# A Two-Stage Strategy for Mitosis Detection Using Improved YOLO11x Proposals and ConvNeXt Classification

**Jie Xiao, Mengye Lyu, and Shaojun Liu**

College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen, 518118, China

**MIDOG 2025 Track 1 requires mitosis detection in whole-slide images (WSIs) containing non-tumor, inflamed, and necrotic regions. Due to the complicated and heterogeneous context, as well as possible artifacts, there are often false positives and false negatives, thus degrading the detection F1-score. To address this problem, we propose a two-stage framework. Firstly, an improved YOLO11x, integrated with EMA attention and LSConv, is employed to generate mitosis candidates. We use a low confidence threshold to generate as many proposals as possible, ensuring the detection recall. Then, a ConvNeXt-Tiny classifier is employed to filter out the false positives, ensuring the detection precision. Consequently, the proposed two-stage framework can generate a high detection F1-score. Evaluated on a fused dataset comprising MIDOG++, MITOS_WSI_CCMCT, and MITOS_WSI_CMC, our framework achieves an F1-score of 0.882, which is 0.035 higher than the single-stage YOLO11x baseline. This performance gain is produced by a significant precision improvement, from 0.762 to 0.839, and a comparable recall. The code is available at https://github.com/xxiao0304/MIDOG-2025-Track-1-of-SZTU.**

**Mitosis Detection | Two-Stage Framework | YOLO11x | ConvNeXt | MIDOG 2025**

**Correspondence:** ***liusj14@tsinghua.org.cn***

## Introduction

Mitotic figure counting is critical for tumor grading in clinical pathology. However, manual annotation of mitosis is time-consuming due to the extra-large size of whole slide images (WSIs). Besides, the morphology of mitotic figures strongly overlaps with similar-looking impostors, leading to annotation variability among pathologists [1]. MIDOG 2025 Track 1 focuses on mitosis detection on *full WSIs* instead of localized regions of interest, introducing three key technical hurdles:

- Small mitotic figures, typically 10–30 pixels, are easily missed in low-resolution WSI patches;
- Necrotic debris and inflamed cells mimic the morphological features of mitoses, leading to high false positives;
- The dataset includes 12 tumor types of human and veterinary specimens. Such domain shifts might reduce the generalization of feature-based models.

Considering the large size of WSIs, the model should be efficient. Single-stage detectors such as YOLO11x [2] prioritize inference efficiency; therefore, they are suitable for this task. However, they usually struggle with false positive suppression in hard regions, including necrotic zones. Indeed, we can decrease the false positives by using a lower confidence threshold. However, this would lead to more false negatives, consequently degrading the detection F1-score.

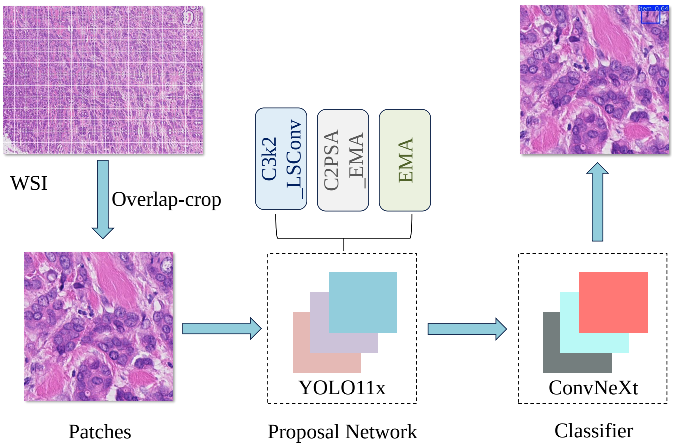

**Fig. 1.** Two-Stage Mitosis Detection Framework Workflow. The pipeline includes patch cropping, candidate generation with an improved YOLO11x [2] as proposal network, and false positive filtering with a ConvNeXt [3] as classification network.

To tackle this problem, we design a two-stage pipeline comprising an improved YOLO11x and a classification network. Specifically, a *recall-oriented improved YOLO11x proposal network*, with a low confidence threshold, is employed to capture all potential mitoses first; and the candidates are filtered with a *precision-oriented ConvNeXt classifier* [3] to reduce false positives. This design can keep the high inference efficiency of YOLO11x while improving the detection F1-score. This synergy between YOLO11x's efficiency and ConvNeXt's robustness directly addresses the unique challenges of MIDOG 2025 Track 1.

## Material and Methods

**1. Datasets and Preprocessing.** Three publicly available data sets are used for this study, with data divided into training, validation, and test sets in a ratio of **7:1:2**. All data usage complies with the MIDOG 2025 Track 1 guidelines [1]. Detailed information about each dataset is as follows:

- **MIDOG++** contains 553 patches sized 7,200×5,400,

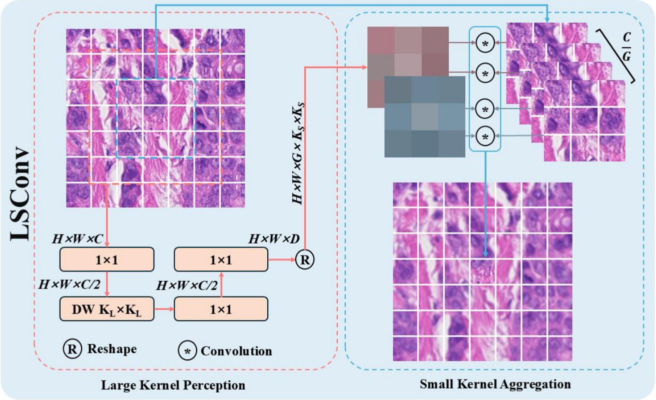

**Fig. 2.** Network Architecture of LSConv [7]. LSConv combines large-kernel perception and small-kernel aggregation. It simulates the dynamic multi-scale visual capability of the human visual system. Large-kernel perception utilizes large-kernel depth-wise convolution to capture extensive contextual information, while small-kernel aggregation performs fine-grained aggregation of features within a small range through dynamic convolution.

from 503 tumor cases spanning 7 tumor types, with 11,937 manually annotated mitotic figures [1].

- **MITOS_WSI_CMC** includes 21 WSIs of canine breast cancer with 13,907 mitotic annotations [4].
- **MITOS_WSI_CCMCT** comprises 32 WSIs of canine mast cell tumors with 44,880 annotations, featuring abundant necrotic and inflamed regions [5].

Since either the patches in MIDOG++ or the WSIs in the rest two datasets are too large for the model to process, we crop them into 512×512 local patches with 20% overlap to preserve the integrity of mitotic figures, avoiding partial mitoses at the patch edges.

To ensure reliable and unbiased evaluation, data augmentation is applied exclusively to the training set; no augmentation is performed on the validation or test sets. The augmentations include:

- Geometric transformations: random rotation in the range of $[0°, 180°]$; random horizontal and vertical flip with a probability of 0.5.
- Mixing strategies: random mixup with another local patch, with a probability of 0.3 and a random weight in the range of $[0, 1]$; mosaic with another 3 random local patches, which is only utilized in the first 20 epochs to stabilize late-stage training.
- Advanced augmentation: RandAugment [6] with default settings; random erasing with a probability of 0.4 and an erase ratio in the range of $[0.02, 0.1]$.

**2. Two-Stage Framework Design.** The framework is illustrated in Fig. 1. Firstly, the WSIs are cropped into local patches, and an improved YOLO11x proposal network is performed under a low confidence threshold to generate mitosis candidates, reducing false negatives. Then, the candidates are filtered by a ConvNeXt classifier network to reduce false positives. Therefore, such a two-stage design can ensure both recall and precision, leading to a high F1-score. The details are described as follows.

***Stage 1: Proposal Network (YOLO11x + LSConv + EMA).*** The base model architecture follows YOLO11x design principles [2]. To enhance small-target recall while reducing model complexity, the original YOLO11x is improved with three key modules:

- **C3k2_LSConv Blocks.** The C3k2 blocks in the P3, P4, and P5 detection heads are replaced with C3k2_LSConv blocks. Specifically, the convolution layers in the C3K2 block are replaced with the LSConv layer [7], whose core architecture is illustrated in Fig. 2. Each LSConv integrates a 7×7 depth-wise convolution to enlarge the receptive field for small targets and a 3×3 dynamic grouped convolution to enhance feature discrimination. This design can improve efficiency while maintaining performance.
- **C2PSA_EMA Module.** It is deployed at the fusion interface between the backbone and neck. The input is divided into several groups and then sequentially fed into a 1×1 convolution, and $n$ PSABloc_EMA units, where each unit incorporates a feed-forward network and EMA attention [8]. Finally, the groups are combined together to restore the original channel dimension.
- **EMA Attention.** To further refine feature representation at the detection stage, this cross-spatial attention mechanism [8] is additionally attached to each detection head, *i.e.*, P3, P4, and P5. It groups channels by a factor of 32 to model dependencies, and through its channel-grouping strategy and cross-spatial interaction, effectively suppresses background noise in necrotic and inflamed regions.

As mentioned above, the proposal network needs to generate as many candidates as possible, aiming to avoid false negatives. Therefore, a low confidence threshold of 0.2 is employed to maximize the recall. Additionally, in the non-maximum suppression procedure, the intersection over union (IoU) threshold is set to 0.3 to reduce redundant candidates.

***Stage 2: ConvNeXt Classifier.*** Since the proposal network is performed under a low confidence threshold, there are many false positives. Therefore, the basic ConvNeXt-Tiny network [3] is employed to further classify the mitosis candidates to reject false positives.

As the candidates are usually of different sizes, they are resized to 64×64 pixels and normalized using ImageNet statistics. During training, the preprocessed proposals are augmented with the following strategies.

- Geometric/color transformations: random horizontal flip with a probability of 0.5; random rotation where the angle is in the range of $[-15°, +15°]$; ColorJitter where the jitter parameters for brightness, contrast, and saturation are 0.2, 0.2, and 0.1, respectively.;

- Mixing strategies: random mixup with another candidate, with a probability of 0.2 and a random weight in the range of $[0,1]$.
- Advanced augmentation: RandAugment [6] with 3 types of operation whose magnitude is 5; random half erasing with a probability of 0.5 and an erase ratio in the range of $[0.02, 0.15]$.

To train the classifier, we use a hybrid loss with focal loss to address the class imbalance issue, and contrastive loss to enhance feature separation between mitosis hard negatives. The loss function is described as (1).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{focal}} + \lambda \cdot \mathcal{L}_{\text{contrastive}}. \tag{1}$$

Here, $\lambda$ is a balancing weight between focal loss $\mathcal{L}_{\text{focal}}$, which is described in (2), and contrastive loss $\mathcal{L}_{\text{contrastive}}$, which is described in (3).

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N}\sum_{i=1}^{N} \alpha_{c_i} \cdot (1-p_{c_i})^{\gamma} \cdot \log(p_{c_i}), \tag{2}$$

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp\left(\text{sim}(f_i, f_i^{+})/T\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(f_i, f_j)/T\right)}\right), \tag{3}$$

where $N$ is the total number of samples in a batch; $\gamma$ is the focusing parameter to down-weight easily classified samples; $c_i$ is the class for $i$-th sample; $\alpha_{c_i}$ is the weight for that class; $p_{c_i}$ is the predicted probability of the $i$-th sample belonging to its true class $c_i$; sim$(\cdot,\cdot)$ is the cosine similarity; $f_i$ and $f_j$ are the feature vectors for the $i$-th and $j$-th samples; $f_i^{+}$ is the feature vector of a sample whose class is the same as sample $i$. $T$ is the temperature parameter adjusting the steepness of the similarity distribution.

## Training and Inference

**1. Training Details.** The training is conducted on a server equipped with 8 NVIDIA A100 80G GPUs. The training parameters are set as follows.

- **Proposal Network.** The loss function is defaultly set according to [2]. The network is optimized using the AdamW optimizer with cosine annealing, where the initial learning rate is $10^{-3}$ and the final learning rate is $10^{-5}$, and a weight decay of $5 \times 10^{-4}$. The model is trained for 300 epochs with a batch size of 960.
- **ConvNeXt-Tiny Classifier.** The parameters in the loss functions are empirically set as: $\alpha_m = 1, \alpha_b = 1.5, \gamma = 2.0, T = 0.2, \lambda = 1.0$. The network is optimized using the AdamW optimizer with cosine annealing, where the initial learning rate is $3 \times 10^{-4}$ and the final learning rate is $10^{-6}$, and a weight decay of $10^{-5}$. The model is trained for 400 epochs with an early stopping strategy monitoring on the validation F1-score, with a patience of 60. The batch size is 960.

**2. WSI Inference Pipeline.**

1. WSIs are split into 512×512 local patches with 20% overlap to ensure continuous coverage;
2. The proposal network generates candidate mitotic regions with confidence scores $\geq 0.2$;
3. Each candidate region is resized to 64×64 pixels, then fed into the ConvNeXt-Tiny classifier; candidates with classification scores $< 0.5$ are rejected;
4. The remaining candidates are merged across overlapping patches when IoU $\geq 0.5$, to generate the final detection results.

**3. Evaluation Metrics.** Aligned with the evaluation guidelines of MIDOG 2025 Track 1 [1], F1-score ($F1$), recall ($R$), and precision ($P$) are employed. Their definitions are:

$$F1 = \frac{2 * P * R}{P + R}, \tag{4}$$

$$R = \frac{TP}{TP + FN}, \tag{5}$$

$$P = \frac{TP}{TP + FP}, \tag{6}$$

where $TP$, $FP$, and $FN$ are the number of true positives, false positives, and false negatives, respectively.

## Results

**Quantitative Results.** We compare the proposed two-stage framework with two model variants: the basic YOLO11x and the improved YOLO11x. Both are single-stage variants. The quantitative results are summarized in Table 1.

**Table 1.** Quantitative results on the test set. The best results are in **bold**, and the second-best results are underlined.

| Model Variant | TP ↑ | FP ↓ | FN ↓ | P ↑ | R ↑ | F1 ↑ |
|---|---|---|---|---|---|---|
| Basic YOLO11x | **17879** | 7165 | **439** | 0.716 | **0.976** | 0.827 |
| Improved YOLO11x | 17441 | 5433 | 877 | 0.762 | 0.952 | 0.847 |
| Two-Stage (Ours) | 17030 | **3272** | 1288 | **0.839** | 0.929 | **0.882** |

***1. Basic YOLO11x.*** The basic YOLO11x achieves the highest recall of 0.976, with the fewest false negatives, demonstrating its strong capability in "broad mitotic capture", a critical requirement for full-WSI detection. However, it suffers from significantly more false positives, resulting in a precision of only 0.716. This is mainly caused by the misclassification of necrotic debris as mitoses, which limits its clinical utility. Though it has the highest recall, the F1-score is degraded significantly by the low precision.

***2. Improved YOLO11x.*** With the integration of EMA attention and LSConv, the improved YOLO11x model avoids false positives to some extent.
FP decreases to 5433, a 24% reduction with respect to the basic YOLO11x. Consequently, the precision improves to 0.762. At the same time, the recall decreases to 0.952. Despite this sacrifice, the F1-score still improves to 0.847, validating the rationality of the "fewer false positives over minor missed detection" design principle.

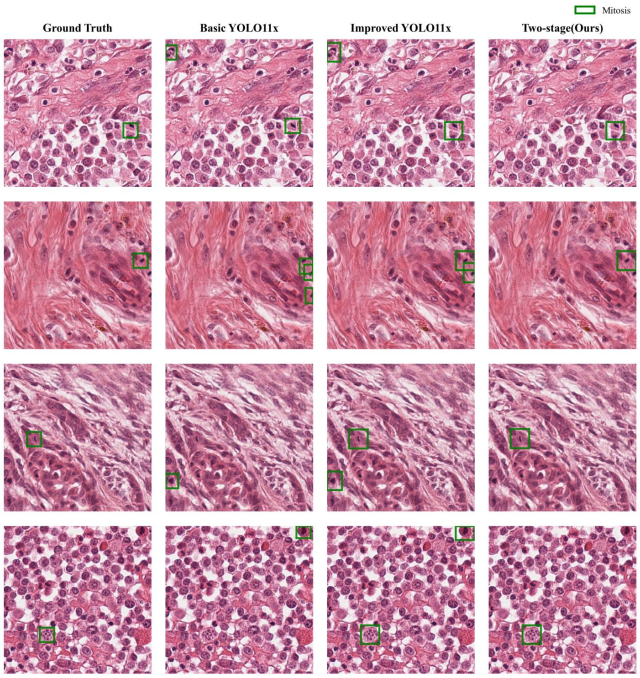

**Fig. 3.** Visual Results of Mitosis Detection. The figure shows a comparison of detection results among Ground Truth, Basic YOLO11x, Improved YOLO11x, and the proposed Two-stage framework. Green boxes indicate correctly detected mitosis regions. There are many false positives for basic YOLO11x; Improved YOLO11x can partially reduce the false positives; The proposed two-stage method can further reject hard false positives.

***3. Two-Stage Framework (Ours).*** Incorporating the ConvNeXt classifier with the improved YOLO11x further enhances precision while maintaining stable recall. FP is reduced to 3272, a 40% reduction with respect to the improved YOLO. Consequently, the precision jumps to 0.839. At the same time, the recall decreases to 0.929, which is still clinically viable for mitotic counting. Despite this sacrifice, the F1-score reaches 0.882, which is 0.055 higher than the basic YOLO11x, and 0.035 higher than the improved YOLO11x.

***Qualitative Results.*** To visually evaluate the performance of the proposed two-stage framework, we compare it with two single-stage model variants, *i.e.*, the basic YOLO11x and the improved YOLO11x. The results are shown in Fig. 3. It can be observed that the basic YOLO11x has more false positives compared to the improved YOLO11x, while the two-stage framework reduces false positives even more than the improved YOLO11x. This coincides with the quantitative results.

## Discussion and Conclusion

**Contribution.** The proposed two-stage “Improved YOLO11x + ConvNeXt” framework delivers two key advantages:

- **Targeted Design.** The “broad recall + precision filtering” workflow directly addresses the inefficiency of full-WSI detection, mitigating false positives in necrotic regions.
- **Superior Performance.** It achieves a higher F1-score, with 54% reduction in false positives with respect to the basic YOLO11x.

**Future Work.** Future research will focus on two directions: 1. Developing a dynamic threshold adjustment mechanism that adapts to WSI-specific features, such as the proportion of necrotic regions, staining intensity, and tumor type; 2. Integrating domain adaptation modules, *e.g.*, domain-adversarial training, to eliminate feature distribution gaps across datasets, further improving the framework’s generalization to diverse clinical scenarios.

## Bibliography

1. Jonas Ammeling, Marc Aubreville, and *et al.* Mitosis Domain Generalization Challenge 2025. In *Medical Image Computing and Computer Assisted Intervention 2025 (MICCAI)*. Zenodo, March 2025. doi: 10.5281/zenodo.15077361.
2. Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.
3. Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.
4. Marc Aubreville, Christof A. Bertram, Taryn A. Donovan, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Scientific Data*, 7(1):417, November 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00756-z.
5. Marc Aubreville, Frauke Wilm, , and *et al.* A comprehensive multi-domain dataset for mitotic figure detection. *Scientific Data*, 10(1):484, July 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02327-4.
6. Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019.
7. Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Lsnet: See large, focus small, 2025.
8. Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5. IEEE, June 2023. doi: 10.1109/icassp49357.2023.10096516.