

Semantic-Guided Multimodal Preprocessing for Vision Transformer-Based Clear Cell Renal Cell Carcinoma Grading

Fatemeh Javadian¹, Zhu Chen^{1,2}, Zahra Aminparast³, and Johannes Stegmaier^{1,2}

¹ Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany

fatemeh.javadian@rwth-aachen.de

² Machine Learning for Medical Data, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

³ Institute of Medical Sciences, Kermanshah University, Medical School, Kermanshah, Iran

Abstract. Clear cell renal cell carcinoma (CCRCC) grading is essential for treatment planning, yet existing approaches either analyze patch-level images directly or focus solely on nuclei-level classification, without linking to final tumor grading. We propose a semantic-guided multimodal preprocessing method that integrates nuclei classification maps from existing pre-trained models with RGB histopathology images for Vision Transformer (ViT)-based CCRCC grading. Our approach employs classification map channel concatenation and multiplicative modulation, with optimized overlays to leverage nuclei grading information, while preserving RGB textural features. Evaluation of multiple preprocessing strategies demonstrates that semantic-guided enhancement achieves 0.916 balanced accuracy, outperforming RGB-only baseline (0.707) and max-voting aggregation from prior studies (0.427). Sensitivity analysis reveals that this 21 percentage point improvement over baseline persists even under simulated perturbation at rates matching current state-of-the-art nuclei classification model error thresholds, suggesting both effective semantic utilization and practical robustness. These findings show that preprocessing-based multimodal fusion can leverage the diagnostic potential of existing imperfect nuclei classifiers, effectively bridging previously isolated fine-grained nuclear-level analysis with coarse-grained ViT-based patch classification. Per-class recall was consistent across grades (0.93, 0.91, 0.91), indicating that gains are not concentrated in the majority class. Because the sensitivity analysis perturbs ground-truth maps rather than predictions from an actual nuclei model, this result characterizes robustness under simulated error rather than deployment with a real upstream model, which remains for future work.

Keywords: Vision Transformers · Semantic guidance · Multimodal preprocessing · CCRCC grading · Nuclei classification · Histopathology image analysis · Computational pathology

1 Introduction

Clear cell renal cell carcinoma (CCRCC) is a common subtype of renal cancer, presenting complex histological features that challenge diagnosis and grading [3]. Grading guides assessment of tumor aggressiveness and treatment decisions [29,10]. The World Health Organization/International Society of Urological Pathology (WHO/ISUP) system grades CCRCC by nucleolar prominence, with higher grades indicating greater aggressiveness [11]. Grades 1–3 form a progressive, nucleus-centered continuum defined by reproducible nuclear features at specific magnifications. Grade 4 represents extreme dedifferentiation (sarcomatoid or rhabdoid morphology) where nuclei no longer follow consistent patterns, and is therefore analyzed separately [29,11]. Among nucleolar-based grades, grade 3 is the most clinically significant, followed by grades 2 and 1 [29,11]. Whole Slide Images (WSIs) are typically segmented into patches that are graded to infer overall tumor grade [4]. The WHO/ISUP system lacks formal quantitative thresholds; pathologists instead weigh the presence, spatial distribution, and morphology of high-grade nuclei [19,22]. This qualitative process shows significant inter-observer variability [3,22], limiting simple aggregation and requiring patch evaluation of both individual nuclei and tissue context [4].

Patch-level strategies are either coarse-grained, grading entire patches directly [27,6], or fine-grained, grading individual nuclei then aggregating [12]. Fine-grained classification captures tumor-cell heterogeneity and subtle morphological variation [13,31], but aggregating nuclei-level grades into patch predictions remains difficult. Many methods use max-voting, assigning the most abundant nuclei class as the patch grade [12,1], which systematically under-grades patches where sparse high-grade nuclei determine clinical significance [8,4].

Fig. 1 illustrates this limitation. Images 1–4 are grade 3 patches in which green (grade 1) and yellow (grade 2) nuclei far outnumber red (grade 3) nuclei; max-voting [12] would misclassify them as grade 1 or 2, whereas pathologists assign grade 3 from the presence and spatial distribution of high-grade nuclei even as a minority. Conversely, Image 5 is a grade 2 patch with scattered grade 3 nuclei, showing that mere presence of high-grade nuclei is insufficient: spatial clustering, density, and morphological context also matter. These subjective, non-linear relationships cannot be captured by simple voting.

1.1 State of the Art

U-Net architectures and variants [23,25,5,13] yield accurate per-nucleus grade predictions, but aggregating them via max-voting discards the spatial and morphological context essential for clinical grading [1]. Separately, coarse-grained methods grade patches directly using deep learning, particularly Vision Transformers (ViTs) that capture local and global patterns [7,26], bypassing nuclei-level analysis and relying on RGB images alone. A critical gap remains: fine-grained nuclei classifiers and coarse-grained patch classifiers stay isolated. Pre-trained nuclei models [12] encode substantial cellular-morphology knowledge unused by patch-level systems. Multimodal fusion in computational pathology [9]

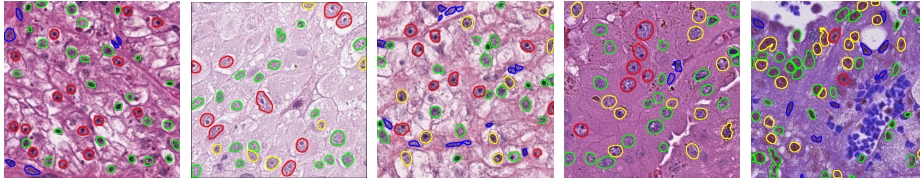


Fig. 1. Limitation of max-voting for CCRCC patch grading. Nuclei are outlined by 4 grades: blue (grade 0, non-tumorous cells), green (grade 1), yellow (grade 2), red (grade 3). **Images 1-4 (left):** Despite predominance of low-grade nuclei (green/yellow), all four patches are assigned grade 3 by the pathologist due to presence and spatial distribution of high-grade nuclei (red). Max-voting (selecting the most abundant grade) would incorrectly classify these as grade 1 or 2. **Image 5 (right):** A grade 2 patch showing that mere presence of scattered grade 3 nuclei is insufficient, grade assignment depends on quantity, spatial distribution, and morphological context, determined subjectively by the pathologist.

integrates complementary sources, but preprocessing-based fusion of nuclei maps with RGB images for patch grading remains, to our knowledge, unexplored.

We bridge this gap with a semantic-guided multimodal preprocessing method⁴ integrating nuclei classification maps from pre-trained models with RGB images for ViT-based CCRCC grading. Multiplicative modulation with perceptually optimized overlays embeds nuclei grading information directly into ViT inputs, letting the model learn grade-determining patterns that combine fine-grained nuclei information with tissue context. Sensitivity analyses quantify how imperfect nuclei classification affects patch-level grading, indicating applicability with existing models. Dataset details are in Section 2.1; subsequent sections present the methodology, results, and workflow implications.

2 Methodology

2.1 Dataset

The dataset comprises 1000 H&E stained patches at 512×512 pixels, with patch-level ground truth annotated by a pathologist following WHO/ISUP guidelines [3,29] for grades 1–3, enabling comparison with aggregation methods such as max-voting. Single-pathologist annotation disregards inter-observer variability; our focus is the methodological advantage of semantic-guided preprocessing, not clinical-grade ground truth. The original distribution was highly imbalanced (grade 1: 66.3%, grade 2: 23.0%, grade 3: 10.7%). Data were split into training (70%), validation (10%), and test (20%). The patches were selected from whole-slide images of the TCGA KIRC and KIRP projects, following the patch selection

⁴ We use *multimodal* to denote the fusion of the RGB image with a derived semantic representation of the same tissue section, rather than distinct imaging modalities; both inputs originate from the same H&E image.

and annotation of [12]. The source dataset is distributed as a flat collection of patches without an explicit patch-to-slide mapping, and its own description specifies only that the data were randomly divided into training, validation and test partitions, without stating whether this division was performed at the slide, case or patch level. Our split therefore follows the same patch-level protocol to remain comparable with the source benchmark, and slide-level grouping could not be verified or enforced. Consequently, patches originating from the same slide may appear in different partitions, and residual correlation arising from shared staining, scanner characteristics and local tissue architecture cannot be excluded. To address imbalance, targeted augmentation (horizontal/vertical flips) was applied to minority classes in the training set only, yielding a near-balanced training distribution (grade 1: 35.7%, grade 2: 35.1%, grade 3: 29.2%).

Table 1. Vision Transformer model comparison for patch-level CCRCC classification on RGB-only histopathology data (without semantic guidance). All models were fine-tuned with identical protocol; performance shown on test set.

Model	F1	Accuracy	Precision	Recall
Google ViT Base Patch32-384	0.8095	0.8077	0.8181	0.8077
Google ViT Large Patch32-384	0.8079	0.8076	0.8112	0.8076
Google ViT Base Patch16-224	0.7766	0.7821	0.7846	0.7821
Google ViT Base Patch16-224-in21k	0.7536	0.7564	0.7546	0.7564
OpenAI CLIP ViT Base Patch32	0.5454	0.5513	0.5538	0.5513

2.2 Vision Transformer Selection and Fine-Tuning

We fine-tuned multiple pre-trained ViT architectures on our CCRCC dataset using RGB images only (without semantic guidance), replacing the classification head with a three-class output layer and updating all weights. All patches were resized to 384×384 pixels. Table 1 compares performance. Google ViT Base Patch32-384, pre-trained on ImageNet-21k, achieved the highest performance and was selected for all experiments; it processes images as 32×32 pixel-token sequences, capturing local and global tissue context [7,26]. Fine-tuning used learning rate 10^{-4} with cosine annealing, batch size 32, AdamW, and 50 epochs with early stopping. The same protocol was applied to all multimodal fusion methods below.

2.3 Semantic-Guided Multimodal Preprocessing

Our approach integrates nuclei classification maps with RGB images through preprocessing, letting the ViT leverage semantic nuclei information without architectural modifications. Maps from pre-trained models [15,12] are combined with RGB patches before input to the fine-tuned ViT. We developed two methods: Classification Map Channel Concatenation and Multiplicative Modulation.

Classification Map Channel Concatenation The classification map channel concatenation (HEC) method uses color deconvolution [24] to separate H&E staining components, creating a three-channel input where nuclei grading information replaces the original blue channel. RGB patches are converted to hematoxylin (H) and eosin (E) grayscale channels via standard deconvolution matrices [24]. The third channel (C) holds the nuclei classification map with five values: background (0), tumor grades 1–3, and grade 0 non-tumorous cells (4). Value 4 denotes a non-tumorous cell label, unrelated to WHO/ISUP grade 4, which is excluded here. Values are linearly mapped to $[0, 255]$ (output = class $\times 63.75$) for compatibility with standard image processing. The ViT is fine-tuned on HEC inputs using the RGB-only protocol.

Multiplicative Modulation Method HEC combines modalities but requires color deconvolution and does not explicitly emphasize clinically important grades. Multiplicative Modulation (MM) is a single configurable method that preserves the original RGB image while emphasizing tumor regions by nuclei grade importance, grounded in information-theoretic fusion principles [9]. Building on multiplicative enhancement in medical imaging [20,18], particularly Retinex-based reflectance/illumination decomposition, MM unifies four components (intensity modulation, grade-specific weighting, spatial smoothing, and color overlay) into one preprocessing step applied before ViT input (Fig. 2).

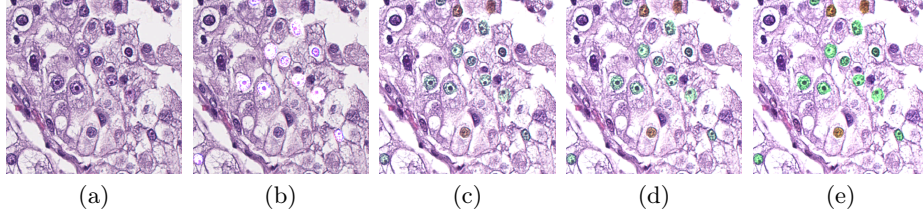


Fig. 2. Preprocessing examples from left to right: (a) Original RGB image, (b) MM ($\alpha = 0.9$, $\beta = 5.0$, $\sigma = 1.5$, overlay=0), (c) MM ($\alpha = 0.5$, $\beta = 2.0$, $\sigma = 2.0$, overlay=0.2), (d) MM ($\alpha = 0.7$, $\beta = 2.5$, $\sigma = 2.0$, overlay=0.35), (e) MM ($\alpha = 0.85$, $\beta = 3.0$, $\sigma = 1.5$, overlay=0.5).

Intensity Modulation: For each pixel (x, y) , enhanced intensity is

$$I'(x, y) = I(x, y) \cdot (1 + \alpha \cdot f(C(x, y))) \quad (1)$$

where $I(x, y)$ is the original RGB intensity, $C(x, y) \in \{0, 1, 2, 3, 4\}$ is the nuclei classification label (0=background, 1–3=tumor grades, 4=non-tumorous $\equiv$ grade 0), f is the grade-dependent weighting below, and $\alpha > 0$ controls modulation strength. Higher α increases nuclei influence; as $\alpha \rightarrow 0$, the output approaches the original RGB. The multiplicative form preserves original gradients, $\nabla I' = (1 + \alpha \cdot w_c) \nabla I$, scaling rather than eliminating edge information and

maintaining texture features critical for ViT attention [14] while adding semantic emphasis via the classification map [30].

Grade-Dependent Weighting: The weighting f in Eq. 2 uses a sigmoid for smooth grade transitions, following ordinal classification principles for hierarchical medical grading [16,28]:

$$f(c) = \frac{\exp(\beta \cdot (c - c_0))}{1 + \exp(\beta \cdot (c - c_0))} \quad (2)$$

where β controls steepness and $c_0 = 1.5$ positions the midpoint between grades 1 and 2 so grade 3 receives maximum emphasis. The sigmoid applies only to tumor grades (1–3), mapping each to a weight in $[0, 1]$ with $f(3) \approx 1$, reflecting grade 3 as most clinically significant [29,11]. Background and non-tumorous cells receive fixed weights outside the sigmoid. In practice, the sigmoid combined with grade-specific base weights w_c yields a clinical-importance hierarchy: background/non-tumorous $w_0 = w_4 = 1.00$; grade 1 $w_1 = 1.10$ – 1.25 ; grade 2 $w_2 = 1.40$ – 1.80 ; grade 3 $w_3 = 1.85$ – 2.0 . These are applied in Eq. 1 through $f(C(x, y))$.

Spatial Smoothing: To avoid harsh boundaries between adjacent nuclei regions, Gaussian smoothing ($\sigma = 1.5$ – 2.0 pixels) [21] is applied to the classification map before fusion, operating only on the classification channel, not the RGB data, preserving tissue texture while softening region boundaries.

Color Overlay: Optionally, perceptually optimized color overlays [17] are blended with the modulated image using coefficient $O \in [0, 1]$, output $(1-O) \cdot I + O \cdot C_{\text{color}} \cdot I$, practical range 0.2–0.5. Distinct colors mark each grade: green (1), yellow (2), red (3) (Fig. 2), giving the model explicit class distinction beyond intensity differences. Fig. 2(b) shows a no-overlay case: modulation and weighting still enhance, but grades remain visually similar, giving spatial segmentation without color-based distinction. Overlay-free configurations achieve notably lower balanced accuracy (Table 2), indicating color overlay is the primary carrier of class-discriminative information.

Parameter Selection: We evaluated MM configurations by grid search over α (0.5–0.9), β (2.0–5.0), σ (0–2.0), overlay (0–0.5), and grade weights, assessing first on validation; Table 2 reports test results for top configurations. The best ($\alpha = 0.85$, $\beta = 3.0$, $\sigma = 1.5$, overlay=0.5) reached 0.9160 balanced accuracy versus 0.707 for the identically trained RGB-only baseline.

2.4 Sensitivity Analysis

Real-world deployment relies on pre-trained nuclei model predictions [12,15] rather than ground truth. We tested robustness to imperfect classification by injecting controlled errors into classification maps: (1) segmentation errors: random nuclei nullified, simulating false negatives; (2) classification errors: grades randomly altered with adjacent-grade preference (e.g., grade 2 $\leftrightarrow$ grade 3), reflecting realistic misclassification [15]. Both were applied simultaneously from 0% to 60%. Models were trained on ground truth; perturbation applied only at evaluation, isolating nuclei-accuracy impact while holding RGB constant. Performance used balanced accuracy and F1.

Table 2. Test Set Performance of Preprocessing Methods (sorted by Balanced Accuracy). MM: Multiplicative Modulation with parameters α (modulation strength), β (sigmoid steepness), σ (spatial smoothing), O (color overlay). Precision, Recall, and Accuracy are micro-averaged; F1 is macro-averaged. All MM configurations maintain high recall, ensuring critical high-grade cases are not missed. **Bold** indicates best performance per metric; underline indicates second best.

Model	α	β	Bal. Acc.	Accuracy	Precision	Recall	F1
Original RGB	N/A	N/A	0.7071	0.7723	0.7723	0.7723	0.7611
HEC	N/A	N/A	0.8612	0.8911	0.8911	0.8911	0.8859
MM ($\sigma=1.5$, O=0.5)	0.85	3.0	0.9160	<u>0.9208</u>	<u>0.9208</u>	<u>0.9208</u>	<u>0.9220</u>
MM ($\sigma=2.0$, O=0.3)	0.65	2.5	<u>0.9125</u>	0.9307	0.9307	0.9307	0.9301
MM ($\sigma=2.0$, O=0.35)	0.70	2.5	0.9067	0.9208	0.9208	0.9208	0.9219
MM ($\sigma=0.0$, O=0.4)	0.70	3.0	0.9009	0.9109	0.9109	0.9109	0.9129
MM ($\sigma=2.0$, O=0.4)	0.75	3.0	0.9009	0.9109	0.9109	0.9109	0.9129
MM ($\sigma=2.0$, O=0.2)	0.50	2.0	0.8822	0.9109	0.9109	0.9109	0.9103
MM ($\sigma=2.0$, O=0)	0.80	4.0	0.7889	0.8317	0.8317	0.8317	0.8240
MM ($\sigma=1.5$, O=0)	0.90	5.0	0.7703	0.8317	0.8317	0.8317	0.8195

This design was chosen deliberately: applying controlled perturbations to ground-truth maps yields a tunable error rate and therefore a continuous characterization of how performance depends on map accuracy, whereas the predictions of a single fixed model provide only one operating point and confound segmentation with classification errors. The perturbation range was calibrated against the error rates measured when running the pre-trained models [12,15] on this dataset, so that the tested interval brackets the accuracy of current nuclei classifiers. Two limitations follow. First, randomly injected errors do not reproduce the structured error patterns of a real upstream model, whose mistakes are correlated with difficult morphology, staining variation, nuclear density and grade, and which may therefore concentrate on precisely the high-grade nuclei that determine the patch label. Second, models were trained on ground-truth maps and perturbed only at evaluation, so the reported degradation characterizes robustness to corruption of a reliable input rather than performance with an imperfect model in the loop at both training and evaluation time. Evaluating the full pipeline with predicted maps in both phases is left for future work. Balanced accuracy is average per-class recall:

$$\text{Balanced Accuracy} = \frac{1}{N} \sum_{i=1}^N \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (3)$$

where N is the number of classes, giving equal importance to all grades regardless of frequency and addressing dataset imbalance [2].

3 Results and Discussion

We evaluated two fusion methods, HEC and MM. Table 2 shows all optimized methods outperform the RGB-only baseline, with MM superior by preserving

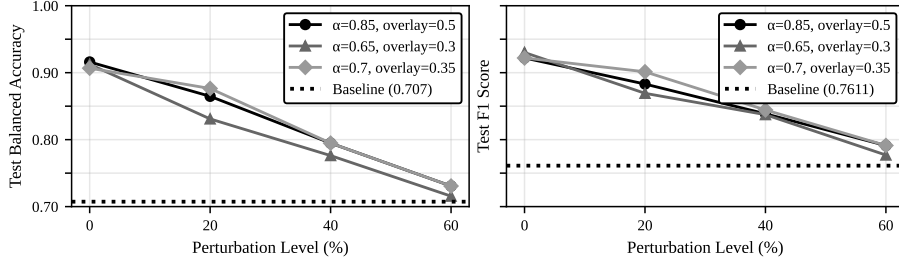


Fig. 3. Perturbation impact on balanced accuracy and F1 score. Performance degrades with increasing simultaneous segmentation and classification errors but remains above RGB-only baselines (0.707 and 0.761, respectively) up to 60% error rate, demonstrating robustness to imperfect nuclei predictions.

gradient information while directly modulating RGB intensities without color deconvolution. The top MM configurations confirm the ranges in Section 2.3: $\alpha = 0.7\text{--}0.85$ was best (lower under-emphasized nuclei, higher over-saturated), $\beta = 2.5\text{--}3.0$ balanced smoothness and separability, $w_3 = 1.85\text{--}2.0$ strengthened the key class without degrading texture, $\sigma = 1.5\text{--}2.0$ softened boundaries (minimal at $\sigma = 0$). Color overlay strength (0.3–0.5) was the most influential parameter, producing perceptually distinct markers against the H&E background; removing it significantly reduced balanced accuracy (Table 2), confirming it as the primary carrier of explicit class-discriminative information, with only spatial cues remaining otherwise (Fig. 2(b)).

Our best method reached 0.916 balanced accuracy and 0.922 F1, versus the RGB-only baseline (0.707, 0.761) and max-voting aggregation (0.427) [12], confirming that max-voting under-grades patches where sparse high-grade nuclei carry clinical significance (Fig. 1).

Table 3. Per-class recall of the best configuration (MM, $\alpha = 0.85$) on the test set.

Grade	Test patches (n)	Correctly classified	Recall
1	114	106	0.930
2	55	50	0.909
3	33	30	0.909
Total	202	186	0.921

Table 3 reports per-class recall for this configuration, showing that performance is not concentrated in the majority class. Classification maps alone (with ground truth) also reach 0.916, but assume perfect classification unavailable in practice. Current nuclei models achieve $\approx 0.64\text{--}0.70$ balanced accuracy [12,15], i.e., 30–36% error; at 30% perturbation our fusion maintains 0.82–0.86, indicating RGB texture provides complementary robustness.

Fig. 3 reports sensitivity results with segmentation and classification errors applied simultaneously. Even at 60% combined perturbation, performance stayed above the RGB-only baseline, demonstrating applicability with existing imperfect models. When semantic modulation is too weak (e.g., $\alpha = 0.3$ with minimal weights), it distorts nuclei RGB appearance without sufficient emphasis and drops below baseline (0.680 vs 0.707), supporting active utilization of nuclei grading rather than general spatial attention. Preliminary experiments examined segmentation and classification errors in selected models [12,15] in isolation. Both types produced similar degradation, with segmentation errors and lower-grade misclassification slightly less impactful than higher-grade misclassification; combined perturbation falls between the isolated cases, making it a practical summary metric, so detailed isolated per-class/per-error analysis is omitted. We attribute this robustness to the ViT leveraging nuclei-derived features alongside RGB texture: the 21 percentage point gain over baseline and monotonic degradation with perturbation both support active semantic utilization. Removing color overlay (overlay=0) eliminates explicit class distinction while preserving segmentation-level spatial information, reducing balanced accuracy to ≈ 0.78 (Table 2). Removing nuclei entirely, leaving only background tissue, eliminates chromatin texture, staining intensity, and nuclear morphology while preserving only size and contour; performance drops substantially below baseline, since WHO/ISUP criteria depend on nuclear characteristics beyond size. Our fusion integrates previously isolated fine-grained classifiers with patch-level grading without architectural modification, enabling practical deployment.

4 Conclusion

This study assessed whether existing imperfect nuclei classification models can improve patch-level grading, and up to which error rate integration remains practically acceptable. Semantic-guided multimodal preprocessing enhances patch-level CCRCC grading without architectural modification. Integrating nuclei maps from pre-trained models with RGB images via multiplicative modulation and perceptually optimized overlays achieved substantial gains over RGB-only processing (0.707) and max-voting aggregation (0.427), the best configuration reaching 0.916 balanced accuracy (Table 2). This lets ViTs exploit both fine-grained nuclei and coarse-grained tissue patterns, evidenced by consistent advantage across all configurations. Sensitivity analysis showed robustness up to 60% combined segmentation and classification perturbation while staying above the RGB-only baseline; current state-of-the-art nuclei models [12,15] exhibit $\approx 30\text{--}36\%$ error, well within this tolerance. Without architectural modification, our approach integrates previously isolated fine-grained classifiers into practical diagnostic pipelines, with further gains possible via architectural adaptation. Because the method requires only a nuclei classification map at the input stage, it can in principle be extended to new, unannotated cohorts using maps produced by an existing nuclei model, with a small annotated subset reviewed by a pathologist to verify map quality and calibrate the expected error rate. Future work will

evaluate the full pipeline with predicted maps at both training and evaluation time, and will test grade-decoupled overlay variants to separate the contribution of the semantic map from that of the colour encoding.

Code Availability. The implementation is publicly available at our GitHub repository.

Acknowledgments. The results published here are in whole or in part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

5 Compliance with Ethical Standards

This study utilized publicly available data from the TCGA Research Network with nuclei annotations from Gao et al. [12]. No additional ethics approval was required.

References

1. Bilal, M., Jewsbury, R., Wang, R., AlGhamdi, H.M., Asif, A., Eastwood, M., Rajpoot, N.: An aggregation of aggregation methods in computational pathology. *Med. Image Anal.* **88**, 102885 (2023)
2. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: *ICPR*. pp. 3121–3124 (2010)
3. Browning, L., Colling, R., Verrill, C.: WHO/ISUP grading of clear cell renal cell carcinoma and papillary renal cell carcinoma; validation of grading on the digital pathology platform and perspectives on reproducibility of grade. *Diagn. Pathol.* **16**, 75 (2021)
4. Campanella, G., Hanna, M.G., Geneslaw, L., Mirafior, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**(8), 1301–1309 (2019)
5. Castro, S., Pereira, V., Silva, R.: Improved segmentation of cellular nuclei using UNet architectures for enhanced pathology imaging. *Electronics* **13**(16), 3335 (2024)
6. Chanchal, A.K., Lal, S., Kumar, R., Kwak, J.T., Kini, J.: A novel dataset and efficient deep learning framework for automated grading of renal cell carcinoma from kidney histopathology images. *Sci. Rep.* **13**, 5728 (2023)
7. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: *CVPR*. pp. 16144–16155 (2022)
8. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F.K., Mahmood, F.: Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: *MICCAI*. vol. 12908, pp. 339–349 (2021)

9. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F.K., Rodig, S.J., Lindeman, N.I., Mahmood, F.: Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **41**(4), 757–770 (2022)
10. Choi, J., Bang, S., Suh, J., Choi, C.I., Song, W., Yuk, H.D., et al.: Survival pattern of metastatic renal cell carcinoma patients according to WHO/ISUP grade: a long-term multi-institutional study. *Sci. Rep.* **14**, 4740 (2024)
11. Delahunt, B., Eble, J.N., Egevad, L., Samaratunga, H.: Grading of renal cell carcinoma. *Histopathology* **74**(1), 4–17 (2019)
12. Gao, Z., Shi, J., Zhang, X., Li, Y., Zhang, H., Wu, J., Wang, C., Meng, D., Li, C.: Nuclei grading of clear cell renal cell carcinoma in histopathological image by composite high-resolution network. In: *MICCAI*. vol. 12908, pp. 132–142 (2021)
13. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019)
14. He, K., Sun, J., Tang, X.: Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(6), 1397–1409 (2013)
15. Javadian, F., Aminparast, Z., Stegmaier, J., Jose, A.: Comparative analysis of unsupervised and supervised autoencoders for nuclei classification in clear cell renal cell carcinoma images. In: *IEEE ISBI*. pp. 1–5 (2025)
16. Le Vuong, T.T., Kim, K., Song, B., Kwak, J.T.: Joint categorical and ordinal learning for cancer grading in pathology images. *Med. Image Anal.* **73**, 102206 (2021)
17. Li, X., Plataniotis, K.N.: A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics. *IEEE Trans. Biomed. Eng.* **62**(7), 1862–1873 (2015)
18. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Shao, L.: Structure preserving stain normalization of histopathology images using self supervised semantic guidance. In: *MICCAI*. vol. 12265, pp. 309–319 (2020)
19. Mercan, C., Balkenhol, M., Salgado, R., Sherman, M., Vielh, P., Vreuls, W., Polónia, A., Horlings, H.M., Weichert, W., Carter, J.M., Bult, P., Christgen, M., Denkert, C., van de Vijver, K., Bokhorst, J.M., van der Laak, J., Ciompi, F.: Deep learning for fully-automated nuclear pleomorphism scoring in breast cancer. *npj Breast Cancer* **8**(1), 120 (2022)
20. Mou, E., Wang, H., Chen, X., Li, Z., Cao, E., Chen, Y., Huang, Z., Pang, Y.: Retinex theory-based nonlinear luminance enhancement and denoising for low-light endoscopic images. *BMC Med. Imaging* **24**(1), 207 (2024)
21. Orlando, J.I., Prokofyeva, E., Blaschko, M.B.: A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE Trans. Biomed. Eng.* **64**(1), 16–27 (2017)
22. Paech, D.C., Weston, A.R., Pavlakis, N., Gill, A., Rajan, N., Barraclough, H., Fitzgerald, B., Van Kooten, M.: A systematic review of the interobserver variability for histology in the differentiation between squamous and nonsquamous non-small cell lung cancer. *J. Thorac. Oncol.* **6**(1), 55–63 (2011)
23. Rasheed, A., Shirazi, S.H., Umar, A.I., Shahzad, M., Yousaf, W., Khan, Z.: Cervical cell’s nucleus segmentation through an improved UNet architecture. *PLoS One* **18**(10), e0283568 (2023)
24. Ruifrok, A.C., Johnston, D.A.: Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **23**(4), 291–299 (2001)

25. Senapati, P., Basu, A., Deb, M., Dhal, K.G.: Sharp dense U-Net: an enhanced dense U-Net architecture for nucleus segmentation. *Int. J. Mach. Learn. Cybern.* **15**(6), 2079–2094 (2024)
26. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In: *NeurIPS*. vol. 34, pp. 2136–2147 (2021)
27. Tabibu, S., Vinod, P.K., Jawahar, C.V.: Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Sci. Rep.* **9**, 10509 (2019)
28. Tang, W., Yang, Z., Song, Y.: Disease-grading networks with ordinal regularization for medical imaging. *Neurocomputing* **545**, 126245 (2023)
29. Warren, A.Y., Harrison, D.: WHO/ISUP classification, grading and pathological staging of renal cell carcinoma: standards and controversies. *World J. Urol.* **36**(12), 1913–1926 (2018)
30. Xu, J., Janowczyk, A., Chandran, S., Madabhushi, A.: A high-throughput active contour scheme for segmentation of histopathological imagery. *Med. Image Anal.* **15**(6), 851–862 (2011)
31. Zhou, Y., Dou, Q., Chen, H., Qin, J., Heng, P.A.: SFCN-OPI: Detection and fine-grained classification of nuclei using sibling fc with objectness prior interaction. In: *AAAI*. pp. 2652–2659 (2018)