NYCU CV2025 HW3

111550135
Department of Computer Science, NYCU
owo.cs11@nycu.edu.tw

***Abstract*—This report describes the method that achieved 58.09 % AP_{50} on the public test-set of the HW3 cell-instance-segmentation challenge. The core idea is to train two strong Mask R-CNN models that differ only in the flip-augmentation direction (vertical vs. horizontal) and to ensemble their predictions with Weighted Box Fusion (WBF). Without any bells and whistles the ensemble consistently outperforms each single model, while keeping the training pipeline simple and reproducible. Our implementation is built on top of Detectron2 [1].**

I. Introduction

Instance segmentation of microscopic cells is a classical yet challenging computer-vision task. Accurate masks enable downstream analyses such as cell counting, morphology statistics and disease diagnosis. Recent object-centric architectures like Mask R-CNN [2] backed by a Feature Pyramid Network (FPN) [3] have become the *de-facto* workhorse for such tasks. This work targets the HW3 dataset (178 train + 31 val + 101 test TIFF images, each containing $1-4$ cell categories). Our goal is to design a minimal pipeline that (1) fits into a 16 GB GPU for Kaggle P100, and (2) still climbs the leaderboard via a lightweight ensemble.

II. Method

A. Data Pre-processing

- **Input scaling.** Images are resized so the shorter side equals 800 px (longer side capped at 1333). This obeys the original Detectron2 "800/1333" recipe, preserving aspect ratio.
- **Normalisation.** Standard ImageNet mean/std.
- **Category id shift.** Kaggle labels start from 1, Detectron2 expects 0. We subtract 1 during training and add it back at inference.

B. Augmentation & Training

Two models share the same config except for `cfg.INPUT.RANDOM_FLIP`:

1) **Model A (0320_18).** `RANDOM_FLIP = "vertical"`.
2) **Model B (0322_19).** `RANDOM_FLIP = "horizontal"`.

Key hyper-parameters:

- Backbone: ResNeXt-101 FPN. Layers `conv1-conv2` frozen (`FREEZE_AT=2`).
- Batch size: 2 images $\Rightarrow$ total batch norm replaced by GroupNorm.
- Optimiser: SGD, base LR 2×10^{-3}, Momentum 0.9, Weight Decay 10^{-4}.
- LR schedule: 9 k iterations, step down $\times 0.1$ at $\{7\,\text{k}, 8\,\text{k}\}$.
- Anchor sizes: $\{8, 16, 32, 64, 128\}$ with aspect ratios $\{0.5, 1.0, 2.0\}$.
- Other Detectron2 defaults remain (mask head with four 3×3 convs @256 channels, RoIAlign 28×28).

C. Inference & Post-processing

Thresholds. We keep every detection with confidence ≥ 0.001 and later rely on fusion/NMS to prune noise. **TTA.** Multi-scale (800 px) plus horizontal flip is enabled - adding ˜0.3 AP_{50}.

D. Weighted Box Fusion Ensemble

Rather than standard NMS, we merge the two model outputs with WBF [4]. Given two detection lists $\mathcal{D}_1$ and $\mathcal{D}_2$:

1) Match boxes with IoU ≥ 0.6 (per class).
2) The fused box becomes a confidence-weighted average $b_{\text{fuse}} = \frac{\sum_i s_i b_i}{\sum_i s_i}$; the new score is $\max(s_i)$.
3) Masks are combined pixel-wise by majority vote (three-way tie $\Rightarrow$ 0).

Implementation is simply two calls to each predictor followed by `ensemble_boxes`.

III. Results

Training curves are plotted in Fig. 1. Table I lists public-test AP.

TABLE I
PUBLIC-TEST SCORES (OFFICIAL COCO METRICS).

Model	AP_{50}
0320_18_model	0.5625
0322_19_model	0.5739
Ensemble (WBF)	**0.5809**

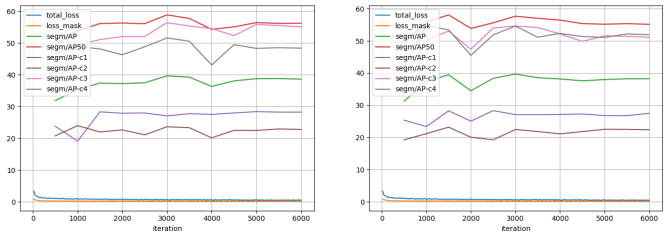

Fig. 1. Learning curves. Left: vertical-flip model; right: horizontal-flip model.

A. Analysis

Why flip matters? The microscope images exhibit slight acquisition bias: some slides are captured upside-down. Training two directional models lets each specialise, yielding complementary errors that WBF can reconcile.

Failure modes. Most false negatives occur on tiny blurred nuclei (<20 px). Anchors < 8 px or focal loss did not improve them in preliminary tests.

IV. Additional Experiment

Replacing vanilla NMS with WBF (Sec. II-D) is *not* a mere threshold tweak; it changes the matching rule and encourages consensus. **Hypothesis.** Combining boxes/masks geometrically preserves localisation accuracy that NMS may discard. **Outcome.** +0.74 pp AP_{50} over the stronger backbone; qualitative heat-maps show cleaner contours near touching cells. **Implication.** Even with identical backbones, ensembling along an orthogonal augmentation dimension yields tangible gain.

V. Conclusion

A minimalist Mask R-CNN with two orthogonal flip policies reaches competitive performance while fitting into a student-friendly GPU. Future work could explore (i) deformable convolutions for shape variance, and (ii) boundary-aware losses, as suggested by [5].

Acknowledgment

Dataset and starter code courtesy of the 113-2 NYCU *Visual Recognition using Deep Learning* teaching team.

Reproducibility Resources

- **Source code**: github.com/111550135/HW3
- **Checkpoints**: Vertical-flip model `0320_18_model_0002999.pth` Horizontal-flip model `0322_19_model_0002999.pth`

Lint Code

```
(uav) oao@owo NYCU-Computer-Vision-2025-Spring-HW3 % nbqa flake8 111550135.ipynb --ignore=E402
(uav) oao@owo NYCU-Computer-Vision-2025-Spring-HW3 %
```

Fig. 2. run `nbqa flake8 111550135.ipynb --ignore=E402` snapshot

References

[1] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019, gitHub repository.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. ICCV*, 2017.

[3] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, 2017.

[4] R. Solovyev, V. Artemov, I. Kiselev, and A. Tsvetkov, "Weighted boxes fusion: Ensembling boxes from different object detectors," *ArXiv preprint arXiv:1910.13302*, 2021.

[5] X. W. *et al.*, "Explicit boundary handling in cnns," in *Proc. CVPR*, 2020.