

Evaluating New AI Cell Foundation Models on Challenging Kidney Pathology Cases Unaddressed by Previous Foundation Models

Runchen Wang^{a†}, Junlin Guo^{b†}, Siqi Lu^b, Ruining Deng^c, Zhengyi Lu^b, Yanfan Zhu^a, Yuechen Yang^a, Chongyu Qu^b, Yu Wang^d, Shilin Zhao^d, Catie Chang^{a,b}, Mitchell Wilkes^b, Mengmeng Yin^e, Haichun Yang^e, and Yuankai Huo^{a,b,e}

^aDepartment of Computer Science, Vanderbilt University, Nashville, TN 37235, USA

^bDepartment of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN 37235, USA

^cDepartment of Radiology, Weill Cornell Medicine, New York, NY 10021, USA

^dDepartment of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37235, USA

^eDepartment of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN 37235, USA

[†]These authors contribute equally to this work.

ABSTRACT

Accurate cell nuclei segmentation is critical for downstream tasks in kidney pathology and remains a major challenge due to the morphological diversity and imaging variability of renal tissues. While our prior work has evaluated early-generation AI cell foundation models in this domain, the effectiveness of recent cell foundation models remains unclear. In this study, we benchmark advanced AI cell foundation models (2025), including CellViT++ variants and Cellpose-SAM, against three widely used cell foundation models developed prior to 2024, using a diverse large-scale set of kidney image patches within a human-in-the-loop rating framework. We further performed fusion-based ensemble evaluation and model agreement analysis to assess the segmentation capabilities of the different models. Our results show that CellViT++ [Virchow] yields the highest standalone performance with 40.3% of predictions rated as “Good” on a curated set of 2,091 challenging samples, outperforming all prior models. In addition, our fused model achieves 62.2% “Good” predictions and only 0.4% “Bad”, substantially reducing segmentation errors. Notably, the fusion model (2025) successfully resolved the majority of challenging cases that remained unaddressed in our previous study. These findings demonstrate the potential of AI cell foundation model development in renal pathology and provide a curated dataset of challenging samples to support future kidney-specific model refinement.

Keywords: Foundation Model, Instance Segmentation, Ensemble Learning, Kidney Pathology, Cell Nuclei, Human-in-the-loop

1. INTRODUCTION

Nuclei instance segmentation is a critical task in computational pathology,^{1,2} serving as a foundational step for downstream analyses such as cell counting,^{3,4} phenotype classification,^{5,6} and spatial transcriptomics.^{7,8} However, traditional deep learning approaches often depend on extensive pixel-level annotations and struggle to generalize across various tissue types, staining protocols, and imaging modalities. This limits their scalability and clinical applicability, especially in high-variance domains such as kidney pathology. These challenges have motivated a shift to foundation models, large-scale pre-trained architectures built for broad adaptability across diverse biomedical tasks.^{4,9,10}

Further author information: (Send correspondence to Yuankai Huo)

E-mail: yuankai.huo@vanderbilt.edu

In our prior work (2024) ,¹¹ we evaluated three widely used cell foundation models—Cellpose,¹² StarDist,¹³ and CellViT¹⁴—on a large-scale kidney pathology dataset consisting of 8,789 high-resolution image patches sampled from 2,542 kidney whole slide images (WSIs). Although these models demonstrated promising zero-shot performance, they still consistently failed in challenging image patches, such as those with low contrast, overlapping nuclei, or atypical morphologies. As large-scale deep learning continues to evolve, a key question arises: **Can newly released cell foundation models overcome the limitations of previous models?**

Motivated by this question, we conduct a comprehensive evaluation of newly released (2025) AI cell foundation models for histopathology, focusing on the task of nuclei instance segmentation in the kidney. These include CellViT++¹⁵ and Cellpose-SAM,¹⁶ which leverage large-scale pretrained Vision Transformers (e.g., SAM,¹⁷ Virchow,¹⁸ HIPT¹⁹) and are integrated into established cell instance segmentation frameworks. This enables better generalization and segmentation accuracy, particularly in the challenging or ambiguous samples identified in our prior assessment. A standardized inference pipeline is used across models and human-in-the-loop quality ratings are incorporated to assess segmentation quality in a realistic setting, as in our prior work. The overall framework for this study is shown in Fig. 1.

In summary, our study makes the following contributions:

- **Benchmarking new AI cell foundation models:** We provide the first comprehensive evaluation of three CellViT++ variants (HIPT, SAM, Virchow) and Cellpose-SAM on challenging kidney pathology patches that prior models failed to segment, highlighting their evolving capabilities.
- **Human-in-the-loop evaluation of hard cases:** We extend our assessment to 2,091 curated difficult samples, rated into “Good”, “Medium”, “Bad” categories, ensuring clinically meaningful performance measurement.
- **Demonstrated robustness and precision:** By targeting hard cases missed in our 2024 study, we show that newer models achieve stronger generalization and higher-quality segmentation.

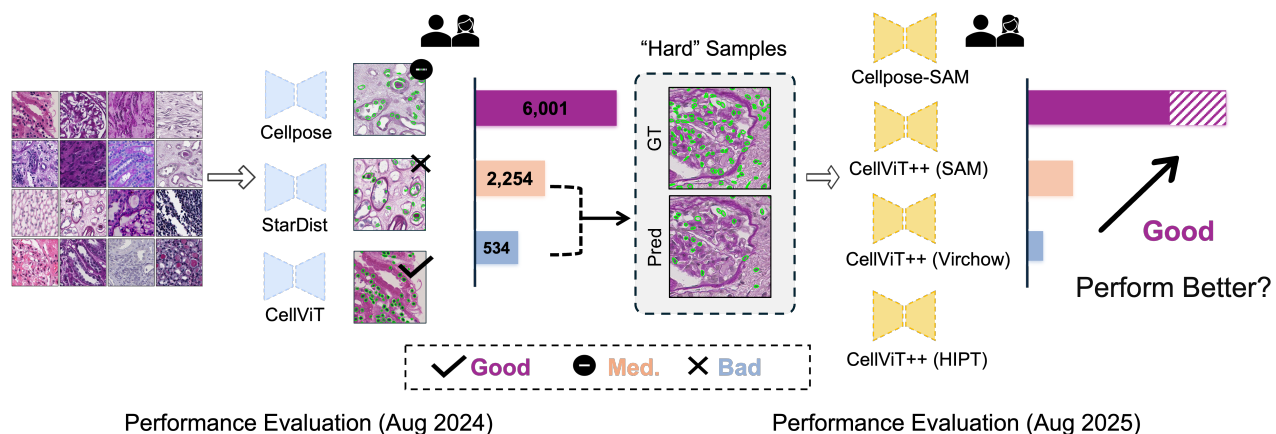


Figure 1: **Overall framework.** The previous study (Aug 2024) performed kidney cell nuclei instance segmentation using three cell foundation models: Cellpose,¹² StarDist,¹³ and CellViT.¹⁴ It revealed segmentation limitations in hard kidney pathology cases. These difficult cases were re-evaluated and categorized in Aug 2025 using four state-of-the-art models: three CellViT++¹⁵ variants and Cellpose-SAM.¹⁶ Model performance was evaluated based on human-in-the-loop rating for each prediction mask.

2. METHODS

Fig. 2 presents the new AI cell foundation models evaluated in this study. On the left, the cell nuclei instance segmentation task is shown. Compared to our earlier work (before August 2024, shown in blue), the new models adopt vision transformer architectures for both encoder and decoder. These encoders (shown in distinct colors) incorporate large-scale pretrained foundation models such as HIPT,¹⁹ Virchow,¹⁸ SAM.¹⁷

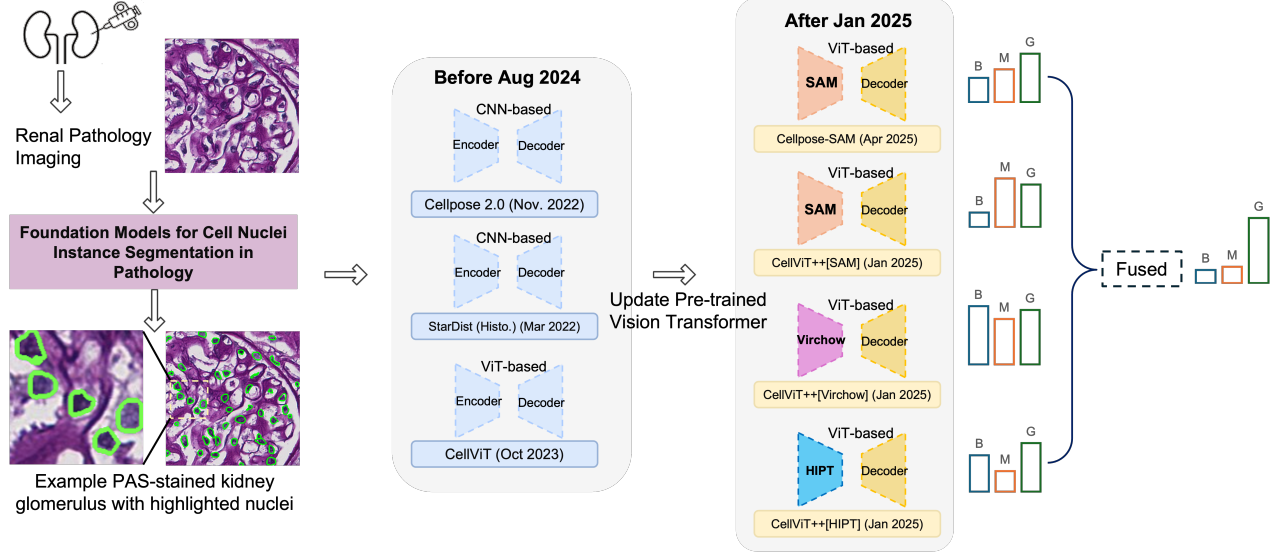


Figure 2: **Illustration of new AI cell foundation models in this evaluation.** Compared to our earlier work (before August 2024, shown in blue), the new models adopt vision transformer architectures for both encoder and decoder. These encoders (shown in distinct colors) incorporate large-scale pretrained foundation models such as HIPT,¹⁹ Virchow,¹⁸ SAM.¹⁷

2.1 Evaluated Cell Nuclei Foundation Models

In this assessment, we benchmark four state-of-the-art cell nuclei vision foundation models on challenging or uncertain samples that were unaddressed in our prior studies.^{11,20} Details of the models are summarized below and in Table 1.

1. **CellViT++[HIPT/SAM/Virchow]:** CellViT++¹⁵ is a generalized cell segmentation framework for H&E-stained images. Compared to its predecessor, CellViT,¹⁴ it leverages large-scale, next-generation vision foundation model pretrained weights (e.g., HIPT,¹⁹ Virchow,¹⁸ SAM¹⁷). Similar to CellViT, CellViT++ is fine-tuned for pathology cell nuclei segmentation using the PanNuke²¹ dataset.
2. **Cellpose-SAM:** Cellpose-SAM¹⁶ integrates the pretrained transformer backbone from the SAM¹⁷ foundation model into the Cellpose framework. Like the original Cellpose,¹² which uses a U-shaped CNN backbone, it outputs horizontal and vertical topological gradients along with semantic predictions. Instance segmentation is derived using the Cellpose gradient flow tracking algorithm.

Table 1: State-of-the-Art Cell Nuclei Segmentation Models in Pathology (2025)

Model	Release Date	Backbone	Post-processing
CellViT++ (SAM) ¹⁵	2025 Jan	ViT (SAM-H) ¹⁷	HoVer-Net ²²
CellViT++ (HIPT) ¹⁵	2025 Jan	ViT (HIPT ₂₅₆) ¹⁹	HoVer-Net ²²
CellViT++ (Virchow) ¹⁵	2025 Jan	ViT (Virchow) ¹⁸	HoVer-Net ²²
Cellpose-SAM ¹⁶	2025 Apr	ViT (SAM) ¹⁷	Gradient Tracking ¹²

2.2 Rating Criteria

To assess segmentation quality, we applied rating criteria from our prior studies,^{11,20} developed by a renal pathologist with over 10 years of experience. As shown in Fig. 3, two trained students reviewed each model’s predictions and assigned one of three labels: “Good”, “Medium”, or “Bad.” This framework ensured consistent, quantitative evaluation. The curated patches also serve as valuable resources for training or fine-tuning future foundation models in kidney pathology.

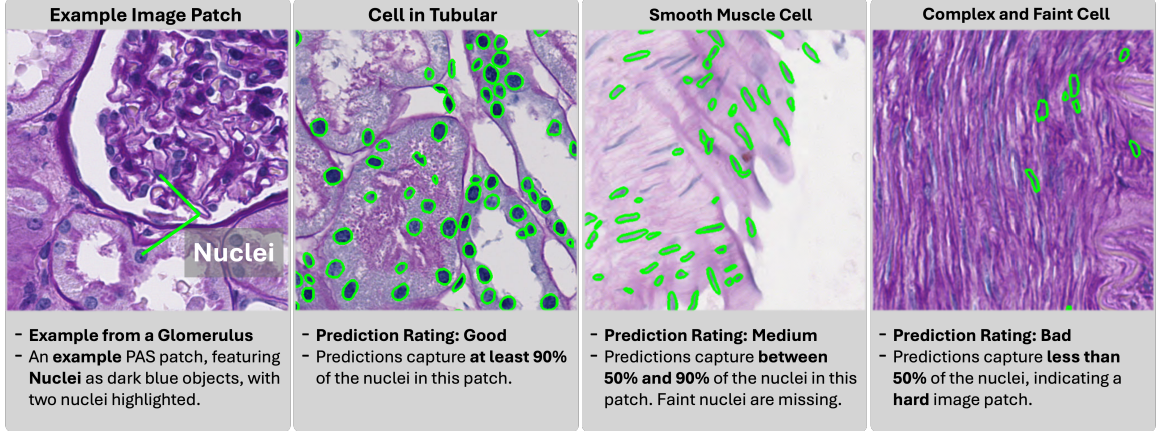


Figure 3: **Rating criteria for this assessment.** Illustrative details of the different rating category examples are provided.

3. DATA AND EXPERIMENTS

Implementations: The model inference pipeline was implemented using Python 3.13.5 and PyTorch 2.7.1, with GPU acceleration provided by CUDA 12.9. All experiments were conducted on an NVIDIA RTX A5000 GPU with 24 GB of memory. The rating was conducted using the Python GUI toolkit, Tkinter.

3.1 Evaluation Dataset

In this study, the evaluation dataset is sourced from our previous assessment, comprising 8,789 image patches sampled from 2,542 WSIs stained with H&E, PAS, or PASM. These WSIs were obtained from public datasets (e.g., KPMP,²³ NEPTUNE,²⁴ HuBMAP²⁵) and in-house data from Vanderbilt University Medical Center. All images were cropped into 512×512 -pixel patches at $40\times$ magnification.

This second-round assessment focuses on difficult samples unaddressed in the prior studies.^{11,20} A curated subset of 2,091 patches—rated as “Medium” or “Bad” and exhibiting greater segmentation uncertainty—was selected as the evaluation dataset.

3.2 Individual and Fused Model Performance

We systematically evaluated the performance of each foundation model across our evaluation dataset and further examined a fused strategy that ensembled their outputs. The individual model evaluation focused on the distribution of different prediction classes (“Good”, “Medium”, “Bad”), providing insights into the performance and behavior of each model. The prediction class assignments were determined through the rating criteria in Fig. 3.

Building on each cell foundation model’s performance, we also constructed a fused model based on the agreement of model-generated ratings. A patch was labeled as “Good” if at least one model’s output was rated as “Good”, and “Bad” only if all models were rated as “Bad”. Patches not meeting either condition were labeled as “Medium.” This fusion scheme simulates the potential benefit of model complementarity and also serves to identify common failure cases—patches rated as “Bad” across all models highlight challenging cases and direct future model improvement.

3.3 Cross-Model Performance Evaluation

Similar to our prior work,¹¹ we conducted a cross-model agreement analysis to assess consistency and divergence among the four cell foundation models. We first computed a **pairwise agreement matrix**, which captures the percentage of image patches where each model pair assigns the same quality label (“Good”, “Medium”, or “Bad”). This quantifies alignment between model predictions and reveals systematic differences in behavior. We further categorized each patch based on consensus levels across all four models—**All Agree**, **Three Agree**, **Two Agree**, and **No Agreement**—to highlight class-wise consistency and guide ensemble strategy design.

3.4 Comparison with Previous Study

To enable a rigorous comparison with our prior 2024 study, we compared the performance of the updated 2025 fusion model with the 2024 fusion results across two evaluation scenarios: (1) a challenging subset of 2,091 image patches curated for hard cases, and (2) the full set of approximately 8,789 image patches from the 2024 evaluation. For consistency, the same rating protocol and annotation schema (“Good”, “Medium”, “Bad”) were applied. This comparative setup ensures that observed performance differences reflect genuine model improvements rather than variations in evaluation design or data distribution.

4. RESULTS

4.1 Individual Model Performance Evaluation.

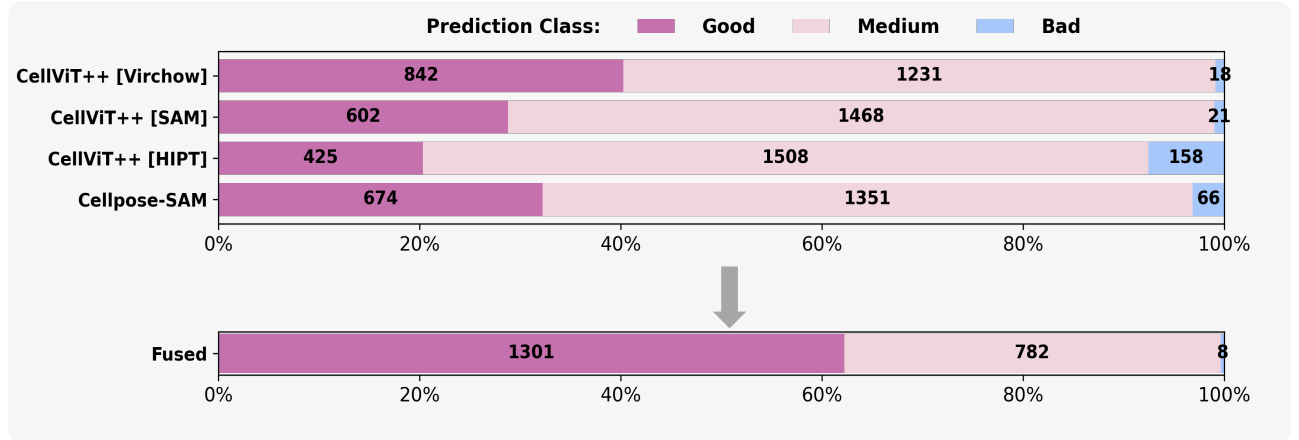


Figure 4: **Distribution of rated predictions from AI cell foundation models across the evaluation dataset.** Each row represents the foundation model’s predictions, with three values corresponding to the number of predictions rated as “Good,” “Medium,” and “Bad,” respectively. Then, data enrichment (shown as “Fused” Model) was performed based on the evaluation results of individual models, resulting in an increase in “Good” image patches and a decrease in “Bad” image patches.

We began by evaluating the standalone performance of each foundation model on the task of cell nuclei instance segmentation on kidney pathology image patches using human ratings. The upper panel of Fig. 4 presents the distribution of predictions categorized as “Good”, “Medium”, or “Bad” based on segmentation quality across four models: CellViT++ [HIPT], CellViT++ [SAM], CellViT++ [Virchow], and Cellpose-SAM.

Among the models, CellViT++ [Virchow] achieved the strongest performance, with 842 patches (40.3%) rated as “Good” and only 18 (0.9%) as “Bad”, reflecting high generalizability to kidney pathology and suggesting a high degree of alignment between its predictions and human judgment. Cellpose-SAM and CellViT++ [SAM] followed closely, with 674 (32.2%) and 602 (28.8%) “Good” ratings, respectively, and fewer than 2.9% “Bad” predictions. In contrast, CellViT++ [HIPT] produced the proportion of “Good” ratings (425, 20.3%) and the highest “Bad” ratio (158, 7.6%), indicating weaker robustness on this evaluation dataset in comparison to other models. Although the majority of predictions for all models were rated as “Medium” (58.9–72.1%), the differences in “Good” and “Bad” outcomes underscore the improvement of instance segmentation quality from new AI cell foundation models. These improvements reflect the impact of large-scale pretraining on model generalization for complex cell nuclei in pathology.

4.2 Fused Model Performance Evaluation.

To evaluate whether combining outputs from multiple foundation models enhances segmentation reliability, we implemented a fused model that aggregates their ratings. As shown in the bottom panel of Fig. 4, the fused model achieved 62.2% “Good” ($n = 1,301$) and 37.4% “Medium” ($n = 781$) predictions, with only 0.4% ($n = 8$)

rated as “Bad”. This represents a substantial reduction in segmentation failures compared to any individual model, suggesting that the fusion strategy effectively mitigates individual model weaknesses.

Notably, this improvement was achieved without retraining or additional supervision, demonstrating the efficacy of ensemble integration in leveraging complementary model strengths. The near-elimination of “Bad” predictions highlights the potential of fusion-based strategies to deliver more robust and dependable cell segmentation in complex kidney pathology settings.

4.3 Cross-Model Performance Agreement.

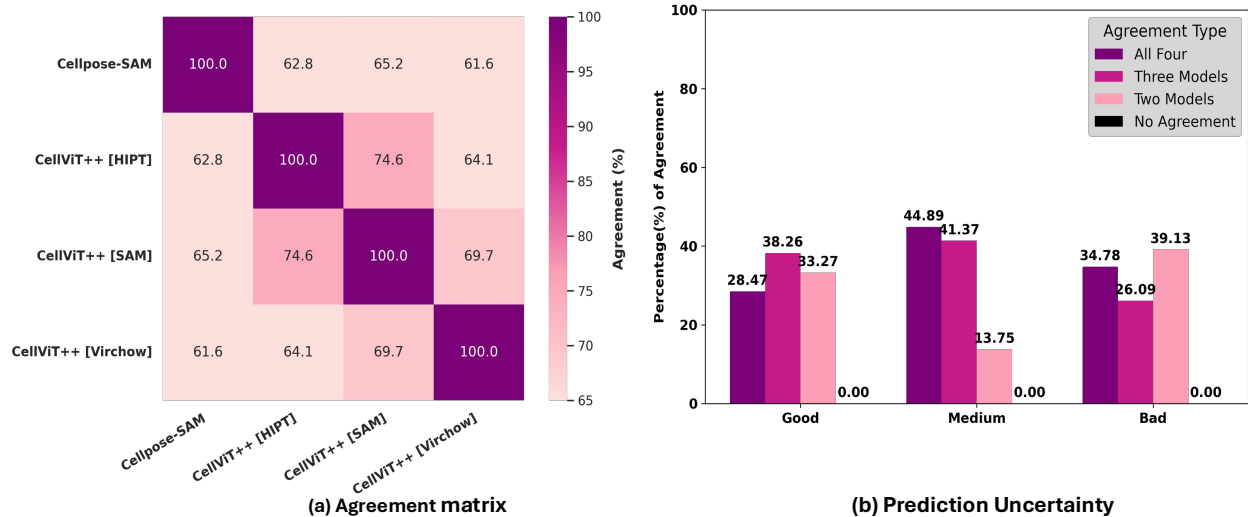


Figure 5: **Cross-model performance rating agreement.** (a) shows the agreement percentages between each pair of foundation models used in this study. To further assess the cross-model performance, (b) shows the percentages of image patches where all four models agree, three agree, two agree, or no agreement, for each rating category (“Good”, “Medium”, “Bad”).

As shown in Fig. 5a, pairwise agreement scores among different models ranged from 61.6% to 74.6%, with the highest between CellViT++ [SAM] and CellViT++ [HIPT], reflecting shared instance segmentation framework and fine-tuning strategy. In contrast, Cellpose-SAM shows lower alignment with CellViT++ variants, reflecting architectural divergence and a distinct training regime.

To further dissect prediction uncertainty, we analyzed agreement per rating category. Fig. 5b shows that “Medium” ratings had the highest 4-model consensus, while “Good” had the lowest (28.47% full agreement), indicating a larger performance gap among models on certain challenging image patches. “Bad” predictions showed moderate agreement. To sum up, these results highlight meaningful prediction diversity across new AI cell foundation models and motivate our use of fusion strategies that combine multiple state-of-the-art cell foundation models to reduce prediction uncertainty and improve nuclei segmentation outcomes.

4.4 Comparative Study with Previous Work.

We evaluated the performance of our 2024 and 2025 fusion models on both challenging cases (2,091 image patches) and the complete evaluation dataset (8,789 image patches) to assess segmentation quality improvements. As shown in Fig. 6a, the 2025 model significantly improved performance on challenging cases: “Good” predictions rose from 134 (6.4%) to 1,301 (62.2%), “Medium” dropped from 1,957 (93.5%) to 782 (37.4%). Fig. 6b shows consistent gains on the full dataset, with “Good” increasing from 6,001 (68.3%) to 7,168 (81.6%), “Medium” decreasing from 2,254 (25.6%) to 1,079 (12.3%), and “Bad” remaining stable at about 6.2%.

These results demonstrate that fusing updated AI cell foundation models improves nuclei instance segmentation precision and robustness, with particular strength in converting previously difficult samples to high-quality segmentation outputs. The substantial reduction in Medium-quality predictions across both evaluation scenarios validates the utility of this fusion strategy, based on multiple new AI cell foundation models, for automated renal histopathology analysis.

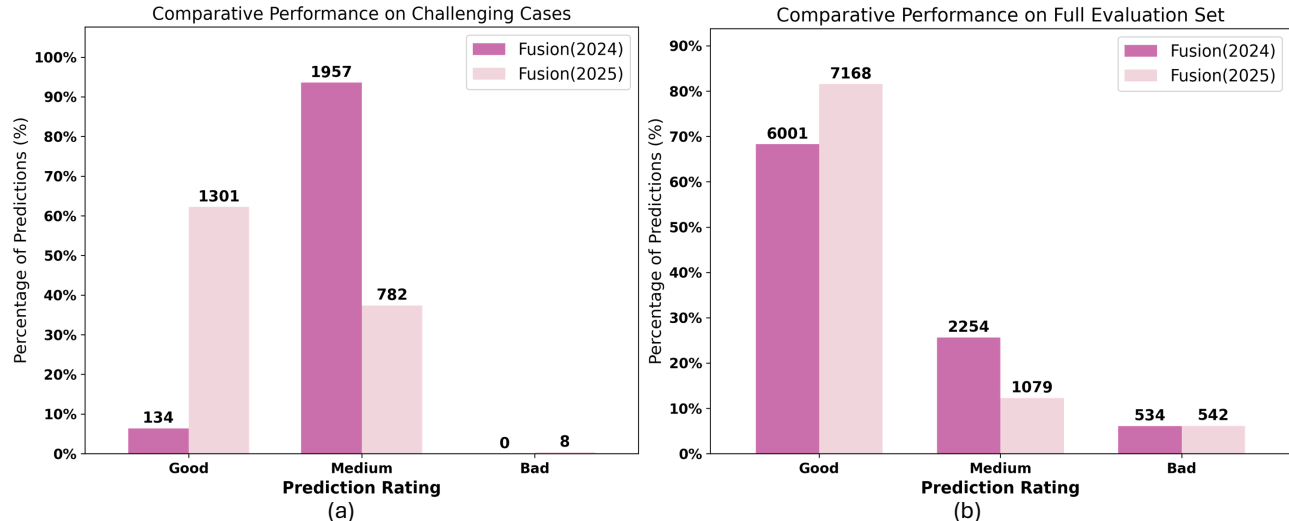


Figure 6: **Comparative analysis of fusion model performance with previous work.** (a) Evaluation of 2024 and 2025 fusion models on 2,091 challenging image patches. (b) Performance comparison on the full dataset (8,789 image patches). In both evaluations, the 2025 fusion model yields more “Good” predictions, fewer “Medium” predictions, while the proportion of “Bad” remains stable.

4.5 Qualitative Results.

Fig. 7 shows two example patches where new AI cell foundation models produced “Good” predictions missed by prior cell foundation models (Cellpose, StarDist, CellViT). In general, as shown in the original model predictions (first row in both examples), these image patches are challenging, with faint nuclei and low image contrast between the nuclei and the background. In Fig. 7A, we can see that Cellpose-SAM reduces false negatives (FN) from StarDist and CellViT. The highlighted yellow box shows that Cellpose-SAM can even detect nuclei in regions with heavy staining, thereby reducing the false positives (FP) from the Cellpose prediction. Similarly, in Fig. 7B, the new AI foundation models demonstrate improved segmentation performance, as most faint nuclei in the lower left corner of the image patch are correctly segmented and the FP from the original Cellpose prediction is eliminated in the updated model predictions.

5. CONCLUSION

In this study, we build upon our prior evaluation of foundation models for kidney cell nuclei segmentation by addressing a key limitation: the failure to robustly segment hard image patches. We benchmark SOTA AI cell foundation models (2025), including three CellViT++ variants and Cellpose-SAM, and introduced a fusion strategy that integrates their strengths. Our results show that the 2025 fusion model substantially outperforms previous models across the full evaluation set and resolves the vast majority of hard cases previously left undressed. Specifically, the updated AI cell foundation models increased the proportion of “Good” segmentation outcomes while nearly eliminating “Bad” predictions on 2,091 difficult samples. These findings highlight the value of ensemble multiple new AI cell foundation models in enhancing generalization and robustness for organ-specific applications. The curated hard image patches benchmark and improved nuclei segmentation outcomes offer a practical foundation for future model refinement and deployment in renal pathology workflows.

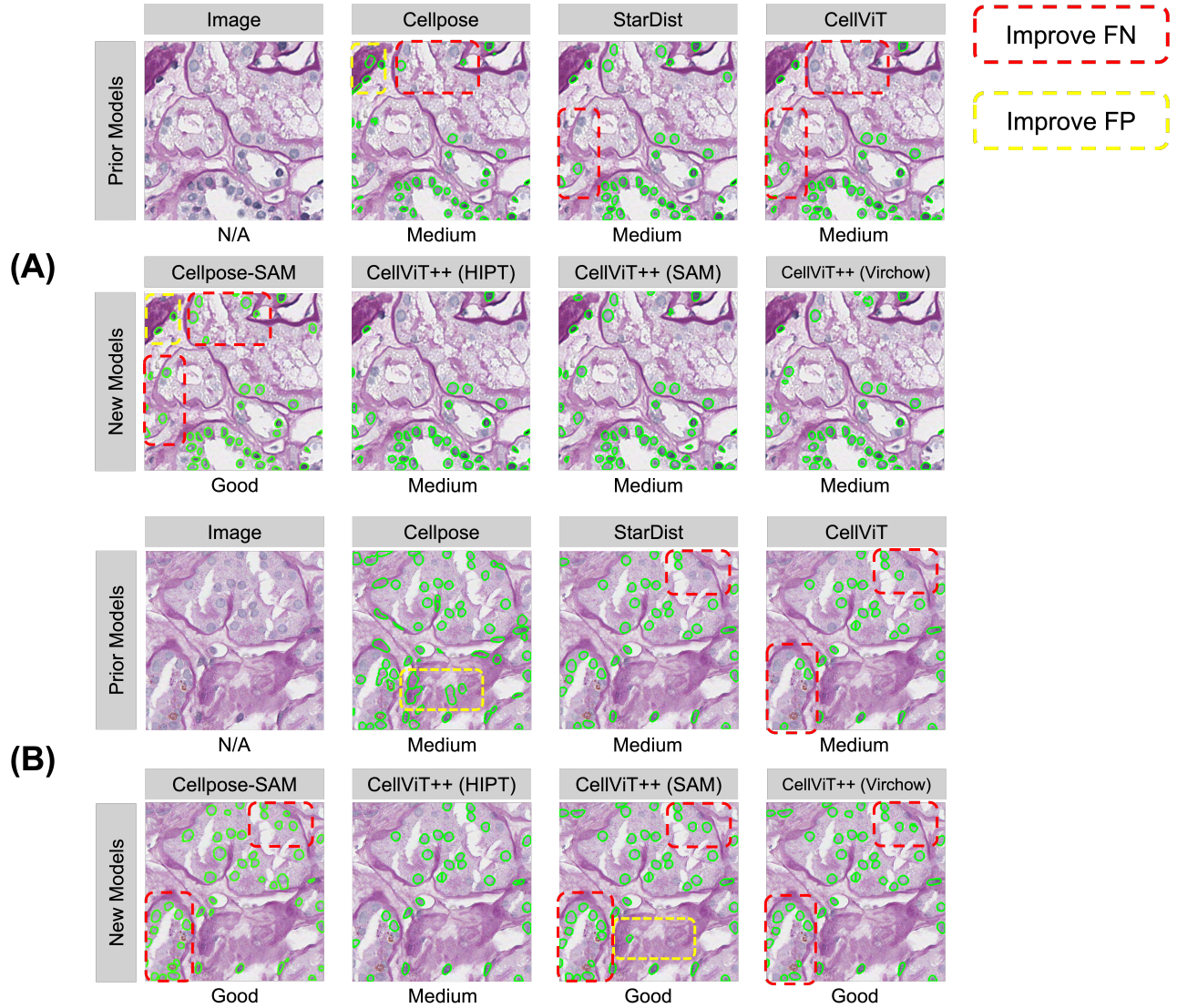


Figure 7: **Qualitative results.** Examples of patches where new AI cell foundation models produced “Good” predictions missed by earlier models. Highlighted regions illustrate improvements in false positives (**FP**) and false negatives (**FN**) compared to prior AI cell foundation models: The **yellow box** marks areas where the new models reduce FP from the original predictions, while the **red box** highlights improved segmentation of FN regions.

ACKNOWLEDGMENTS

This research was supported by NIH R01DK135597 (Huo), DoD HT9425-23-1-0003 (HCY), and KPMP Glue Grant. This work was also supported by Vanderbilt Seed Success Grant, Vanderbilt Discovery Grant, and VISE Seed Grant. This project was supported by The Leona M. and Harry B. Helmsley Charitable Trust grant G-1903-03793 and G-2103-05128. This research was also supported by NIH grants R01EB033385, R01DK132338, REB017230, R01MH125931, and NSF 2040462. We extend gratitude to NVIDIA for their support by means of the NVIDIA hardware grant. This work was also supported by NSF NAIRR Pilot Award NAIRR240055.

REFERENCES

- [1] Deng, R., Yang, Y., Pisapia, D. J., Liechty, B., Zhu, J., Xiong, J., Guo, J., Lu, Z., Wang, J., Yao, X., et al., “Casc-ai: Consensus-aware self-corrective ai agents for noise cell segmentation,” *arXiv e-prints*, arXiv-2502 (2025).
- [2] Chen, Y., Huang, Q., Geng, M., Wang, Z., and Han, Y., “A systematic review on cell nucleus instance segmentation,” *IET Image Processing* **19**(1), e70129 (2025).
- [3] Ke, J., Lu, Y., Shen, Y., Zhu, J., Zhou, Y., Huang, J., Yao, J., Liang, X., Guo, Y., Wei, Z., et al., “Clusterseg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets,” *Medical Image Analysis* **85**, 102758 (2023).
- [4] Cui, C., Deng, R., Guo, J., Liu, Q., Yao, T., Yang, H., and Huo, Y., “Pfps: Prompt-guided flexible pathological segmentation for diverse potential outcomes using large vision and language models,” *arXiv preprint arXiv:2407.09979* (2024).
- [5] Pratapa, A., Doron, M., and Caicedo, J. C., “Image-based cell phenotyping with deep learning,” *Current opinion in chemical biology* **65**, 9–17 (2021).
- [6] Yang, Y., Wang, Y., Yao, T., Deng, R., Yin, M., Zhao, S., Yang, H., and Huo, Y., “Pyspatial: A high-speed whole slide image pathomics toolkit,” *arXiv preprint arXiv:2501.06151* (2025).
- [7] Zhu, J., Deng, R., Yao, T., Xiong, J., Qu, C., Guo, J., Lu, S., Yin, M., Wang, Y., Zhao, S., et al., “Asign: an anatomy-aware spatial imputation graphic network for 3d spatial transcriptomics,” in [*Proceedings of the Computer Vision and Pattern Recognition Conference*], 30829–30838 (2025).
- [8] Zhu, J., Deng, R., Yao, T., Xiong, J., Qu, C., Guo, J., Lu, S., Tang, Y., Xu, D., Yin, M., et al., “Magnet: Multi-level attention graph network for predicting high-resolution spatial transcriptomics,” *arXiv preprint arXiv:2502.21011* (2025).
- [9] Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B., “Segment anything in medical images,” *Nature Communications* **15**, 654 (2024).
- [10] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J., “Llava-med: Training a large language-and-vision assistant for biomedicine in one day,” in [*Thirty-seventh Conference on Neural Information Processing Systems*], (2023).
- [11] Guo, J., Lu, S., Cui, C., Deng, R., Yao, T., Tao, Z., Lin, Y., Lions, M., Liu, Q., Xiong, J., et al., “Assessment of cell nuclei ai foundation models in kidney pathology,” in [*Medical Imaging 2025: Image Perception, Observer Performance, and Technology Assessment*], **13409**, 76–82, SPIE (2025).
- [12] Stringer, C., Michaelos, M., and Pachitariu, M., “Cellpose: a generalist algorithm for cellular segmentation,” *bioRxiv* (2020).
- [13] Weigert, M. and Schmidt, U., “Nuclei instance segmentation and classification in histopathology images with stardist,” in [*2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*], IEEE (Mar. 2022).
- [14] Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., and Kleesiek, J., “Cellvit: Vision transformers for precise cell segmentation and classification,” (2023).
- [15] Hörst, F., Rempe, M., Becker, H., Heine, L., Keyl, J., and Kleesiek, J., “Cellvit++: Energy-efficient and adaptive cell segmentation and classification using foundation models,” *arXiv preprint arXiv:2501.05269* (2025).
- [16] Pachitariu, M., Rariden, M., and Stringer, C., “Cellpose-sam: superhuman generalization for cellular segmentation,” *bioRxiv*, 2025–04 (2025).
- [17] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R., “Segment anything,” *arXiv:2304.02643* (2023).
- [18] Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Yang, E., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y. K., Kunz, J. D., Lee, M. C. H., Bernhard, J. H., Godrich, R. A., Oakley, G., Millar, E., Hanna, M., Wen, H., Retamero, J. A., Moye, W. A., Yousfi, R., Kanan, C., Klimstra, D. S., Rothrock, B., Liu, S., and Fuchs, T. J., “A foundation model for clinical-grade computational pathology and rare cancers detection,” *Nature Medicine* (2024).

- [19] Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., and Mahmood, F., “Scaling vision transformers to gigapixel images via hierarchical self-supervised learning,” in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 16144–16155 (2022).
- [20] Guo, J., Lu, S., Cui, C., Deng, R., Yao, T., Tao, Z., Lin, Y., Lions, M., Liu, Q., Xiong, J., et al., “How good are we? evaluating cell ai foundation models in kidney pathology with human-in-the-loop enrichment,” *arXiv preprint arXiv:2411.00078* (2024).
- [21] Gamper, J., Koohbanani, N. A., Benes, K., Khuram, A., and Rajpoot, N., “Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification,” in [*European Congress on Digital Pathology*], 11–19, Springer (2019).
- [22] Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., and Rajpoot, N., “Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images,” (2019).
- [23] “Kidney precision medicine project data.” <https://www.kpmp.org>. Funded by the National Institute of Diabetes and Digestive and Kidney Diseases (Grant numbers: U01DK133081, U01DK133091, U01DK133092, U01DK133093, U01DK133095, U01DK133097, U01DK114866, U01DK114908, U01DK133090, U01DK133113, U01DK133766, U01DK133768, U01DK114907, U01DK114920, U01DK114923, U01DK114933, U24DK114886, UH3DK114926, UH3DK114861, UH3DK114915, UH3DK114937).
- [24] Barisoni, L., Nast, C. C., Jennette, J. C., Hodgin, J. B., Herzenberg, A. M., Lemley, K. V., Conway, C. M., Kopp, J. B., Kretzler, M., Lienczewski, C., et al., “Digital pathology evaluation in the multicenter nephrotic syndrome study network (neptune),” *Clinical Journal of the American Society of Nephrology* **8**(8), 1449–1459 (2013).
- [25] Howard, A., Lawrence, A., Sims, B., Tinsley, E., Kazmierczak, J., Borner, K., Godwin, L., Novaes, M., Culliton, P., Holland, R., Watson, R., and Ju, Y., “Hubmap - hacking the kidney,” (2020).