Cross-Scale Path Aggregation Transformer for Joint Pulmonary Lesion Segmentation and Gene Expression Pattern Prediction in Precision Oncology

Vikram A. Batra
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
vikramabatra120@uab.edu

Jan L. Bush
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
bushjan@colostate.edu

Abstract

Precision oncology increasingly relies on the integration of radiological phenotypes and molecular signatures to guide individualized treatment. While deep learning has achieved remarkable performance in pulmonary lesion segmentation and in predicting gene expression patterns from medical images, these tasks are typically addressed in isolation, leaving unexploited synergistic representations that could enhance both anatomical delineation and molecular characterization. This paper presents a system-level analysis of a novel architecture that jointly performs pulmonary lesion segmentation and gene expression pattern prediction through a cross-scale path aggregation transformer. The design couples a hierarchical vision transformer backbone with a bidirectional path aggregation mechanism that fuses multi-scale feature maps without resorting to simplistic skip connections, enabling simultaneous refinement of fine-grained boundary information and high-level semantic abstractions. A dual-head decoder produces a dense segmentation mask and a vector of predicted expression levels for clinically relevant oncogenes. We examine the structural trade-offs inherent in multi-task training, including gradient interference, loss weighting strategies, and latent representation entanglement, and we articulate how the cross-scale aggregation reduces representation misalignment across tasks. Beyond model architecture, we discuss deployment considerations such as computational footprint, robustness under domain shift, federated learning for privacy-preserving multi-institutional collaboration, and alignment with regulatory frameworks. By situating the technical contribution within a broader socio-technical infrastructure, we address fairness, interpretability, and sustainability requirements. The discussion offers a forward-looking perspective on how tightly coupled imaging–genomic models can be operationalized in clinical workflows while maintaining safety, equity, and governance standards.

Keywords

cross-scale aggregation, transformer, pulmonary lesion segmentation, gene expression prediction, precision oncology, path aggregation, joint learning.

1. Introduction

Lung cancer remains a leading cause of cancer mortality worldwide, driving an urgent need for tools that can simultaneously resolve anatomical abnormalities and characterize their

molecular underpinnings. Radiological imaging, particularly computed tomography (CT), provides non-invasive structural information, while gene expression profiling offers insight into tumor biology and potential therapeutic targets. The convergence of these modalities has given rise to radiogenomics, wherein quantitative image features are linked to molecular phenotypes [1]. Despite progress, clinical decision-making still largely treats image analysis and molecular testing as independent, sequential processes, delaying integrated interpretation. A unified computational framework that jointly segments pulmonary lesions and predicts gene expression patterns directly from imaging could accelerate diagnostic workflows and enable more timely precision oncology interventions.

Deep convolutional neural networks have become the cornerstone of medical image segmentation, with architectures such as U-Net providing robust encoder–decoder pathways [2] and attention mechanisms enhancing the saliency of relevant structures [3]. The introduction of hierarchical vision transformers, exemplified by the Swin Transformer [4], has further expanded the capacity to model long-range dependencies and multi-scale representations. In parallel, task-specific advances have yielded segmentation models tailored to pulmonary nodules, including designs that employ path aggregation and dual attention to improve boundary fidelity [5] and transformer-based frameworks that recast segmentation as a sequence-to-sequence prediction problem [6]. Concurrently, the feasibility of inferring molecular characteristics from imaging has been demonstrated through radiogenomic mapping studies that associate CT-derived features with transcriptomic signatures [9]. More recently, transformer-augmented survival models have been applied to forecast adverse events in clinical trials, underscoring the versatility of attention mechanisms for structured health data [7].

However, existing architectures share a common limitation: they are designed for either segmentation or molecular prediction, not both. Joint modeling of these tasks could leverage shared inductive biases, but it introduces complex system-level challenges. The representations beneficial for precise boundary delineation may partially conflict with those suited for capturing global tumor-level features that correlate with gene expression. Moreover, aggregating information across spatial scales without introducing distortion or feature misalignment is non-trivial. To address these difficulties, we investigate a cross-scale path aggregation transformer that fuses features from different hierarchical stages through a bidirectional aggregation neck, distributing refined representations to a dual-task decoder head. The resulting system treats lesion morphology and molecular pattern inference as mutually informative supervisory signals, potentially improving robustness and interpretability.

This paper adopts a systems-oriented perspective, emphasizing the architectural rationale, governance mechanisms, and deployment implications of such an integrated model. Rather than presenting isolated accuracy metrics, we delve into the design trade-offs that govern multi-task learning, scalability, fairness, regulatory compliance, and sustainability. By doing so, we aim to provide a comprehensive framework for translating this joint learning paradigm into trustworthy clinical AI infrastructure.

2. Related Work and System Design Rationale

The landscape of medical image segmentation has been shaped by successive waves of methodological innovation. The foundational U-Net architecture [2] established the effectiveness of symmetric encoder–decoder structures with long skip connections, enabling precise localization while preserving context. Attention-augmented variants [3] improved

upon this by learning to focus on salient regions, mitigating the influence of irrelevant background texture. More recently, nnU-Net demonstrated that systematic hyperparameter configuration could achieve state-of-the-art performance across a wide array of biomedical segmentation tasks without task-specific manual tuning [4]. These methods, however, rely predominantly on convolutional operations with limited receptive fields, which can struggle to incorporate global anatomical context essential for distinguishing lesion types or predicting systemic molecular phenotypes.

Vision transformers have emerged as a compelling alternative. The Swin Transformer [4] introduces a hierarchical architecture with shifted window attention, generating multi-scale feature maps that naturally align with the needs of dense prediction tasks. This shift from pure convolutional backbones to transformer encoders enables stronger context aggregation and more flexible cross-scale interactions. Early adaptations for medical segmentation, such as TransUNet [6], combined a transformer encoder with a convolutional decoder, illustrating the benefits of hybrid designs. For lung nodules specifically, the PDU-Net approach integrated path aggregation and dual attention to refine segmentation boundaries in CT scans [5], highlighting the utility of explicit multi-scale feature fusion strategies.

Parallel to segmentation advances, radiogenomics has matured into a recognized research paradigm. Publicly available datasets such as the NSCLC Radiogenomics collection [9] link CT volumes with matched gene expression profiles, enabling the development of models that predict oncogene status from imaging. Early efforts used handcrafted radiomic features; deep learning methods have since been employed to learn predictive representations end-to-end. The task of forecasting adverse events in clinical trials using transformer-augmented survival analysis [7] further exemplifies how attention architectures can handle time-to-event outcomes and structured covariates that share conceptual ground with gene expression pattern prediction.

Despite these developments, the vast majority of architectures are optimized for a single output modality. Multi-task learning offers a principled way to share representations across related tasks, potentially improving each task through shared inductive biases and increased effective sample size. In computational pathology and radiology, joint learning has shown promise for tasks such as simultaneous segmentation and classification. Our work extends this idea to the pairing of lesion segmentation and gene expression prediction, two tasks that differ in granularity—pixel-wise versus patient-level—yet share common imaging features. The central challenge is designing an architecture that can maintain high-resolution detail for accurate lesion boundaries while also consolidating global semantic information into a compact molecular signature.

Path aggregation networks, originally developed for instance segmentation in natural images [8], offer a solution. By establishing both top-down and bottom-up pathways between multi-scale feature layers, a path aggregation module enables information to flow bidirectionally, enriching low-level features with semantic guidance and high-level features with spatial precision. When integrated with a transformer backbone, this mechanism can alleviate the representation misalignment that often arises when simply concatenating skip connections. The cross-scale path aggregation transformer thus brings together hierarchical self-attention, bidirectional feature fusion, and dual-task supervision into a unified system.

3. Proposed Architecture and Cross-Scale Aggregation Mechanism

The overall system receives a volumetric CT scan, preprocessed into axial slices or a sub-volume centered on a suspected lesion. Input slices pass through a hierarchical vision transformer that produces feature representations at four spatial scales, mirroring the pyramidal structure of convolutional backbones. Each stage applies shifted window self-attention, progressively reducing spatial resolution while increasing channel dimensionality. This design captures both local texture cues critical for delineating lesion borders and long-range dependencies that may correlate with gene expression signatures.

The cross-scale path aggregation module forms the conceptual core of the architecture. Unlike standard skip connections that simply concatenate encoder feature maps with decoder paths, we adopt a bidirectional aggregation neck inspired by path aggregation concepts [8] but adapted to transformer-derived tensors. A top-down pathway propagates rich semantic information from the deepest transformer layers to earlier, higher-resolution features through lateral connections, while a subsequent bottom-up pathway further refines these enriched feature maps by reintroducing fine spatial details. At each aggregation node, a channel–spatial attention gate selectively modulates the contribution of different scales, suppressing noise and emphasizing lesion-relevant patterns. This gating mechanism draws conceptually from dual attention formulations [5], extended here to operate within the transformer feature space and across non-adjacent scales.

The bidirectional aggregation yields a multi-scale feature pyramid in which every level has been exposed to both global context and local detail. This pyramid is then routed to two task-specific heads. The segmentation head consists of a lightweight upsampling decoder that progressively fuses the aggregated feature pyramid to produce a pixel-wise lesion probability map. The design avoids excessive parameters, recognizing that the richness of the aggregated features reduces the burden on the decoder. The gene expression prediction head globally pools the aggregated pyramid, applying a channel-wise attention block that learns which feature scales and channels are most informative for gene expression. The resulting compact descriptor is passed through a fully connected regressor that outputs predicted expression levels for a predefined panel of oncogenes, such as EGFR, KRAS, and ALK.

Training uses a weighted multi-task loss combining a segmentation term, typically a combination of Dice loss and focal cross-entropy, with a gene expression regression term, such as mean squared error or correlation-based loss. A critical system-level choice is how to balance these losses. Joint optimization can suffer from conflicting gradient directions, where improving segmentation performance harms gene expression prediction and vice versa. We adopt homoscedastic task uncertainty weighting [10], which learns task-specific observation noise parameters that dynamically adjust the contribution of each loss during training. This approach has been shown to stabilize multi-task training and reduce the need for manual tuning. In settings where gene expression labels are scarce, the framework can be extended to semi-supervised regimes by incorporating consistency regularization on unlabeled imaging data, leveraging the segmentation task as a source of inductive structure.

4. System-Level Trade-Offs and Representation Entanglement

Designing a model that faithfully performs joint dense prediction and patient-level regression involves navigating several structural tensions. One key trade-off concerns capacity allocation. The shared backbone and aggregation module must encode features that serve two disparate objectives: precise spatial localization for segmentation and holistic abstraction for gene expression inference. Over-compressing spatial detail risks degradation of lesion boundary accuracy, while preserving overly fine-grained features may introduce noise that confounds

the holistic molecular predictor. The cross-scale path aggregation mechanism mitigates this tension by allowing the downstream heads to selectively attend to the scale most relevant to their task—lower-level feature maps for segmentation, higher-level for gene expression—while still enabling cross-task feature reuse.

The entanglement of representations also has implications for interpretability. Clinicians need to understand the basis for both the segmented lesion contour and the predicted molecular profile. Because the dual heads share a common feature pyramid, attribution methods such as gradient-weighted class activation mapping [13] can highlight which image regions contribute to gene expression predictions, potentially revealing radiogenomically salient zones that align with known tumor habitats. However, this shared attribution can also conflate evidence, making it harder to disentangle whether a high EGFR prediction stems from the segmented lesion core or from peritumoral tissue. From a governance perspective, this necessitates careful documentation and validation in accordance with the FDA's proposed framework for adaptive AI/ML-based software as a medical device [14], ensuring that the system's outputs can be audited and explained.

Robustness to domain shift constitutes another critical system property. When deployed across different scanner types, acquisition protocols, or patient populations, the model must maintain stable performance. Dropping the requirement for paired imaging–genomic training data in every deployment setting, federated learning architectures [11] can enable collaborative model refinement without centralized data pooling, preserving patient privacy under regulations such as HIPAA and GDPR [15]. Simultaneously, lightweight compression techniques such as quantization and pruning [16] can reduce the model's computational footprint to facilitate edge deployment on clinical workstations, though pruning must be performed carefully to avoid exacerbating biases against demographic subgroups underrepresented in training data.

Furthermore, the sustainability of training large transformer-based models warrants scrutiny. The energy consumption associated with self-attention over high-resolution 3D volumes can be substantial, motivating the development of efficient attention variants and the use of mixed-precision training [17]. From a policy standpoint, funding agencies and healthcare institutions increasingly require environmental impact assessments for AI deployments, making energy-aware architectural choices a governance imperative as much as a technical optimization.

5. Integration into Clinical and Regulatory Infrastructure

Operationalizing a joint segmentation–gene expression model in real-world oncology workflows demands more than high accuracy. The model must interface seamlessly with picture archiving and communication systems, radiology information systems, and electronic health records. This requires standardized input–output protocols such as DICOM and FHIR, as well as robust data pre-processing pipelines that handle variations in slice thickness, reconstruction kernels, and contrast timing. The output should be presented in a clinically intuitive format—a segmentation overlay on the CT study with an associated gene expression likelihood panel—allowing the oncology team to integrate the findings with histopathological and liquid biopsy results.

Regulatory classification of such a model adds further complexity. Because the system provides both quantitative image analysis and molecular predictions that may influence treatment decisions, regulators are likely to consider it a medical device under frameworks for

software as a medical device (SaMD). The FDA's evolving approach to AI/ML-based SaMD emphasizes the need for predetermined change control plans and real-world performance monitoring [14]. A model that learns continuously after deployment would require additional oversight, including periodic re-validation against reference standards and monitoring for performance drift across demographic groups. The joint nature of the output compounds this challenge: a degradation in segmentation fidelity could introduce cascading errors in gene expression prediction, or vice versa, necessitating multi-faceted quality assurance protocols.

Fairness is a non-negotiable dimension of this infrastructure. Studies have demonstrated that clinical AI systems can exhibit racial and socioeconomic biases when trained on non-representative data [12]. In lung cancer, smoking history, environmental exposures, and genetic ancestry intersect in complex ways that influence both lesion appearance and molecular profiles. Ensuring equitable performance requires stratified evaluation across self-reported race, ethnicity, sex, and age, as well as deliberate oversampling or reweighting strategies during training. The model's reliance on imaging features that may correlate with body habitus or scan quality must be audited to avoid indirect discrimination. Institutional governance boards, including ethics committees and data access committees, should oversee the deployment pipeline from data collection to model serving, enforcing transparency and accountability.

6. Evaluation Framework and Sustainability Considerations

Comprehensive evaluation must go beyond reporting overall Dice scores or mean absolute gene expression error. We advocate a multi-dimensional benchmarking protocol that assesses segmentation accuracy using overlap and distance metrics on public datasets such as the LUNA16 challenge for nodule detection [18] and the NSCLC Radiogenomics collection [9] for paired imaging–genomic data. Gene expression prediction can be evaluated using Pearson correlation coefficients and concordance indices between predicted and measured expression levels. Ablation studies should isolate the contributions of the cross-scale path aggregation module, the bidirectional attention gates, and the multi-task learning formulation, comparing performance against strong baselines that treat segmentation and gene expression prediction separately.

Robustness testing, including adversarial evaluation with synthetic noise, contrast perturbations, and simulated low-dose CT protocols, is essential to gauge real-world reliability. Cross-institutional validation, whether through federated simulation or actual multi-site trials, can expose vulnerabilities to scanner-specific biases. Moreover, longitudinal monitoring protocols must be established so that if the model is updated or retrained, its behavior against historic cases is systematically reviewed.

Sustainability encompasses both environmental and economic dimensions. The energy and carbon footprint of model training can be reported using established tools, incentivizing the adoption of sparse attention mechanisms and model distillation. From a health economics perspective, the cost-effectiveness of adding gene expression prediction to routine CT analysis must be compared against standalone molecular testing. A systems-level analysis should quantify potential savings from earlier identification of actionable mutations, reduced biopsy rates, and more precise therapy selection, balanced against the computational infrastructure investments.

7. Conclusion

The cross-scale path aggregation transformer represents a step toward tightly integrating anatomical delineation with molecular inference within a single computational framework. By uniting a hierarchical vision transformer backbone with bidirectional feature aggregation and dual-task learning, the architecture addresses several longstanding tensions in medical image analysis: the need for both fine-grained spatial precision and high-level semantic abstraction, the challenge of task interference in multi-task settings, and the requirement for representations that are both robust and interpretable. Our system-level discussion has underscored that such a model cannot be evaluated solely on accuracy benchmarks. Its value to precision oncology will be determined by how it interfaces with clinical workflows, respects privacy and fairness, adheres to regulatory standards, and operates sustainably within healthcare ecosystems. Future work should focus on closed-loop clinical trials that measure the impact of joint imaging–genomic AI on treatment decisions and patient outcomes, while continuing to advance model architectures that are simultaneously powerful, efficient, and equitable.

References

1. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., ... & Aerts, H. J. (2017). Radiomics: The bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology, 14(12), 749–762.

2. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234–241). Springer.

3. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

4. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012–10022).

5. Chang, C., Fu, M., Chen, X., Feng, S., Zhang, M., Zhou, X., ... & Liu, Z. (2025, November). Research on PDU-Net Lung Nodule Segmentation Algorithm Based on Path Aggregation and Dual Attention. In 2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 1897-1900). IEEE.

6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

7. Wang, Y. (2025, April). Efficient adverse event forecasting in clinical trials via transformer-augmented survival analysis. In Proceedings of the 2025 International Symposium on Bioinformatics and Computational Biology (pp. 92-97).

8. Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8759–8768).

9. Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., ... & Napel, S. (2018). A radiogenomic dataset of non-small cell lung cancer. Scientific Data, 5, 180202.

10. Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7482–7491).

11. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., ... & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), 12598.

12. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

13. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).

14. U.S. Food and Drug Administration. (2019). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). Discussion paper.

15. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43.

16. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In International Conference on Learning Representations.

17. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

18. Setio, A. A. A., Traverso, A., de Bel, T., Berens, M. S., Bogaard, C. v. d., Cerello, P., ... & van Ginneken, B. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical Image Analysis, 42, 1–13.