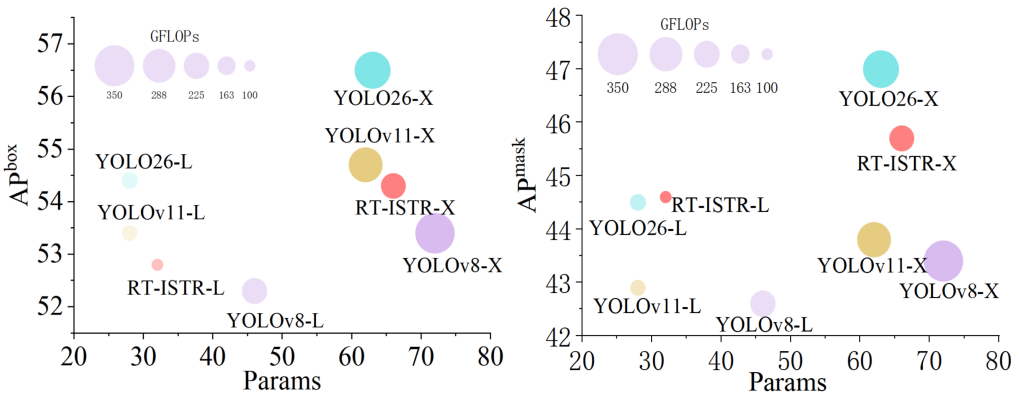

Figure 2: Real-time object detection and instance segmentation performance comparison of each model on the test split of the COCO dataset. The specific original data can be viewed in the Table 6.

3.2 Main Results

The performance of the proposed model is systematically evaluated on the COCO dataset and the TCM dataset.

Table 1: Comparison of target detection and instance segmentation performance of each model on test split of the COCO dataset. $*$ means Cascade Mask-RCNN. $\star$ represents the re measured results using the code provided by mmdetection Chen et al. (2019b). ▲ indicates that the report refers to Mask DINO.

Model	AP^{box}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{S}	AP^{mask}_{M}	AP^{mask}_{L}	GFLOPS(G)	Params(M)	FPS
MaskRCNN-R50▲	48.2	42.5	-	-	23.8	45.0	60.0	207	40	10
CMRCNN-R50 $*$	44.5	38.6	60.0	41.7	21.7	40.8	49.6	1787$\star$	77$\star$	10
HTC-R50	43.6	38.4	60.0	41.5	20.4	40.7	51.2	1735$\star$	80$\star$	3
Mask DINO-R50	49.8	45.4	67.9	49.3	25.2	48.3	65.8	280	52	15
D-Fine-L	54.0	-	-	-	-	-	-	91	31	125
RT-ISTR-L	52.8	44.6	67.3	48.0	21.9	48.5	66.5	102	32	23
CMRCNN-R101 $*$	45.7	39.8	61.6	43.0	22.4	42.2	50.5	1863$\star$	96$\star$	9
HTC-R101	45.3	39.7	61.8	43.1	21.0	42.2	53.5	1803$\star$	99$\star$	2
D-Fine-X	55.8	-	-	-	-	-	-	202	62	78
RT-ISTR-X	54.3	45.7	68.9	49.0	22.1	50.0	68.5	217	66	21

3.2.1 Comparison on the COCO Dataset

We evaluate RT-ISTR against state-of-the-art methods on the COCO dataset. As shown in Table 1, RT-ISTR-L achieves competitive performance under substantially lower computational budgets. It attains 52.8% AP^{box} and 44.6% AP^{mask}, surpassing MaskRCNN, Cascade Mask R-CNN and HTC by large margins (*e.g.,* $+2.1\%$ AP^{mask} over MaskRCNN). Notably, RT-ISTR-L requires only 102 GFLOPs and 32M parameters, far fewer than Mask DINO (280 GFLOPs, 52M), and runs at 23 FPS, demonstrating a superior accuracy-efficiency trade-off. This validates our core innovation: high-quality instance segmentation is achievable without heavy parameter counts, challenging the prevailing assumption in the field that high accuracy inherently requires heavy models.

For object scale analysis, RT-ISTR-L excels on medium and large objects ($AP^{\text{mask}}_{M} = 48.5\%$, $AP^{\text{mask}}_{L} = 66.5\%$), yet underperforms on small objects. We attribute this to the MME module's reliance on high-level semantics, where shallow details are fused only via element-wise addition and may be overridden. This limitation suggests a future direction for finer-grained feature preservation. Among real-time methods (Figure 2), RT-ISTR-L achieves the highest AP^{mask} (44.6%) among its peers, surpassing YOLOv8-L (42.6%), YOLOv11-L (42.9%), and YOLO26-L (44.5%), with competitive GFLOPs and FPS. The larger variant RT-ISTR-X attains 54.3% AP^{box} and 45.7% AP^{mask}, with significantly lower complexity than YOLO26-X and Mask-RT-DETR-X. These results confirm

that RT-ISTR delivers real-time, high-accuracy instance segmentation with minimal resource overhead, a critical advantage for deployment in resource-constrained microscopy environments.

Table 2: Comparison of target detection and instance segmentation performance of each model on the test split of the TCM Dataset.

model	AP^{box} ↑	AP^{box}_{50} ↑	AP^{box}_{75} ↑	AP^{mask} ↑	AP^{mask}_{50} ↑	AP^{mask}_{75} ↑	GFLOPS ↓	Params(M)↓
Yolact-R50	43.5	67.6	51.4	47.6	68.5	52.5	68	35
MaskRCNN-R50	60.1	88.5	70.5	67.2	89.5	76.5	260	44
Yolact-R101	68.5	94.9	82.0	72.2	95.2	82.0	92	54
MaskRCNN-R101	65.6	95.8	73.4	71.3	94.8	82.4	336	63
ConvNeXt	75.8	97.6	86.7	78.4	97.6	88.9	262	48
YOLOv8-L	84.0	96.6	91.9	80.3	96.8	90.6	221	46
YOLOv8-X	82.9	97.6	92.0	79.6	97.6	91.2	329	72
YOLOv11-L	81.5	97.3	91.5	77.9	97.3	89.5	132	28
YOLOv11-X	81.0	96.5	87.8	77.3	96.5	86.9	296	62
YOLO26-L	78.0	95.2	86.7	77.3	94.8	87.6	140	28
YOLO26-X	73.6	92.3	83.5	72.4	91.7	84.2	313	63
D-Fine-L	85.4	98.2	93.7	-	-	-	91	31
D-Fine-X	83.7	96.8	92.4	-	-	-	202	62
RT-ISTR-L	86.3	99.0	93.8	82.1	99.1	94.1	102	32
RT-ISTR-X	86.3	97.9	93.2	81.5	97.9	90.4	216	66

Table 3: Performance comparison of models with different scales on the TCM Dataset.

model	AP^{box} ↑	AP^{box}_{50} ↑	AP^{box}_{75} ↑	AP^{mask} ↑	AP^{mask}_{50} ↑	AP^{mask}_{75} ↑	GFLOPS	Params(M)
RT-ISTR-N	82.4	96.0	89.4	78.4	96.3	88.7	11	5
RT-ISTR-S	85.4	98.3	91.8	81.8	97.9	92.8	34	11
RT-ISTR-M	86.2	98.4	92.6	81.5	98.5	90.2	67	20
RT-ISTR-L	86.3	99.0	93.8	82.1	99.1	94.1	102	32
RT-ISTR-X	86.3	97.9	93.2	81.5	97.9	90.4	216	66

3.2.2 Comparison on the TCM Dataset

Table 2 compares object detection and instance segmentation performance on the TCM microscopic cell dataset. RT-ISTR achieves an AP^{box} of 86.3%, substantially outperforming YOLACT (68.5%), Mask R-CNN (65.6%), and ConvNeXt-t (75.8%), while slightly surpassing YOLOv8l-seg (84.0%). Under the stricter AP^{box}_{75} metric, RT-ISTR reaches 93.8%, again leading all competitors, demonstrating its superior localization accuracy. Notably, although YOLOv11 and YOLO26 excel on COCO, they underperform on this microscopic dataset compared to YOLOv8.

For instance segmentation, RT-ISTR attains an AP^{mask} of 82.1%, exceeding YOLOv8l-seg (80.3%) and YOLOv8x-seg (79.6%), and significantly surpassing traditional methods like Mask R-CNN and YOLACT. At the high IoU threshold of 0.75, RT-ISTR achieves 92.2% AP^{mask}_{75}, confirming its ability to capture fine-grained cellular boundaries. We further evaluate five model scales (N to X) from D-FINE (Table 3). The smallest variant (N) uses only 5M parameters and 11 GFLOPs yet delivers 82.4 AP^{box} and 78.4 AP^{mask}. Accuracy consistently improves with model size, demonstrating the scalability and lightweight deployment flexibility of RT-ISTR across diverse hardware configurations.

3.3 Ablation Experiments

According to the ablation results in Figure 3(a), after introducing the hierarchical fusion module, RT-ISTR achieves significant improvements in mask segmentation tasks. Specifically, the addition of hierarchical fusion increases AP^{mask} from 34.9 to 82.1, indicating that hierarchical fusion effectively integrates multi-scale features, particularly markedly improving mask prediction quality.

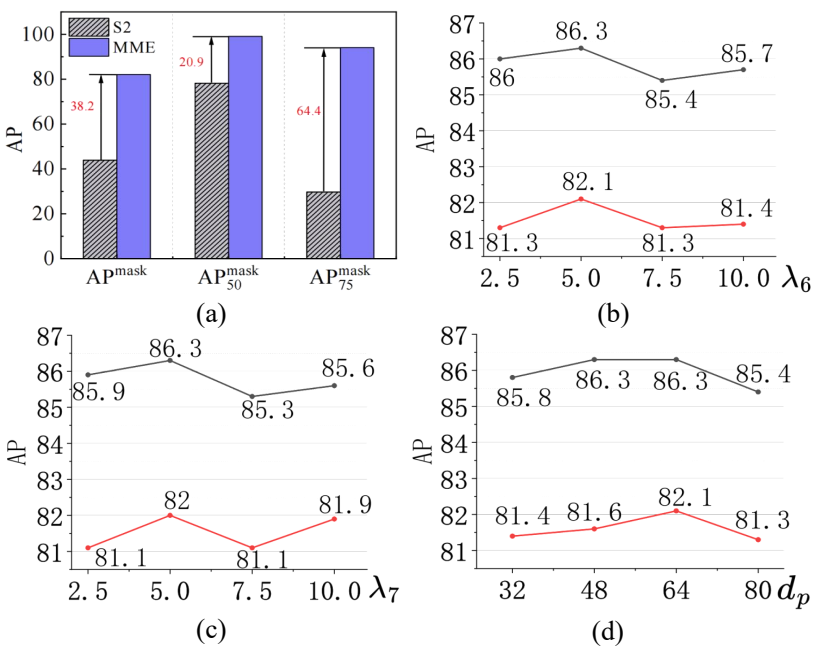

Figure 3: Ablation studies of the proposed modules and hyperparameters on TCM dataset. (a) Ablation Experiment on MME module; (b) Influence of λ_6 in $\mathcal{L}_{\text{focal}}$; (c) Influence of λ_7 in $\mathcal{L}_{\text{dice}}$; (d) Effect of prototype mask dimension d_p.

Furthermore, AP_{75}^{mask} increases from 29.7 to 94.1, further validating the advantage of this method in fine boundary segmentation.

Figures 3(b) and 3(c) present the effects of the loss weights λ_6 and λ_7 on the performance of traditional Chinese medicine cell detection. Based on the two sets of experiments, the optimal weight for both $\mathcal{L}_{\text{focal}}$ and $\mathcal{L}_{\text{dice}}$ is 5, at which the model achieves balanced and optimal results in both box detection and mask segmentation tasks.

Figure 3(d) shows the influence of the dimension of the prototype mask feature map output by MME module on the experimental results. According to the table, the best results are obtained when the prototype mask dimension is 64. This suggests that if the output mask dimension is too low, the mask feature map cannot contain sufficient basic information of the image, leading to loss of some image features. Conversely, if the output mask dimension is too high, the information contained in the mask feature map becomes too dispersed, and the attention capacity of the mask query embeddings is limited, thereby reducing the efficiency and accuracy of instance object queries.

4 Conclusion

This paper presents RT-ISTR, an end-to-end Transformer-based framework for real-time object detection and instance segmentation. The core innovation lies in the lightweight Multi-scale Mask Enhancement Module (MMEM), which efficiently fuses high-resolution spatial features from the backbone with encoder-enhanced semantics to produce high-quality prototype masks, and the reuse of detection query embeddings for dynamic mask querying at minimal extra cost. Extensive experiments on COCO and a TCM microscopic cell dataset demonstrate that RT-ISTR achieves competitive or superior segmentation accuracy (e.g., 44.6% AP^{mask} on COCO and 82.1% on TCM) with substantially lower computational overhead (102 GFLOPs and 32M parameters for the L variant) and real-time inference speed (64 FPS), outperforming both heavy CNN-based and Transformer-based counterparts. Ablation studies validate the effectiveness of each component.

REFERENCES

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9157–9166, 2019.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229, 2020.

Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4974–4983, 2019a.

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, 2019b.

Yuqian Chen, Yang Song, Chaoyi Zhang, Fan Zhang, Lauren O'Donnell, Wojciech Chrzanowski, and Weidong Cai. Celltrack R-CNN: A novel end-to-end deep neural network for cell segmentation and tracking in microscopy images. In *2021 IEEE 18th international symposium on biomedical imaging*, pp. 779–782, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yifan Feng, Jiangang Huang, Shaoyi Du, Shihui Ying, Jun-Hai Yong, Yipeng Li, Guiguang Ding, Rongrong Ji, and Yue Gao. Hyper-YOLO: When visual object detection meets hypergraph computation. *IEEE transactions on pattern analysis and machine intelligence*, 47(4):2388–2401, 2024.

Daniel Franco-Barranco, Zudi Lin, Won-Dong Jang, Xueying Wang, Qijia Shen, Wenjie Yin, Yutian Fan, Mingxing Li, Chang Chen, Zhiwei Xiong, et al. Current progress and challenges in large-scale 3d mitochondria instance segmentation. *IEEE transactions on medical imaging*, 42(12): 3956–3971, 2023.

Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Martin Weigert, Uwe Schmidt, Wenhua Zhang, Jun Zhang, Sen Yang, Jinxi Xiang, Xiyue Wang, et al. CoNIC challenge: Pushing the frontiers of nuclear detection, segmentation, classification and counting. *Medical image analysis*, 92:103047–103061, 2024.

Anubha Gupta, Shiv Gehlot, Shubham Goswami, Sachin Motwani, Ritu Gupta, Álvaro García Faura, Dejan Štepec, Tomaž Martinčič, Reza Azad, Dorit Merhof, et al. SEGPC-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images. *Medical image analysis*, 83:102677, 2023.

Shuaihao Han, Tingfa Xu, Peifu Liu, and Jianan Li. MODA: The first challenging benchmark for multispectral object detection in aerial images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pp. 4574–4582, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. ISTR: End-to-end instance segmentation with transformers, 2021.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring R-CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6409–6418, 2019.

Glenn Jocher and Jing Qiu. Ultralytics YOLOv11. `https://github.com/ultralytics/ultralytics`, 2024. Accessed: 2024.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. `https://github.com/ultralytics/ultralytics`, 2023. Accessed: 2023.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified real-time end-to-end vision models, 2026.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3041–3050, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pp. 740–755, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, 2016.

Jinquan Mo, Zizhen Peng, Wanying Li, Chengfeng Wei, Guangyao Pang, and Wei Fu. Microscopic detection approach of chinese herbal medicine base on YOLOv8. In *2024 International conference on ubiquitous computing and communications*, pp. 226–230, 2024.

Authors PaddlePaddle. PaddleDetection, object detection and instance segmentation toolkit based on paddlepaddle. `https://github.com/PaddlePaddle/PaddleDetection`, 2019. Accessed: 2019.

Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement. In *International conference on learning representations*, pp. 1–17, 2025.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. RF-DETR: neural architecture search for real-time detection transformers. In *Proceedings of the conference on International Conference on Learning Representations*, pp. 1–24, 2026.

Kaushiki Roy, Debapriya Banik, Gordon K Chan, Ondrej Krejcar, and Debotosh Bhattacharjee. 2pclpr: A two-phase clump profiler for segmentation of cancer cells in fluorescence microscopic images. *IEEE Transactions on instrumentation and measurement*, 72:1–14, 2023.

Jeongun Ryu, Aaron Valero Puche, JaeWoong Shin, Seonwook Park, Biagio Brattoli, Jinhee Lee, Wonkyung Jung, Soo Ick Cho, Kyunghyun Paeng, Chan-Young Ock, et al. OCELOT: Overlapped cell on tissue dataset for histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23902–23912, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3, 2025.

Ching-Wei Wang, Kun-Lin Lin, Hikam Muzakky, Yi-Jia Lin, and Tai-Kuang Chao. Weakly supervised bilayer convolutional network in segmentation of her2 related cells to guide her2 targeted therapies. *Computerized medical imaging and graphics*, pp. 102270–102286, 2023.

Chengfeng Wei, Zizhen Peng, Linhe Yang, Guiguang Li, and Yincheng Kong. A microscopic image object detection method based on deformable convolution and multi-scale convolution attention. In *2025 IEEE International conference on high performance computing and communications*, pp. 1–6, 2025.

Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. UnitBox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520, 2016.

Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An IOU-Aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8514–8523, 2021.

Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2582–2593, 2022.

Yanning Zhou, Hao Chen, Huangjing Lin, and Pheng-Ann Heng. Deep semi-supervised knowledge distillation for overlapping cervical cell instance segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 521–531, 2020.

A Appendix

B Experimental details

B.1 Implementation details

Table 4 presents the hyperparameter configurations of our models. This section details the experimental setups for the RT-ISTR variants of different scales (N, S, M, L, and X). All models adopt the HGNetv2 series as the backbone network, with specific versions B0, B0, B2, B4, and B5, respectively. The backbone return indices vary across models: RT-ISTR-N returns stages 1, 2, and 3, whereas all other variants return stages 0, 1, 2, and 3. The number of decoder layers ranges from 3 to 6: N and S use 3 layers, M uses 4, L uses 6, and X uses 4. Regarding the hidden dimension, N is set to 128, S, M, and L to 256, and X to 384. The number of training epochs decreases with model size: N is trained for 160 epochs, S and M for 132 epochs, and L and X for 80 epochs. During training on the COCO dataset, the batch size is adjusted according to GPU memory capacity: N uses 32, S and M use 16, and L and X use 8. All models are trained on NVIDIA A40 GPUs; specifically, RT-ISTR-M, RT-ISTR-L, and RT-ISTR-X are trained on two GPUs, whereas all other models are trained on a single GPU. On the TCM dataset, the batch size is uniformly set to 8 for all models, and all are trained on a single NVIDIA A40 GPU. These configurations reflect the trade-offs between computational resources and model scales.

Table 4: Hyperparameter configurations for different RT-ISTR models.

Setting	RT-ISTR-N	RT-ISTR-S	RT-ISTR-M	RT-ISTR-L	RT-ISTR-X
Backbone Name	HGNetv2-B0	HGNetv2-B0	HGNetv2-B2	HGNetv2-B4	HGNetv2-B5
Backbone Return idx	(1, 2, 3)	(0, 1, 2, 3)	(0, 1, 2, 3)	(0, 1, 2, 3)	(0, 1, 2, 3)
Decoder Layers	3	3	4	6	4
Hidden Dim	128	256	256	256	384
Epoch	160	132	132	80	80
COCO Batch	32	16	16	16	8
COCO GPU Type	NVIDIA A40	NVIDIA A40	NVIDIA A40 × 2	NVIDIA A40 × 2	NVIDIA A40 × 2
TCM Batch	8	8	8	8	8
TCM GPU Type	NVIDIA A40	NVIDIA A40	NVIDIA A40	NVIDIA A40	NVIDIA A40

B.2 Evaluation Metrics.

To evaluate the performance of the proposed method, the Average Precision (AP) metric is adopted. Specifically, the primary focus is on the AP computed across different IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. To distinguish between the average precision of object detection and instance segmentation, AP^{box} denotes the average precision of the model for the bounding box regression task, and AP^{mask} denotes that for the instance segmentation task. Additionally, other metrics include AP^{box}_{50} and AP^{mask}_{50}, which represent the average precision measured at an IoU threshold of 0.5 to evaluate performance under looser matching conditions, as well as AP^{box}_{75} and AP^{mask}_{75}, which represent the average precision evaluated at an IoU threshold of 0.75 to reflect performance under stricter matching conditions. Finally, AP^{mask}_{S}, AP^{mask}_{M}, and AP^{mask}_{L} are also employed to evaluate the segmentation performance on small, medium, and large objects in the COCO dataset. Furthermore, the number of parameters, computational complexity, and detection speed (FPS) of various models are reported to enable a more systematic comparison of model performance.

B.3 Comparison Methods.

We carried out numerous experiments for comparison with other state-of-the-art models. On the COCO dataset, the proposed method is compared with Mask RCNN He et al. (2017), Cascade Mask-RCNN Cai & Vasconcelos (2019), HTC Chen et al. (2019a), Mask DINO Li et al. (2023), RF-DETRRobinson et al. (2026), and YOLO series methods Jocher et al. (2023); Jocher & Qiu

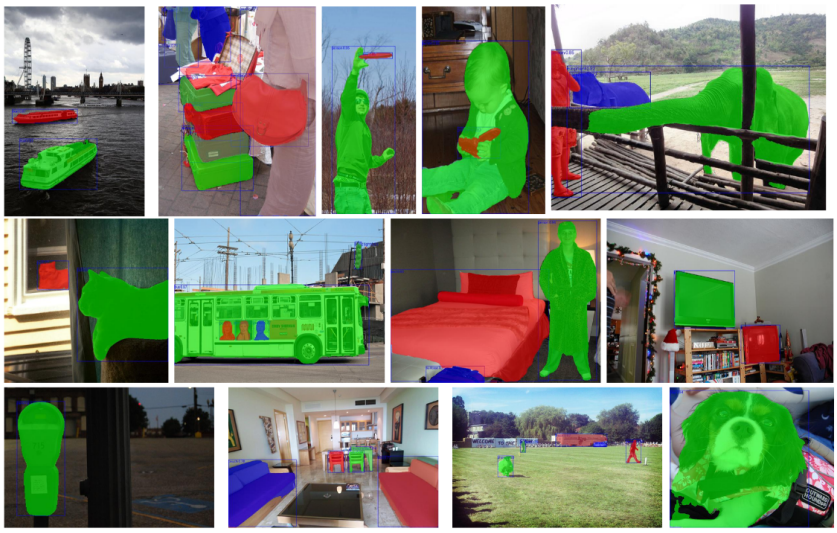

Figure 4: Visualization of detection results of the model on the COCO dataset.

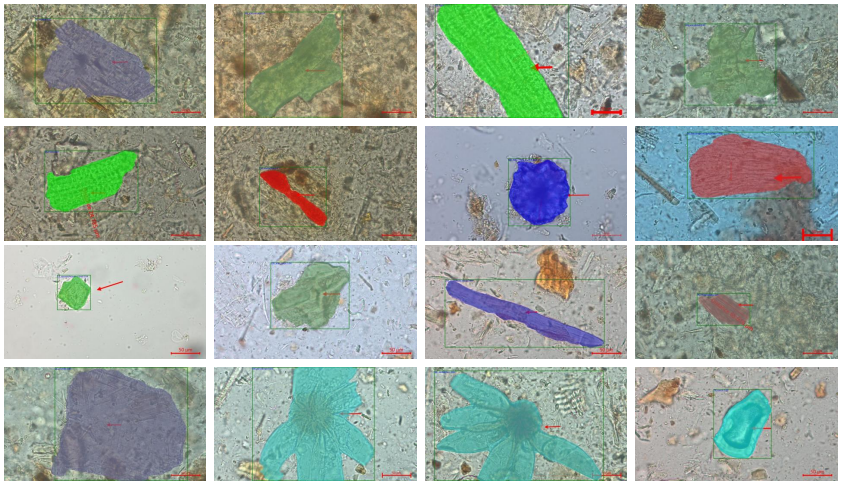

Figure 5: Visualization of detection results of the model on the TCM dataset.

(2024); Jocher et al. (2026); Feng et al. (2024). The results of these methods are referenced from the respective literature. On the self-built traditional Chinese medicine microscopic cell dataset, YOLACT Bolya et al. (2019), Mask RCNN, ConvNeXt Liu et al. (2022), and YOLO seriesJocher et al. (2023); Jocher & Qiu (2024); Jocher et al. (2026) methods are re-implemented for comparison.

B.4 Supplementary Experiment

Table 5 reports the experimental results of our three model variants (N, S, and M) on the COCO2017 instance segmentation dataset. The primary baseline for comparison is the YOLO series. All our models were trained and tested following the experimental settings detailed in Table 4. For the YOLO models, since the official reports do not provide AP_{50} and AP_{75} values, we downloaded the publicly released model parameters and evaluated them on the COCO test set using an NVIDIA A40 GPU with a batch size of 16.

Among the lightweight N-scale models, RT-ISTR-N achieves the highest detection and segmentation accuracy, with a bounding box average precision (AP^{box}) of 40.3%, surpassing YOLOv8-N,

Table 5: Performance comparison of different scale models on coco dataset. Among them, the result of the Yolo model is the reasoning result of the officially published model parameters on the NVIDIA A40 GPU with batch 8 in the test set.

Model	$AP^{box}\uparrow$	$AP^{box}_{50}\uparrow$	$AP^{box}_{75}\uparrow$	$AP^{mask}\uparrow$	$AP^{mask}_{50}\uparrow$	$AP^{mask}_{75}\uparrow$	GFLOPS	Params(M)
YOLOv8-N	36.2	51.5	39.3	30.2	48.8	31.8	12.0	3.4
YOLO11-N	38.3	53.7	41.7	31.9	50.9	33.8	9.7	2.8
YOLO26-N	39.2	54.9	42.5	32.9	52.3	34.9	9.1	2.7
RT-ISTR-N	40.3	56.3	43.3	34.8	54.1	36.7	10.8	4.7
YOLOv8-S	44.1	60.6	48.1	36.7	57.7	39.2	40.1	11.8
YOLO11-S	46.1	62.6	50.0	38.0	59.6	40.6	33.0	10.1
YOLO26-S	46.8	63.7	50.9	38.6	60.7	41.3	34.2	10.4
RT-ISTR-S	45.7	61.4	50.0	38.8	59.2	41.6	34.4	11.2
YOLOv8-M	49.4	66.0	54.2	40.7	62.9	43.8	104.5	27.3
YOLO11-M	51.0	67.8	55.2	41.9	64.8	45.0	113.2	22.4
YOLO26-M	51.7	68.9	56.9	42.1	65.5	45.6	121.5	23.6
RF-DETR-Seg N	-	-	-	40.3	63.0	42.6	50.0	33.6
RF-DETR-Seg S	-	-	-	43.1	66.2	45.9	70.6	33.7
RT-ISTR-M	51.2	67.6	55.7	43.2	65.3	46.4	66.8	20.5

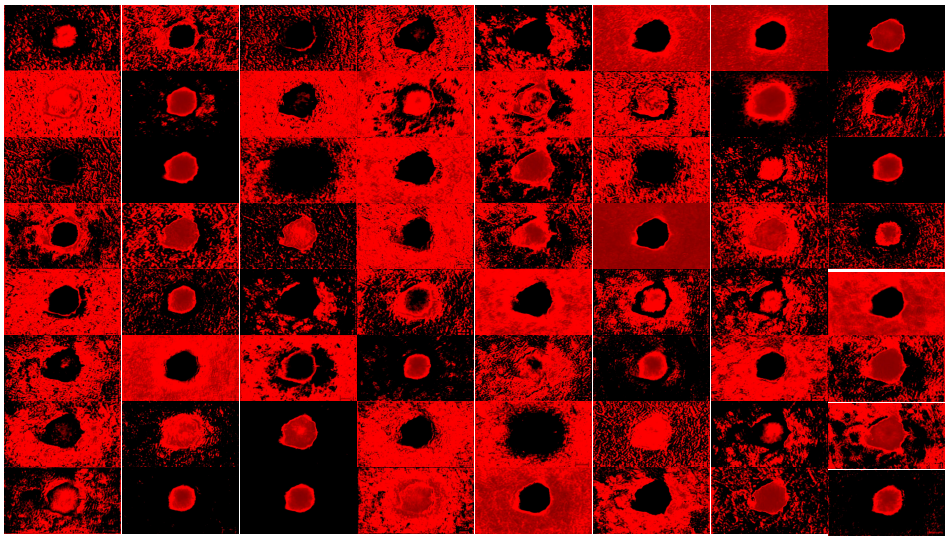

Figure 6: High-response results of the output prototype masks of the model on a single dimension.

Table 6: Real-time object detection and instance segmentation performance comparison of each model on the test split of the COCO dataset.

model	AP^{box}	AP^{mask}	GFLOPS	Params
YOLOv8-L	52.3	42.6	221	46
YOLOv11-L	53.4	42.9	132	28
YOLO26-L	54.4	44.5	140	28
RT-ISTR-L	52.8	44.6	102	32
YOLOv8-X	53.4	43.4	344	72
YOLOv11-X	54.7	43.8	296	62
YOLO26-X	56.5	47.0	314	63
RT-ISTR-X	54.3	45.7	217	66

YOLO11-N, and YOLO26-N by 4.1, 2.0, and 1.1 percentage points, respectively. Under the more stringent AP^{box}_{75} metric, RT-ISTR-N (43.3%) also outperforms competing models such as YOLO26-N (42.5%). For the segmentation task, RT-ISTR-N attains an AP^{mask} of 34.8%, which is 1.9 percentage points higher than that of YOLO26-N (32.9%). Although RT-ISTR-N has a larger parameter count (4.7M) than YOLO26-N (2.7M), its computational cost (10.8 GFLOPs) lies between those of YOLO11-N (9.7) and YOLOv8-N (12.0), thereby striking a favorable balance between accuracy and efficiency overall.

In the S-scale models, RT-ISTR-S achieves an AP^{box} of 46.6%, slightly lower than that of YOLO26-S (46.8%) but higher than those of YOLOv8-S (44.1%) and YOLO11-S (46.1%). However, RT-ISTR-S exhibits a clear advantage in mask segmentation: its AP^{mask} reaches 39.5%, surpassing YOLOv8-S (36.7%), YOLO11-S (38.0%), and YOLO26-S (38.6%), and it also leads all compared S-scale models in AP^{mask}_{75} (42.4%). The computational cost of RT-ISTR-S is 34.4 GFLOPs, close to that of YOLO26-S (34.2), while its parameter count (11.2M) is comparable to that of YOLOv8-S (11.8M).

For the M-scale models, YOLO26-M achieves the best detection accuracy with a box AP of 51.7%, demonstrating the strong scalability of its backbone network with a large parameter budget. Nevertheless, RT-ISTR-M stands out in segmentation (mask AP of 43.2%) and exhibits a substantial advantage in computational efficiency: its GFLOPs (66.8) are only 55.0% of those of YOLO26-M (121.5), and its parameter count (20.5M) is the smallest among the same group.

C VISUALIZATION OF RT-ISTR PREDICTION

Figure 4 and Figure 5 present the performance of the proposed model on the COCO dataset and the microscopic cell dataset, respectively. As shown in Figure 4, our model accurately segments instances in complex scenes of the COCO dataset. Figure 5 also displays the segmentation effects for other categories of microscopic cells. As can be observed, our model achieves high-precision segmentation in complex microscopic scenarios.

Figure 3(d)quantitatively illustrates the impact of the number of mask dimensions on mask segmentation accuracy. In Figure 6, the high-response regions of 64 feature dimensions are visualized. It can be observed that the high-response regions of the prototype masks generated by the model cover the basic shapes of the target objects, preserving the shape characteristic information of the targets, which is consistent with the analysis in Figure 3(d).