Dual-Query Multiple Instance Learning for Dynamic Meta-Embedding based Tumor Classification

Simon Holdenried-Krafft[1]
simon.krafft@uni-tuebingen.de
Peter Somers[3]
somers@isys.uni-stuttgart.de
Ivonne A. Montes-Majarro[2]
ivonne.montes@med.uni-tuebingen.de
Diana Silimon[2]
dianasilimon@icloud.de
Cristina Tarín[3]
cristina.tarin-sauer@isys.uni-stuttgart.de
Falko Fend[2]
falko.fend@med.uni-tuebingen.de
Hendrik P. A. Lensch[1]
hendrik.lensch@uni-tuebingen.de

[1] Institute for Computer Graphics, University of Tübingen, Tübingen, Germany
[2] Institute of Pathology and Neuropathology, University Hospital of Tübingen, Tübingen, Germany
[3] Institute for System Dynamics, University of Stuttgart, Stuttgart, Germany

Summary

Whole slide image (WSI) assessment is a challenging and crucial step in cancer diagnosis and treatment planning. WSIs require high magnifications to facilitate subcellular analysis. Precise annotations for patch- or even pixel-level classifications in the context of gigapixel WSIs are tedious to acquire and require domain experts. Coarse-grained labels, on the other hand, are easily accessible, which makes WSI classification an ideal use case for multiple instance learning (MIL). In our work, we propose a novel embedding-based Dual-Query MIL pipeline (DQ-MIL). We contribute to both the embedding and aggregation steps. Since all-purpose visual feature representations are not yet available, embedding models are currently limited in terms of generalizability. With our work, we explore the potential of dynamic meta-embedding based on cutting-edge self-supervised pre-trained models in the context of MIL. Moreover, we propose a new MIL architecture capable of combining MIL-attention with correlated self-attention. The Dual-Query Perceiver design of our approach allows us to leverage the concept of self-distillation and to combine the advantages of a small model in the context of a low data regime with the rich feature representation of a larger model. We demonstrate the superior performance of our approach on three histopathological datasets, where we show improvement of up to 10% over state-of-the-art approaches. GitHub repository: https://github.com/cgtuebingen/DualQueryMIL

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

1 Introduction

Histopathological slide assessment is the gold standard for grading and treatment planning for almost all types of cancer [12]. In computational pathology, slide scanners convert tissue specimens on glass slides into digital images. Due to the required subcellular details, the scanned slide specimens, also called whole slide images (WSIs), can have more than a hundred thousand pixels in each dimension. Processing such gigapixel images entirely is computationally intractable. Hence, WSIs are subdivided into patches, reducing the computational burden and allow for processing each patch with well-established architectures such as a convolutional neural network (CNN) or Transformers [33]. Unfortunately, precise annotations for patch- or even pixel-level classifications in the context of gigapixel images are labor intensive to acquire and require expert knowledge. Instead, slide-level labels, such as tissue type, cancer grade, or molecular subtype are widely available and less time-consuming to collect. Multiple instance learning (MIL), a subset of weakly supervised learning introduced by Dietterich et al. [10], can make use of such coarse-grained labels and has shown its effectiveness in the field of WSI classification in a variety of recent studies [3, 14, 15, 19, 21, 35].

MIL defines one sample as a bag of instances and there are two major categories: instance-based or embedding-based [1, 15]. Different studies indicate that embedding-based MIL has superior performance compared to instance-based MIL [3, 15, 19, 24, 27]. Embedding-based approaches first transform all instances into learned feature vectors, aggregate them into a joint bag representation, and conclude with a bag-level classification. The initial step of acquiring robust visual features is demanding, especially for WSI classification, where relevant features depend on the cancer entity [29]. But even for the same entity, WSIs can vary from hospital to hospital and show severe differences in appearance due to slightly different staining chemicals [28]. Thus, out-of-distribution generalization remains a challenge for embedding models, and as the quality of the feature representations directly affects the performance on the downstream task [19], it is not negligible.

Although aggregation models can supplement the embedding architecture by leveraging the supervised training signal [18] to enrich the representations, they come with inherent issues. In classical MIL, a WSI is defined as a bag and its corresponding patches are assumed to be independent and identically distributed (i.i.d.) instances. Given a binary cancer classification task, the whole bag is labeled as cancerous as soon as a single patch is cancerous. For highly unbalanced bags, where only a small fraction of patches are actually positive (diseased), the training signal diminishes due to the dominance of negative instances [19, 40]. In such cases, simple models tend to misclassify. While larger models can still learn rich bag representations, they tend to overfit in small data regimes, common in medical image analysis. Another dubious aspect of classical MIL in the context of WSI classification is the i.i.d. assumption [27, 32]. In fact, pathologists exploit structural information to enrich smaller areas with the surrounding context. Thus, correlating instances seems natural, but due to a large number of instances within one bag, it can be computationally demanding, especially for Transfomer-based approaches [27].

In our work, we address the various topics previously mentioned. We conduct extensive experiments to validate the benefits of our approach and evaluate our method based on three different publicly available medical datasets on tumor classification and cancer subtyping. Our contributions are threefold:

- We introduce a novel MIL architecture, named Dual-Query MIL inspired by the Perceiver [16]. Due to its design, the Perceiver decouples the input size from the dimen-

sionality of a latent representation, eliminating the quadratic scaling problem of the classical Transformer architecture [33]. Our dual-query design in the cross-attention layer combines i.i.d. MIL attention [15] with correlative self-attention [33] in one architecture.

- We introduce a self-distillation loss function, which allows us to leverage both the advantages of a small and a larger aggregation model in one framework, preventing overfitting while simultaneously acquiring rich feature representations.
- We explore the potential of dynamic meta-embedding [18, 31] based on three state-of-the-art self-supervised learning (SSL) methods in the context of MIL. Our experiments show the superiority of dynamic meta-embedding compared to individual embeddings and indicate a step towards robust visual representations in the context of medical image analysis.

2 Related Work

As our work focuses on deep MIL-based histopathological slide assessment, we want to provide an overview of the most recent trends and the role of SSL embedding specific to this field of research. For further literature, we refer to [1, 4, 34].

Deterministic MIL pooling operations, such as max or mean pooling, are limited in terms of performance. Therefore, Ilse et al. [15] base the pooling operation on DNNs, which assign attention scores to i.i.d. instances, defining the contribution of each instance to the final bag representation. Lu et al. [21] extended the idea of attention-based instance scoring. They utilize instance-level clustering to guide the learning and to constrain the feature space by creating class-specific pseudo-labels and subsequent class-specific attention branches for multi-class settings. All these methods neglect correlations between instances, whereas graph or capsule-based architectures [32, 37] incorporate contextual information between instances. This resembles a pathologist's proceeding that connects local characteristics such as the nucleus shape and size with the global context, e.g. surrounding cell architecture. Most recent architectures use non-local attention. Dual-stream MIL (DS MIL) [19] consists of one branch, which detects the most significant instance using a max-pooling operation, and a second branch correlating the detected characteristic instance with all remaining instances using a Transformer-like one-to-all attention mechanism. The one-to-many approach by Bergner et al. [2] similarly consists of two stages: an iterative patch selection (IPS) and a small cross-attention Transformer stage. Using the IPS module drastically reduces the number of patches per bag, accelerating the aggregation step. Shao et al. [27] utilize an approximated all-to-all self-attention by utilizing the Nyström method [36]. This allows for large inputs, as is crucial for WSI classification, and reduces the computational cost of multi-head self-attention.

Our proposed method is based on the Perceiver model by Jaegle et al. [16]. Instead of the classical all-to-all Transformer self-attention [33] with its quadratic scaling problem, the Perceiver relies on an asymmetric attention mechanism. This reduces the computational complexity and decouples the input size from the depth of the architecture. The Perceiver exploits a cross-attention layer to transform the input into a condensed latent array. The all-to-all self-attention is only applied in this latent array. Furthermore, we combine this approach with the one-to-all query design from Bergner et al. [2]. Our adaptation combines MIL and Transformer attention in one architecture. Whereas most other methods rely on cross-entropy (CE) loss with bag-labels as the training signal [15, 19, 27], we utilize the

concept of self-distillation [38, 39] to fully exploit the potential of our approach. Besides the aggregation model, we also explore new ways of feature extraction or merging. As indicated by Tendle and Hasan [30], the generalization of SSL representations is improved compared to supervised learning (SL) representations. However, instead of training an embedding model with histopathological samples using SSL [7, 19], we explore the potential of dynamic meta-embedding in the context of MIL based on three of the most recent pre-trained SSL methods (SwAV [6], DINO [6], DINOv2 [23]). This idea from the vibrant field of natural language processing showed increased robustness and generalization by combining multiple embedding techniques complementary to one another [18, 31].

3 Methodology

During classical supervised training, a model learns to estimate the given label y corresponding to input image $x \in \mathbb{R}^{h \times w \times 3}$. Instead, multiple-instance learning is set-based. Each set consists of multiple inputs, or instances, and is called a bag $\mathcal{B} = \{x_1, ..., x_N\}$. The number of instances N within the bag can vary between bags. Moreover, we assume that there exists a label y_n with $n = 1, ..., N$ for each instance within the bag, which is unknown. Only one global label $\mathcal{Y}$ is given for the whole bag $\mathcal{B}$. In a binary MIL classification task, the bag label is positive as soon as a single instance label is positive. To estimate the final label of bag $\mathcal{B}$, multiple instance learning requires suitable transformations represented by f and g. The choice of f and g defines whether it is an instance-based or embedding-based approach [1, 15]. In instance-based MIL, f transforms each instance into scores, and function g is a pooling operation, such as max- or mean-pooling, aggregating the scores. In embedding-based MIL, f first projects the instances into a newly learned embedding space, and function g afterward distills all instances corresponding to one bag into a joint bag representation.

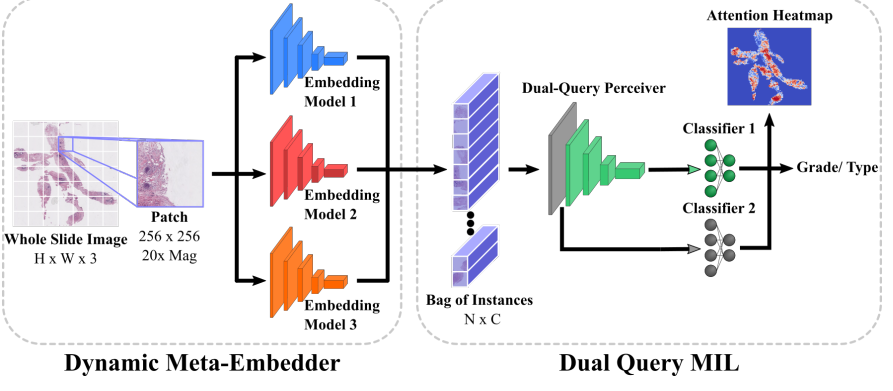

Figure 1: DQ-MIL pipeline. The Dynamic Meta-Embedder (DME) combines three different feature representations per patch and creates one joint feature vector. The bag of DME representations is then processed by the Dual-Query (DQ) Perceiver in two different pathways, exploiting the advantages of MIL- and self-attention.

In our proposed method, illustrated in Figure 1, we touch upon both transformations f and g of the embedding-based procedure. First, we introduce the concept of meta-embedding in the context of multiple instance learning with the Dynamic Meta-Embedder (DME), corresponding to projection f. Furthermore, we propose the Dual-Query (DQ) Perceiver rep-

resenting function g, based on the Perceiver architecture [16]. Our method leverages the flexibility of the Perceiver and joins MIL and self-attention in one framework.

3.1 Dynamic Meta-Embedding for Multiple Instance Learning

Instance-embedding models transform a raw input patch x_i into a feature vector $\mathbf{h}_i = f(x_i)$. We utilize three SSL pre-trained encoding models to distill the raw image into a single feature representation. Rather than just concatenate the embeddings, we utilize the training signal of the downstream task to dynamically learn the new representation [18, 31]. Our Dynamic Meta-Embedder consists of two ResNet50 architectures [13], and one Vision Transformer [11] (ViT-L/14). The two ResNet models were pre-trained on the ImageNet dataset [9], whereas the ViT used the LVD-142M dataset [23].

Besides the architecture, all three embedders differ in terms of the SSL technique used for pre-training. One ResNet model was pre-trained using SwAV [5], the other one utilizes the DINO approach [6]. The ViT model was pre-trained with the most recent SSL method DINOv2 [23]. DINOv2 joins ideas from various SSL methods, the image-level loss of DINO [6], the masked image modeling of iBOT [41], the Sinkhorn-Knopp centering of SwAV [5], and more. After piping the input patch through each of the embedding models, the Dynamic Meta-Embedder projects all three embeddings of various lengths to the same dimensionality using separate linear layers per embedder. This step, in which the three SSL models are frozen, allows exploiting the training signal from the bag label to fine-tune the representations and to extract task- and domain-specific features. Figure 2 depicts the DME module.

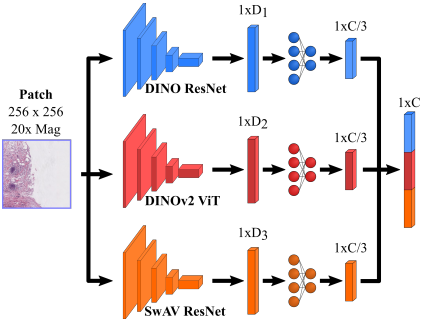

Figure 2: Dynamic Meta-Embedder. The different embedding models first condense each patch into a representation vector, then each of them gets processed in three different trainable linear layers and concatenated to a single vector.

3.2 Dual-Query Perceiver

As an aggregator model to summarize the bag, we propose the Dual-Query Perceiver. This architecture is based on the Perceiver and Perciever IO idea [16, 17]. We leverage the flexibility of the proposed querying mechanism and propose the novel MIL architecture in Figure 3. The key components of our method are the Dual-Query Cross-Attention Module and the Latent Transformer. Both include a query-key-value (QKV) attention block, the core element in all Transformer-like architectures. It transforms the input into queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ by piping the input through three multi-layer perceptrons (MLPs). The general attention operation itself can be expressed as:

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\tau}\right)\mathbf{V}, \tag{1}$$

where the temperature τ scales the dot-product of $\mathbf{Q}$ and $\mathbf{K}^T$. There are two types of attention, self-attention and cross-attention. In self-attention, the queries originate from the same

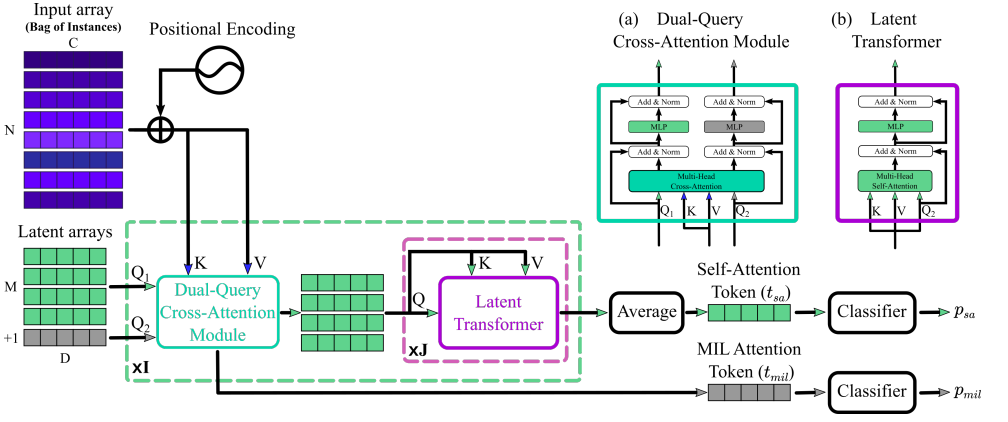

Figure 3: Dual-Query Perceiver. Consisting of two pathways, the DQ Perceiver combines MIL- and self-attention in one framework. The two main components, Dual-Query Cross-Attention Module **(a)**, and Latent Transformer **(b)**, follow the typical structure proposed by Vaswani et al. [33]. However, the proposed Cross-Attention module consists of two pathways based on two separate queries $\mathbf{Q_1}$ and $\mathbf{Q_2}$. Both branches share keys $\mathbf{K}$ and values $\mathbf{V}$.

source as keys and values. While in cross-attention, queries do not share the same origin.

The DQ Perceiver combines both attention categories. In the cross-attention module, keys $\mathbf{K} \in \mathbb{R}^{N \times d_k}$, and values $\mathbf{V} \in \mathbb{R}^{N \times d_k}$ are projections of the input array (bag-of-instances) with shape $N \times C$. The queries $\mathbf{Q}_1 \in \mathbb{R}^{M \times d_k}$ and $\mathbf{Q}_2 \in \mathbb{R}^{1 \times d_k}$ are projections of two learned latent arrays, one of size $M \times D$ and the other of shape $1 \times D$, with $M + 1 \ll N$. As the query defines the shape of the output, the input array is distilled into a latent array of fixed size.

The dual-query module, illustrated in Figure 3a, leverages this behavior and creates two pathways. The first pathway is based on the regular Perceiver pipeline. Here we use $\mathbf{Q}_1$ to compress the input into a latent array, which afterward gets processed by the Latent Transformer. This module, shown in Figure 3b, performs self-attention on the latent array. The latent array is piped through the Latent Transformer J times to improve the features. In the final step of this pathway, the latent array is averaged along the instance dimensions M to obtain the self-attention token t_{sa}.

The second pathway is based on the idea of MIL-attention, where an attention-score a weights each instance, so the aggregation function g corresponds to weighted sum, see Equation 2. This can be transferred to a single query cross-attention. Similar to the proposed method by Bergner et al. [2], the query ($\mathbf{Q}_2$) of size $1 \times d_k$ is used to predict attention scores for each projected instance $\mathbf{k}_i = \mathbf{W}_k \mathbf{h}_i$. Afterward, a second projected version of the instance $\mathbf{h_i}$, $\mathbf{v}_i = \mathbf{W}_v \mathbf{h}_i$ is scaled by the predicted attention score a_i and summed up to build the MIL-attention token t_{mil}.

$$t_{mil} = \sum_{i=1}^{N} a_i \mathbf{v}_i = \sum_{i=1}^{N} a_i \mathbf{W}_v \mathbf{h}_i = \sum_{i=1}^{N} \frac{\exp\Big(s\big(\mathbf{Q}_2, \mathbf{k}_i\big)\Big)}{\sum_{k=1}^{N} \exp\Big(s\big(\mathbf{Q}_2, \mathbf{k}_k\big)\Big)} \mathbf{W}_v \mathbf{h}_i, \tag{2}$$

where $s\big(\cdot,\cdot\big)$ denotes the scaled dot-product, given by $s\big(\mathbf{Q}_2, \mathbf{k}\big) = \frac{\mathbf{Q}_2 \mathbf{k}^T}{\tau}$ with temperature τ used as scaling factor. Each of the final bag representations t_{sa} and t_{mil} is processed in a separate MLP-based classifier in combination with a softmax operation to acquire the

corresponding probability distribution p_{sa} and p_{mil}, where p_{sa} was determined by the self-attention based Perceiver branch and p_{mil} predicted by the MIL pathway. The final output during inference is derived with a simple, balanced weighting mechanism, which can be expressed as $p = bp_{sa} + (1-b)p_{mil}$, with b as a hyper-parameter. This combination of outputs enables an architecture immanent supervision and the utilization of a self-distillation-based learning strategy.

3.2.1 Self-distillation Loss

Self-distillation exploits components within an architecture to set up a knowledge-distillation-like learning scheme, in which shallow parts of a network are treated as an independent student architecture [38, 39]. For the DQ Perceiver, the final loss function $\mathcal{L}_{SD}$, is a combination of the main Cross-Entropy (CE) loss, $\mathcal{L}_{CE}(p_{sa}, \mathcal{Y})$ of the deepest part (Perceiver branch) and three additional self-distillation losses of the shallow part (MIL branch).

Like the main branch, the Cross-Attention Module also receives supervision by the bag label $\mathcal{Y}$. Moreover, the deeper Perceiver pathway supervises the Cross-Attention Module, using the Kullback-Leibler divergence between p_{mil} and p_{sa}. This is complemented by an L2 loss, also called hint [26], inducing the MIL-attention token t_{mil} to fit the self-attention token t_{sa}. Hyper-parameter α and λ are used for balancing the contributions of the different loss terms. In our experiments, we empirically found the weighting factors $\alpha = 0.7$ and $\lambda = 0.03$ worked best for varying data sets.

$$\mathcal{L}_{SD} = \mathcal{L}_{CE}(p_{sa}, \mathcal{Y}) + \alpha \mathcal{L}_{CE}(p_{mil}, \mathcal{Y}) + (1-\alpha)\mathcal{L}_{KL}(p_{mil}, p_{sa}) + \lambda \|t_{sa} - t_{mil}\|_2^2 \quad (3)$$

4 Experiments and Results

4.1 Experimental Design

We thoroughly evaluate the DQ-MIL on three different histopathological datasets (Camelyon16, TCGA-BRCA, and TCGA-BLCA). The tasks are cancer classification and subtyping. Details regarding the different datasets, their curation, as well as about implementation are covered in the supplementary material. We report our evaluation using two metrics, area under the curve (AUC), and accuracy scores. Furthermore, we evaluated the localization of the most significant instances qualitatively. During pre-processing, each WSI is subdivided into patches, $x_i \in \mathbb{R}^{256\times256\times3}$ in magnification of $20\times$. Patches with background or artifacts are sorted out by combining threshold-based filtering [21] with a pre-trained tissue segmentation U-Net [25].

4.2 Tumor Classification and Cancer Subtyping

The two tasks we use for evaluation, tumor classification, and cancer subtyping, cover complementary challenges. Slides from the Camelyon16 dataset are highly unbalanced, where less than 10% of the tissue area per slide covers positive instances (cancer) [19]. In contrast, The Genome Cancer Atlas (TCGA) datasets [20], which we use to test performance in cancer subtyping, require that the network does not just focus on small regions within the tissue but rather evaluate the global appearance of WSIs. For cancer subtyping, we use two publicly available datasets of different entities, breast cancer (BRCA) and bladder cancer (BLCA). The results of the classification are summarized in Table 1. We realized that

the pre-processing step, especially the Otsu-based filtering, has a strong impact on the final evaluation metrics. Thus, all values represented are based on experiments we run under the exact same conditions, using the dynamic meta-embedding approach for all of the different methods.

Aggregation Method	Camelyon16		TCGA-BRCA		TCGA-BLCA	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
DS MIL [19]	0.8527	0.8605	0.9434	0.8814	0.7312	0.8061
TransMIL [27]	0.8559	0.8450	0.9308	0.9040	0.6769	<u>0.8673</u>
CLAM-SB [21]	<u>0.8946</u>	<u>0.8915</u>	**0.9455**	**0.9266**	<u>0.7448</u>	0.8061
DQ-MIL-SD	**0.9594**	**0.9457**	<u>0.9441</u>	**0.9266**	**0.8461**	**0.9184**

Table 1: Performance evaluation of different MIL architectures on three medical datasets. The best performance is written in bold digits, and the second-best is underlined.

Our proposed self-distilled DQ-MIL achieves state-of-the-art performance on the TCGA-BRCA dataset. For the Camelyon16 dataset, we achieve an improvement of up to 6.4% in AUC and 5.4% in accuracy. For the BLCA dataset, the improvement is even more significant, with up to 10.1% in AUC and 5.1% in accuracy compared to the second-best performing networks per metric.

To assess whether the DQ Perceiver is able to localize the most relevant areas with regard to the classification task, we also conduct a qualitative analysis, illustrated in Figure 4. We utilize the pixel-wise annotations of the Camelyon16 dataset to evaluate the match between patches with top 5% attention scores (highlighted in red Figure 4 **(c-f)**) and cancerous regions, annotated by domain experts (green contours in Figure 4 **(c-f)**). We can see that the DQ Perceiver is congruent with the annotated regions and is even able to detect small cancer areas, as shown in Figure 4 **(c)**.

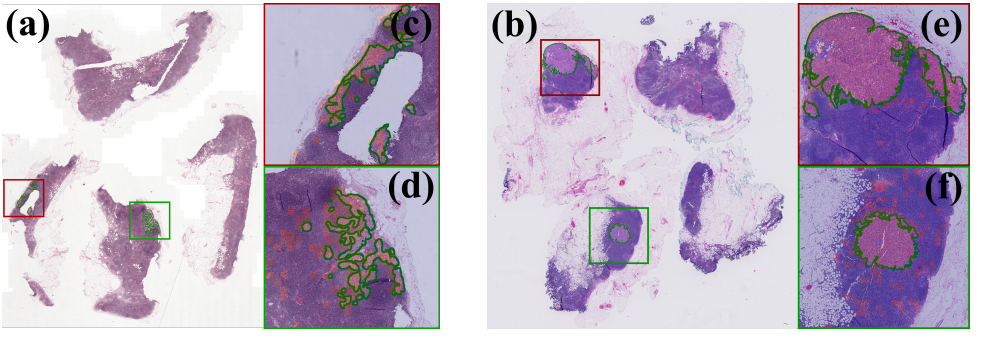

Figure 4: Visualization of the most significant attention scores. **(a)** and **(b)** show two WSIs from the Camelyon16 dataset. The red and green bounding boxes indicate the position of the cropped regions shown in **(c-f)**. The green contours in **(c-f)** indicate the cancerous regions annotated by pathologists. The attention scores are normalized per slide to [0,1], where the red colored regions in **(c-f)** highlight the patches with attention scores higher than 0.95.

4.3 Ablation Study

4.3.1 Effects of the Individual Components of the Dual-Qyery Perceiver

In this section, we evaluate how the individual components of our approach perform on the different classification tasks, the results are given in Table 2. Each sub-component, namely pure MIL Cross-Attention, the original Perceiver proposed by Jaegle et al. [16], and the Dual-Query Perceiver (DQ-MIL) without additional self-distillation is tested on all of the three medical datasets. For all sub-networks, we use a single cross-entropy loss during training. The final logits of the regular DQ-MIL are derived with the weighting mechanism, mentioned above, with $b = 0.5$, leading to $p = \frac{1}{2}(p_{sa} + p_{mil})$. For this ablation study, we also use the dynamic meta-embedding strategy.

Aggregation Method	Camelyon16		TCGA-BRCA		TCGA-BLCA	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
MIL Cross-Attention	0.9497	0.9380	0.8940	0.9040	0.8172	0.8673
Perceiver	0.9439	0.9147	**0.9464**	0.8475	**0.8679**	0.8571
DQ-MIL	0.9099	0.9147	0.9362	0.8418	0.8303	0.8571
DQ-MIL-SD	**0.9594**	**0.9457**	0.9441	**0.9266**	0.8462	**0.9184**

Table 2: Comparison of the derivatives of the DQ-MIL-SD approach.

We can see that the DQ-MIL-SD approach is slightly better or on par with its sub-components. The table also indicates that the self-distillation loss is the key element to boost the performance of the DQ-MIL-SD architecture and to join the benefits of the small MIL Cross-Attention model and the larger Perceiver. The table point out that for the cancer subtyping task, a correlation between the instances is slightly beneficial to improve on the AUC metric, whereas the MIL-attention approach achieves higher accuracy values.

4.3.2 Effect of the Dynamic Meta-Embedding Strategy

We also conduct an ablation study to indicate the benefits and advantages of dynamic meta-embedding. Here we used the DQ-MIL-SD approach as our fixed evaluation model. We trained the model using the different embedding methods shown in Table 3. Each embedding model varies in terms of architecture and SSL strategy. Furthermore, we compare the out-of-domain embedding methods (*) with two in-domain pre-trained embedding methods (†). The in-domain methods are a ResNet18 pre-trained on the Camelyon16 dataset using SimCLR [8, 19] and a Vision Transformer (ViT) pre-trained on a large and comprehensive TCGA dataset covering multiple entities [7].

The performance evaluations show the advantage of the proposed Dynamic Meta-Embedder. It also indicates that the aggregation model can compensate for the embedding models' lack of domain knowledge. Although we were surprised to observe that the in-domain methods did not generalize well across our evaluation datasets, it resonates recent findings by McBee et al. [22].

Embedding Method	Camelyon16		TCGA-BRCA		TCGA-BLCA	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
ViT-B/8 Dino† [7]	0.7298	0.7519	0.8900	0.8701	0.7611	0.8163
ResNet18 SimCLR† [19]	0.9136	0.9225	0.8443	0.8475	0.7928	0.7857
ResNet50 SwAV*	0.9406	0.9302	0.9201	0.8927	0.8045	0.8776
ResNet50 DINO*	0.8543	0.8837	0.9347	0.8814	0.7747	0.8265
ViT-L/14 DINO v2*	0.7474	0.7984	**0.9704**	**0.9266**	0.7405	0.8367
Dynamic Meta-Embedder*	**0.9594**	**0.9457**	0.9441	**0.9266**	**0.8462**	**0.9184**

Table 3: Comparison of different embedding methods evaluated with a fixed DQ-MIL-SD aggregation model. The Dynamic Meta-Embedder utilizes all three SSL methods, pre-trained on ImageNet (*). The (†) indicates in-domain methods pre-trained on WSI patches.

5 Conclusion and Future Work

In our work, we present a novel MIL approach called DQ-MIL-SD to the field of histopathological slide assessment. We introduce a dual-query cross-attention layer to combine single-token MIL-cross-attention with multi-token Perceiver cross- and self-attention in one architecture. By introducing a self-distillation loss, we can leverage the advantages of a small and a larger aggregation model. The proposed DQ Perceiver outperforms recent state-of-the-art approaches or is on par. In addition, combining multiple pre-trained embedders by the Dynamic Meta-Embedder ensures consistent performance across datasets. The next step will be to extend this approach to a multi-modal setting, allowing us to fully leverage the flexibility of the Perceiver and to explore its potential in the field of molecular subtyping.

Acknowledgements This work was sponsored by the Graduate School 2543/1 "Intraoperative Multisensory Tissue Differentiation in Oncology" (project ID 40947457), funded by the German Research Foundation (DFG - Deutsche Forschungsgemeinschaft). This work was also supported in part by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A.

References

[1] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 8 2013. ISSN 0004-3702. doi: 10.1016/J.ARTINT.2013.06.003.

[2] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. In *The Eleventh International Conference on Learning Representations*, 2 2023.

[3] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine 2019 25:8*, 25:1301–1309, 7 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0508-1.

[4] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2017.10.009.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Advances in Neural Information Processing Systems*, 2020-December, jun 2020. ISSN 10495258.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[7] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155, June 2022.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-3:1575–1585, 2 2020.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1 1997. ISSN 0004-3702. doi: 10.1016/S0004-3702(96)00034-3.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[12] Metin N Gurcan, Senior Member, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2, 2009. doi: 10.1109/RBME.2009.2034865.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:2424–2433, 4 2015. ISSN 10636919. doi: 10.1109/CVPR.2016.266.

[15] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *35th International Conference on Machine Learning, ICML 2018*, 5: 3376–3391, 2 2018. doi: 10.48550/arxiv.1802.04712.

[16] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR, 18–24 Jul 2021.

[17] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022.

[18] Douwe Kiela, Changhan Wang, and Kyunghyun Cho. Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1477, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1176.

[19] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 11 2020. ISSN 10636919. doi: 10.48550/arxiv.2011.08939.

[20] Jianfang Liu, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173:400–416.e11, 4 2018. ISSN 10974172. doi: 10.1016/j.cell.2018.02.052.

[21] Ming Y. Lu, Drew F.K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *Nature Biomedical Engineering*, 5:555–570, 4 2020. ISSN 2157846X. doi: 10.48550/arxiv.2004.09666.

[22] Payden McBee, Nazanin Moradinasab, Donald E Brown, and Sana Syed. Pre-training segmentation models for histopathology. In *Medical Imaging with Deep Learning, short paper track*, 2023.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, Piotr Bojanowski, and Meta Ai Research. DINOv2: Learning Robust Visual Features without Supervision. apr 2023.

[24] Ziniu Qian, Kailu Li, Maode Lai, Eric I.Chao Chang, Bingzheng Wei, Yubo Fan, and Yan Xu. Transformer based multiple instance learning for weakly supervised histopathology image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13432 LNCS:160–170, 2022. ISSN 16113349. doi: 10.1007/978-3-031-16434-7_16/TABLES/4.

[25] Abtin Riasatian, Maral Rasoolijaberi, Morteza Babaei, and H. R. Tizhoosh. A comparative study of u-net topologies for background removal in histopathology images. *Proceedings of the International Joint Conference on Neural Networks*, 6 2020. doi: 10.1109/IJCNN48605.2020.9207018.

[26] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015.

[27] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 3:2136–2147, 6 2021. ISSN 10495258. doi: 10.48550/arxiv.2106.00908.

[28] Milad Sikaroudi, Maryam Hosseini, Ricardo Gonzalez, Shahryar Rahnamayan, and H. R. Tizhoosh. Generalization of vision pre-trained models for histopathology. *Scientific Reports 2023 13:1*, 13(1):1–14, apr 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-33348-z.

[29] Karin Stacke, Jonas Unger, Claes Lundström, and Gabriel Eilertsen. Learning representations with contrastive self-supervised learning for histopathology applications. *Machine Learning for Biomedical Imaging*, 1:1–33, 2022. ISSN 2766-905X.

[30] Atharva Tendle and Mohammad Rashedul Hasan. A study of the generalizability of self-supervised representations. *Machine Learning with Applications*, 6:100124, dec 2021. ISSN 26668270. doi: 10.1016/J.MLWA.2021.100124.

[31] Tuan Truong, Sadegh Mohammadi, and Matthias Lenga. How transferable are self-supervised features in medical image classification tasks? In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 54–74. PMLR, 04 Dec 2021.

[32] Ming Tu, Jing Huang, Xiaodong He, and Bowen Zhou. Multiple instance learning with graph neural networks. In *ICML 2019 workshop on Learning and Reasoning with Graph-Structured Representations*, jun 2019.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[34] Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10:213–247, 6 2022. ISSN 21686831. doi: 10.1109/MGRS.2022.3198244.

[35] Wei Wu, Zhonghang Zhu, Baptiste Magnier, and Liansheng Wang. Clustering-based multi-instance learning network for whole slide image classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13574 LNCS:100–109, 2022. ISSN 16113349. doi: 10.1007/978-3-031-17266-3_10.

[36] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 16: 14138–14148, 2 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i16.17664.

[37] Yongluan Yan, Xinggang Wang, Jiemin Fang, Wenyu Liu, Junzhou Huang, Jun Zhu, and Ichiro Takeuchi. Deep multi-instance learning with dynamic pooling. In *Proceedings of Machine Learning Research*, volume 95, pages 662–677. PMLR, 11 2018.

[38] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 3712–3721, 2019. ISBN 9781728148038. doi: 10.1109/ICCV.2019.00381.

[39] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-Distillation: Towards Efficient and Compact Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, aug 2022. ISSN 19393539. doi: 10.1109/TPAMI.2021.3067100.

[40] Xiaoxian Zhang, Sheng Huang, Yi Zhang, Xiaohong Zhang, Mingchen Gao, and Liu Chen. Dual space multiple instance representative learning for medical image classification. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[41] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.