

Cross-Modal Contrastive Learning for the Retrieval of Immunotherapy-Associated Molecular Signatures from Histopathology

Sigrid Vila-Bagaria¹, Mar Teixidó², Miquel Piñol², Felip Vilardell²,
Robert Montal², and Veronica Vilaplana¹

¹ Signal Theory and Communications Department (TSC)
Universitat Politècnica de Catalunya - BarcelonaTech (UPC)
{sigrid.vila,veronica.vilaplana}@upc.edu

² Cancer Biomarkers Research Group (GReBiC) - IRB Lleida
{mteixido,rmontal}@irbllleida.cat

Abstract. Gastric Adenocarcinoma is a leading cause of cancer mortality. Although “Inflamed/Non-Inflamed” subtypes have been proposed to predict immunotherapy response, their identification relies on a costly 10-gene RNA signature. We propose a Cross-modal Contrastive Multiple Instance Learning (CCMIL) framework for cross-modal retrieval, imputing these molecular signatures directly from standard Hematoxylin & Eosin (H&E) slides. By leveraging a supervised contrastive objective, CCMIL aligns visual morphological patterns with molecular phenotypes into a shared latent space. This establishes an interpretable search-by-case retrieval engine, enabling pathologists to query a whole slide image to surface transcriptomically coherent neighbors and approximate RNA signatures without genomic sequencing at inference. Our results demonstrate that this retrieval-first approach captures the continuous phenotypic spectrum of tumor inflammation and yields clinically interpretable attention heatmaps. Furthermore, the learned representation also supports competitive downstream classification, providing a practical molecular pre-screening strategy.

Keywords: Cross-Modal Retrieval · Contrastive Learning · Gastric Cancer · Molecular Subtyping · Multiple Instance Learning.

1 Introduction

Gastric adenocarcinoma (GAC) is a leading cause of cancer mortality globally [22], typically diagnosed at advanced stages with poor prognosis [1]. While immune checkpoint inhibitors (ICIs) have become a standard of care [8,9,18], heterogeneous responses demand better predictive biomarkers [2]. Recently, a distinct “Inflamed/Non-Inflamed” GAC phenotype, characterized by high immunogenicity, was identified as a strong predictor of ICI response [19]. However, determining this class requires a costly 10-gene RNA signature, limiting its routine clinical use. By learning to impute this molecular signature directly from

low-cost, standard H&E features, our framework bypasses the need for genomic sequencing at inference, providing a scalable pre-screening tool.

Deep learning (DL) in computational pathology offers an accessible alternative to molecular assays [15]. However, learning the “Inflamed/Non-Inflamed” phenotype directly from H&E faces two key challenges. First, paired histology-genomic datasets are scarce, increasing the risk of overfitting. Second, binarizing a continuous gene expression signature into a label introduces noise and discards the patient’s true position along the biological phenotypic spectrum. To address these challenges, we investigate metric learning to build a highly interpretable search-by-case diagnostic tool. The contributions of this work are as follows.

- We propose a Cross-modal Contrastive Multiple Instance Learning (CCMIL) framework that learns a shared histology-transcriptomics embedding.
- We demonstrate that the learned representation captures biologically coherent relationships between patients, enabling interpretable RNA signature imputation and competitive downstream classification from H&E alone.

2 Related Work

Contrastive learning in computational pathology: Contrastive learning optimizes latent spaces by pulling positive pairs together and pushing negative pairs apart [5]. Frameworks like ConVIRT [26] and CLIP [17] proved this highly effective for cross-modal retrieval in the vision-language domain. Medical adaptations, such as MedCLIP [23], apply these principles to healthcare, maximizing similarity between matched modalities (e.g., histology and genomics) to capture patient-specific biological associations. Crucially, establishing this tightly clustered joint manifold provides the foundational architecture required for robust retrieval tasks.

Cross-modal retrieval in computational pathology: Accurate tumor subtyping is fundamental to precision oncology, yet unimodal MIL frameworks (e.g., ABMIL [6], CLAM [15], TransMIL [20]) often fail to capture the full transcriptomic complexity of heterogeneous tumors. To address this, multimodal frameworks like MCAT [3] fuse histological and genomic data, but strictly require both modalities at inference, maintaining high sequencing costs. Emerging approaches instead leverage contrastive learning to align modalities [12], integrating spatial transcriptomics [16] or enforcing semantic consistency [24] to enhance robustness. Building upon these advances, our framework exploits this aligned multimodal manifold specifically to enable cross-modal retrieval, allowing clinicians to impute molecular signatures using only histology at inference.

3 Method

3.1 Cross-modal contrastive architecture

The framework, as shown in Figure 1a, consists of two parallel encoding streams that map histological and genomic data to a shared joint embedding space.

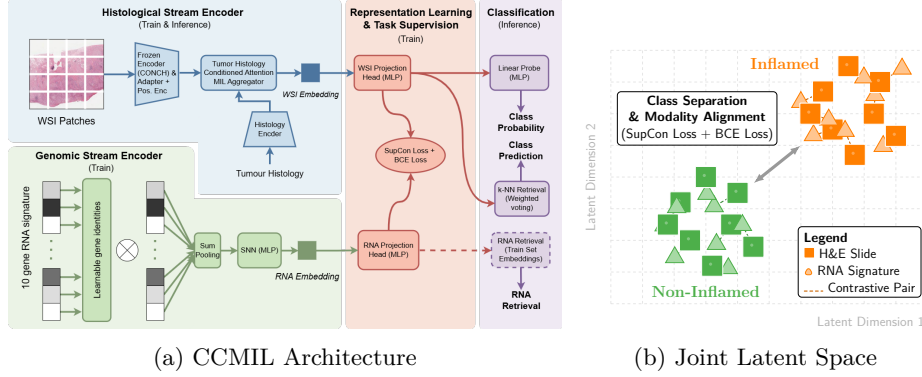


Fig. 1: The CCMIL framework. (a) Dual-stream architecture aligning WSI and RNA features during training to enable inference from H&E alone. (b) The composite objective aligns modalities and separates “Inflamed/Non-Inflamed” cases in the joint latent space.

Histological stream: Visual features extracted from the frozen CONCH backbone [13] are refined by a visual adapter (a 2-layer MLP with layer normalization and ReLU) into a task-specific latent space. To preserve spatial structure within the permutation-invariant MIL setting, we employ Fourier Positional Encodings, mapping normalized patch coordinates to a high-dimensional frequency domain. Next, because different tumor histologies present distinct regions of interest, we compute the final slide-level embedding using our Tumor Histology Conditioned Attention (THCA), a lightweight gated-attention MIL module that conditions each patch’s attention score on a learned histology-type embedding, yielding a histology-weighted sum of the adapted features. THCA guides the network to learn histology-specific attention profiles. Finally, an MLP projects these features into a shared space with the RNA embeddings, while a concurrently trained probe (2-layer MLP) enables direct WSI classification at inference.

Genomic stream: We encode the 10-gene expression signature into a latent representation utilizing a permutation-invariant embedding scheme followed by a Self-Normalizing Neural Network (SNN) [10]. We assign a learnable identity vector to each gene. The initial aggregation is computed as the expression-weighted sum of these identities, since it learns a gene-specific semantic space while preserving permutation invariance. The aggregated vector is refined by an SNN composed of linear layers with Scaled Exponential Linear Units (SELU) and Alpha Dropout. Finally, similar to the histological stream, the features are projected to a common space with the RNA features using an MLP.

Table 1: Quantitative performance comparison of the proposed framework against state-of-the-art baselines. All values are reported as mean $\pm$ standard deviation across 5-fold cross-validation.

Retrieval						
Model	Q $\rightarrow$ Tgt	Retrieval (Image $\rightarrow$ RNA)			k-NN Classification	
		R@1	R@5	MRR	Acc.	F1
Baseline	Te $\rightarrow$ Tr	–	–	.71 $\pm$.106	.57 $\pm$.129	.53 $\pm$.109
Baseline	Te $\rightarrow$ Te	.67 $\pm$.115	.89 $\pm$.086	.78 $\pm$.067	.67 $\pm$.107	.65 $\pm$.074
CCMIL	Te $\rightarrow$ Tr	–	–	.75 $\pm$ 0.049	.68 $\pm$.080	.61 $\pm$.138
CCMIL	Te $\rightarrow$ Te	.72 $\pm$.085	.92 $\pm$.051	.80 $\pm$.067	.71 $\pm$.155	.67 $\pm$.133

Classification						
Modality	Model	Acc.	AUC	F1	Prec.	Rec.
Unimodal	CLAM	.73 $\pm$.035	.79 $\pm$.034	.72 $\pm$.051	.68 $\pm$.031	.76 $\pm$.065
	ABMIL	.72 $\pm$.021	.79 $\pm$.078	.70 $\pm$.022	.68 $\pm$.048	.72 $\pm$.073
	TransMIL	.52 $\pm$.117	.72 $\pm$.082	.62 $\pm$ 0.027	.51 $\pm$ 0.104	.85 $\pm$.191
Multimod.	MCAT	.66 $\pm$ 0.065	.68 $\pm$.072	.66 $\pm$.101	.61 $\pm$.076	.72 $\pm$.231
Cross-mod.	CCMIL	.72 $\pm$.157	.75 $\pm$.135	.76 $\pm$.112	.68 $\pm$.117	.92 $\pm$.083

3.2 Losses

To address label ambiguity and data scarcity, we optimize a composite objective to promote the latent space to be structured by molecular phenotype and separable for downstream classification.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{SupCon}} \quad (1)$$

Representation Learning (Supervised Contrastive Loss): We use a multi-modal Supervised Contrastive Loss ($\mathcal{L}_{\text{SupCon}}$) over the joint representations. Within a mini-batch of WSI and RNA embeddings, all features are first L_2 normalized and concatenated into a single set I of size $2N$ (where N is the batch size). For any anchor $i \in I$ with embedding z_i and discrete label y_i , the set of positive keys $P(i)$ consists of all other samples in the batch sharing the same label ($p \in I \setminus \{i\} : y_p = y_i$). Because the WSI and RNA embeddings from the same patient inherently share the same label, this formulation naturally groups both *instance-level cross-modal pairs* and *class-level pairs* (any intra- or inter-modal embeddings with the same phenotype). The loss is formulated as:

$$\mathcal{L}_{\text{SupCon}} = \frac{1}{2N} \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in I \setminus \{i\}} \exp(z_i \cdot z_a / \tau)} \quad (2)$$

where $z_i \cdot z_p$ represents the cosine similarity between the embeddings, and the temperature parameter is $\tau = 0.07$.

Table 2: Ablation study of the framework evaluating the Tumour Histology Conditioned Attention (THCA) and loss combinations for Classification and Retrieval.

Components			Classification			Retrieval	
THCA	SupConL	ClassL	Acc.	AUC	F1	R@1	R@5
✓	✓	✗	.54 ± .092	.45 ± .156	.66 ± .069	.69 ± .068	.88 ± .068
✓	✗	✓	.68 ± .103	.68 ± .073	.72 ± .076	.63 ± .119	.84 ± .068
✗	✓	✓	.72 ± .106	.74 ± .084	.74 ± .080	.60 ± .069	.84 ± .107
✓	✓	✓	.72 ± .157	.75 ± .135	.76 ± .112	.72 ± .085	.92 ± .051

Task Supervision (Classification Loss): To enforce separability and refine the decision boundaries of the shared projection manifold, we apply a classification probe to both the aggregated WSI features and the genomic RNA features. The total classification loss is computed as the unweighted average of the Binary Cross-Entropy (BCE) losses from both modalities. This equal weighting encourages that both the histological and genomic representations independently maintain high diagnostic utility and are mapped correctly with respect to the target task’s decision boundary.

3.3 Inference strategy

At inference, the genomic stream is detached, enabling unimodal operation on standard H&E WSIs to bypass sequencing costs while retaining multimodal benefits. The resulting embedding can be exploited in two complementary ways.

Direct Classification: The pre-trained probe projects the histological slide-level embedding onto the decision boundary to output a probability score.

Retrieval Strategy: Given a query H&E slide, its embeddings are projected into the shared latent space and used to retrieve the most similar RNA embeddings from the reference dataset. We evaluated several cross-modal strategies: Inverted Softmax [21], α -CSLS [11], cosine k -NN, Diffusion [7], and AQE [4]. We found CSLS combined with Reciprocal reranking to provide the best retrieval performance, mitigating hubness and refining cross-domain alignment. By retrieving the top- k ($k = 5$) nearest neighbors using this strategy we can aggregate their corresponding RNA profiles to reliably impute a continuous molecular signature for the queried slide.

4 Experimental setup

4.1 Dataset and preprocessing

We utilized a curated subset of the TCGA Gastric Adenocarcinoma cohort comprising $N = 265$ diagnostic WSIs, manually reviewed by expert pathologists to

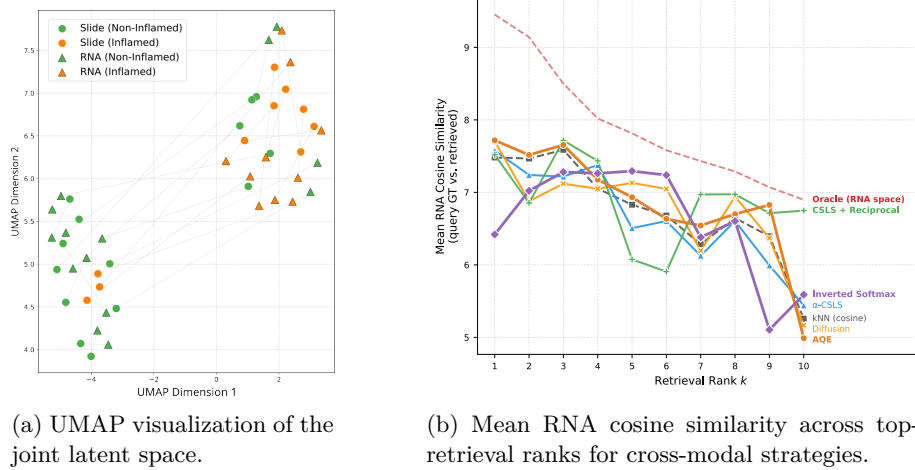


Fig. 2: Evaluation of the proposed framework. **(a)** Latent space visualization via UMAP demonstrating a structured joint representation. **(b)** Mean RNA Cosine Similarity between the query ground truth and retrieved cases across retrieval rank k . CSLS + Reciprocal strategy remains closest to the Oracle (retrieval in the RNA space) upper bound.

ensure optimal tissue quality. The dataset consists of 142 “Non-Inflamed” and 123 “Inflamed” cases, with ground truth labels derived from a 10-gene RNA signature (*ADORA2A*, *BLM*, *CCL20*, *CD3E*, *CD5*, *CD74*, *GLS*, *HLA-E*, *ITGAL*, *RPL23*) [19]. Additionally, cases were stratified by expert pathologists according to the histology Lauren classification into 173 Intestinal, 65 Diffuse, and 16 Mixed subtypes. WSI processing was standardized using the Trident [25] framework for tissue segmentation and non-overlapping patching at $40\times$. Finally, patch-level embeddings were extracted using the CONCH [14] foundation model.

4.2 Training setup

Framework and hardware All experiments were implemented using the PyTorch framework, leveraging the pytorch-metric-learning library for loss formulations. Training was accelerated on a single NVIDIA A40 GPU (48GB VRAM). Due to the computational overhead of processing dense WSI feature bags, we utilized a batch size of $B = 8$ patients with gradient accumulation of 4 steps, which was the maximum capacity allowed by memory constraints.

Optimization The network was trained for 50 epochs per fold using a 5-fold cross-validation scheme with early-stopping. Each fold was composed of a train, validation and test sets of 80%, 10% and 10% of the cohort respectively. We employed the Adam optimizer with a fixed learning rate of 5×10^{-5} and the loss defined in Section 3.2.

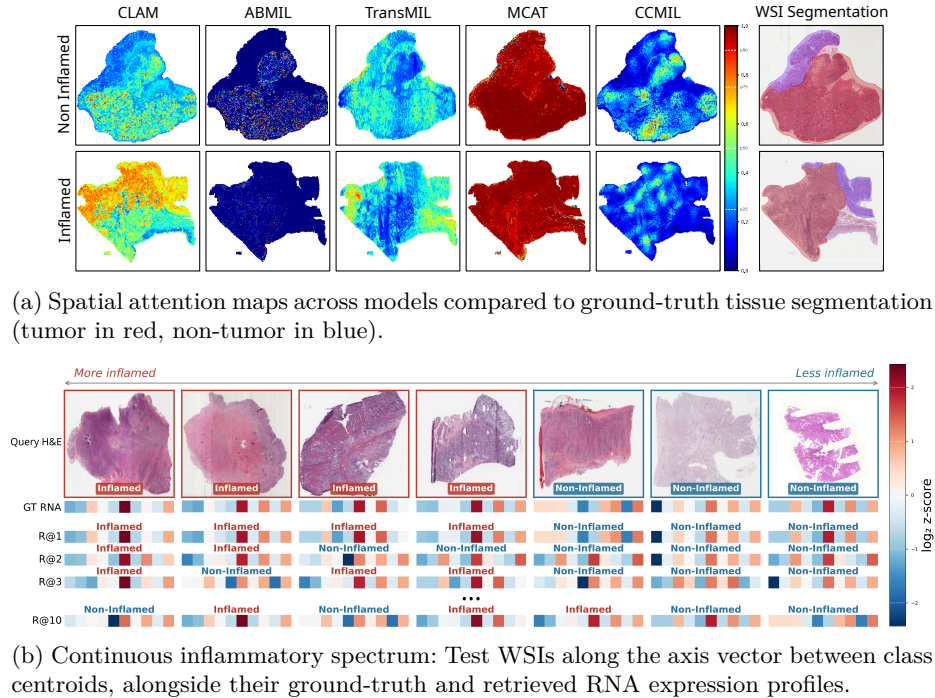


Fig. 3: Qualitative evaluation of CCMIL. **(a)** CCMIL generates dense, highly localized attention hotspots compared to the baseline models. **(b)** The joint latent space captures a continuous inflammatory gradient, successfully retrieving transcriptomically coherent neighbors.

Baselines We compare CCMIL against representative state-of-the-art unimodal (H&E-only) and multimodal (H&E + RNA) approaches. We evaluate three widely used unimodal MIL models: ABMIL [6], CLAM-SB [15], and TransMIL [20]. For multimodal fusion, we compare against MCAT [3], which explicitly co-attends and fuses WSI and genomic features. For fair comparison, all models utilize identical CONCH embeddings and are evaluated across the same 5-fold splits. To isolate the contribution of our class-aware supervised contrastive framework, we evaluate a retrieval baseline that utilizes the same spatial architecture but relies solely on instance-level alignment (CLIP loss [17]). This setup mirrors standard multimodal contrastive methods widely used in the field.

5 Results

5.1 Quantitative Evaluation

We first select the retrieval strategy by evaluating multiple metrics on the validation set. As demonstrated in Figure 2b, CSLS + Reciprocal reranking mitigates

high-dimensional hubness and maintains the highest RNA cosine similarity to the Oracle upper bound for reliable signature imputation. Using this retrieval strategy, the 5-fold cross-validation results (Table 1) demonstrate the effectiveness of CCMIL as a cross-modal retrieval engine. We report performance across two retrieval settings: Test $\rightarrow$ Train, mimicking real-world clinical inference by querying a new test slide against an established set of training cases, and Test $\rightarrow$ Test, which retrieves entirely within the test set to accurately compute R@1 and R@5 scores (percentage of queries that retrieve a match in the top k results). Compared with the CLIP baseline, our class-aware formulation achieves a Test $\rightarrow$ Test R@5 of 0.92 and MRR of 0.80. Aggregating these retrieved WSI via weighted k -NN voting in the Test $\rightarrow$ Train setting yields a classification Accuracy of 0.72 and an F1-score of 0.76.

This learned latent space also enables effective downstream classification. CLAM achieves comparable Accuracy (0.73) but lacks clinical balance, while TransMIL and multimodal MCAT yield poor Precision (0.51, 0.61) due to domain gaps. Conversely, CCMIL achieves the highest F1-score (0.76) and Recall (0.92) indicating improved sensitivity for ICI pre-screening. Our ablation study (Table 2) confirms our architectural choices: representation learning alone yields poor decision boundaries (Accuracy 0.54), omitting the contrastive signal degrades retrieval (R@1 0.63), and replacing THCA with standard attention impairs both fine-grained cross-modal retrieval (R@1 0.60) and overall F1-score (0.74).

5.2 Qualitative Evaluation

Spatial attention heatmaps (Fig. 3a) highlight the interpretability of our framework. Compared to the baseline methods, which exhibit sparse under-activation (ABMIL), severe attention saturation (MCAT), or diffuse noise (CLAM), CCMIL generates dense, localized hotspots that align closely with ground-truth tumor segmentations. Because the “Inflamed” signature is visually indistinguishable to pathologists, these biologically coherent activations suggest CCMIL captures subtle morphological patterns associated with the “Inflamed” phenotype.

Furthermore, we show that the embedding space captures a continuous phenotype (Fig. 3b). By projecting test slides onto an inflammation axis (the vector between class centroids), we selected seven evenly-spaced query slides. Displaying these alongside their retrieved RNA expression profiles (via Slide $\rightarrow$ RNA CSLS+Reciprocal reranking at ranks 1, 2, 3, and 10) confirms that CCMIL organizes slides along a biologically meaningful gradient, surfacing transcriptomically coherent neighbors even at larger retrieval ranks.

6 Discussion and Conclusions

We introduce CCMIL, a Cross-modal Contrastive MIL framework that enables interpretable retrieval of molecular signatures directly from standard H&E slides. By integrating a composite loss with our THCA module, CCMIL learns a shared

histology-transcriptomics latent space during training, improving cross-modal alignment relative to the evaluated baselines. Beyond binary classification, CCMIL acts as an interpretable search-by-case engine, allowing pathologists to retrieve transcriptomically coherent neighbors and impute a patient’s continuous inflammatory phenotype.

This learned retrieval-oriented representation also supports effective downstream classification, yielding a highly sensitive classifier for immunotherapy pre-screening. While our cross-validation demonstrates consistent performance within the TCGA cohort, this proof-of-concept has been evaluated only on a retrospective cohort. Future work will focus on validating the framework in an independent external cohort to assess generalizability. Ultimately, by providing biologically meaningful spatial heatmaps and bypassing genomic sequencing at inference, CCMIL offers a scalable framework for precision immuno-oncology.

Acknowledgments. This work was supported by the research project PID2023-148614OB-I00, funded by MICIU/AEI/10.13039/501100011033 and by FEDER, EU. The research conducted by Sigrid Vila-Bagaria is supported by the AGAUR-FI predoc-toral programme (2025 FI-STEP 00032) of the Secretariat for Universities and Research of the Department of Research and Universities of the Generalitat of Catalonia, and by the European Social Fund Plus. This study has been funded by Instituto de Salud Carlos III (ISCIII) through the research project PI24/01017, and co-funded by the European Union. MTM acknowledges the support from Diputació de Lleida La força dels municipis and from the Intramural Research Program of Institut de Recerca Biomèdica de Lleida (IRBLleida). RM acknowledges the support from Instituto de Salud Carlos III (Juan Rodés contract: JR21/00039) co-funded by the European Union.

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

References

1. Ajani, J.A., Lee, J., Sano, T., Janjigian, Y.Y., Fan, D., Song, S.: Gastric adenocarcinoma. *Nature reviews Disease primers* **3**(1), 1–19 (2017)
2. Borcoman, E., Kanjanapan, Y., Champiat, S., Kato, S., Servois, V., Kurzrock, R., Goel, S., Bedard, P., Le Tourneau, C.: Novel patterns of response under immunotherapy. *Annals of Oncology* **30**(3), 385–396 (2019)
3. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4015–4025 (2021)
4. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: *2007 IEEE 11th international conference on computer vision*. pp. 1–8. IEEE (2007)
5. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*. vol. 2, pp. 1735–1742. IEEE (2006)
6. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: *International conference on machine learning*. pp. 2127–2136. PMLR (2018)

7. Iscen, A., Tolias, G., Avrithis, Y., Furon, T., Chum, O.: Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2077–2086 (2017)
8. Janjigian, Y.Y., Kawazoe, A., Bai, Y., Xu, J., Lonardi, S., Metges, J.P., Yanez, P., Wyrwicz, L.S., Shen, L., Ostapenko, Y., et al.: Pembrolizumab plus trastuzumab and chemotherapy for her2-positive gastric or gastro-oesophageal junction adenocarcinoma: interim analyses from the phase 3 keynote-811 randomised placebo-controlled trial. *The Lancet* **402**(10418), 2197–2208 (2023)
9. Janjigian, Y.Y., Shitara, K., Moehler, M., Garrido, M., Salman, P., Shen, L., Wyrwicz, L., Yamaguchi, K., Skoczylas, T., Bragagnoli, A.C., et al.: First-line nivolumab plus chemotherapy versus chemotherapy alone for advanced gastric, gastro-oesophageal junction, and oesophageal adenocarcinoma (checkmate 649): a randomised, open-label, phase 3 trial. *The Lancet* **398**(10294), 27–40 (2021)
10. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. *Advances in neural information processing systems* **30** (2017)
11. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: International conference on learning representations (2018)
12. Li, T., Zhou, X., Xue, J., Zeng, L., Zhu, Q., Wang, R., Yu, H., Xia, J.: Cross-modal alignment and contrastive learning for enhanced cancer survival prediction. *Computer Methods and Programs in Biomedicine* **263**, 108633 (2025)
13. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863874 (2024)
14. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F.K., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19764–19775 (June 2023)
15. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**(6), 555–570 (2021)
16. Min, W., Shi, Z., Zhang, J., Wan, J., Wang, C.: Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics* **25**(6), bbae551 (2024)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. Rha, S.Y., Oh, D.Y., Yañez, P., Bai, Y., Ryu, M.H., Lee, J., Rivera, F., Alves, G.V., Garrido, M., Shiu, K.K., et al.: Pembrolizumab plus chemotherapy versus placebo plus chemotherapy for her2-negative advanced gastric cancer (keynote-859): a multicentre, randomised, double-blind, phase 3 trial. *The Lancet Oncology* **24**(11), 1181–1195 (2023)
19. Rodriguez, J.V., Piñol, M., Sorolla, M.A., Parisi, E., Sorolla, A., Santacana, M., Ruiz, M., Parra, G., Bernabeu, M., Iglesias, M., et al.: Comprehensive immunophenotyping of gastric adenocarcinoma identifies an inflamed class of tumors amenable to immunotherapies. *Journal for Immunotherapy of Cancer* **13**(3), e010024 (2025)

20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems* **34**, 2136–2147 (2021)
21. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859* (2017)
22. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **71**(3), 209–249 (2021)
23. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. vol. 2022, p. 3876 (2022)
24. Yu, S., Kim, Y., Kim, H., Lee, S., Kim, K.: Contrastive learning for omics-guided whole-slide visual embedding representation. *bioRxiv* pp. 2025–01 (2025)
25. Zhang, A., Jaume, G., Vaidya, A., Ding, T., Mahmood, F.: Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750* (2025)
26. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: *Machine learning for healthcare conference*. pp. 2–25. PMLR (2022)