


## Article

# Dual-Attention Multiple Instance Learning Framework for Pathology Whole-Slide Image Classification

Dehua Liu <sup>1</sup>, Chengming Li <sup>2,3,\*</sup>, Xiping Hu <sup>1,4,\*</sup> and Bin Hu <sup>1,4,\*</sup>

<sup>1</sup> Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China; liudh21@lzu.edu.cn

<sup>2</sup> Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China

<sup>3</sup> Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China

<sup>4</sup> School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

\* Correspondence: licm@smbu.edu.cn (C.L.); huxp@bit.edu.cn (X.H.); bh@bit.edu.cn (B.H.)

**Abstract:** Conventional methods for tumor diagnosis suffer from two inherent limitations: they are time-consuming and subjective. Computer-aided diagnosis (CAD) is an important approach for addressing these limitations. Pathology whole-slide images (WSIs) are high-resolution tissue images that have made significant contributions to cancer diagnosis and prognosis assessment. Due to the complexity of WSIs and the availability of only slide-level labels, multiple instance learning (MIL) has become the primary framework for WSI classification. However, most MIL methods fail to capture the interdependence among image patches within a WSI, which is crucial for accurate classification prediction. Moreover, due to the weak supervision of slide-level labels, overfitting may occur during the training process. To address these issues, this paper proposes a dual-attention-based multiple instance learning framework (DAMIL). DAMIL leverages the spatial relationships and channel information between WSI patches for classification prediction, without detailed pixel-level tumor annotations. The output of the model preserves the semantic variations in the latent space, enhances semantic disturbance invariance, and provides reliable class identification for the final slide-level representation. We validate the effectiveness of DAMIL on the most commonly used public dataset, Camelyon16. The results demonstrate that DAMIL outperforms the state-of-the-art methods in terms of classification accuracy (ACC), area under the curve (AUC), and F1-Score. Our model also allows for the examination of its interpretability by visualizing the dual-attention weights. To the best of our knowledge, this is the first attempt to use a dual-attention mechanism, considering both spatial and channel information, for whole-slide image classification.

**Keywords:** attention; computational pathology; multiple instance learning; whole-slide image; computer-aided diagnosis; WSI classification



**Citation:** Liu, D.; Li, C.; Hu, X.; Hu, B. Dual-Attention Multiple Instance Learning Framework for Pathology Whole-Slide Image Classification. *Electronics* **2024**, *13*, 4445. <https://doi.org/10.3390/electronics13224445>

Received: 9 October 2024

Revised: 1 November 2024

Accepted: 11 November 2024

Published: 13 November 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Histopathology is the gold standard for tumor diagnosis. Tissue sections are typically scanned to create whole-slide images (WSIs), which serve as important references for pathologists [1–3]. Deep learning-based WSI processing and analysis play a crucial role in developing tumor computer-aided diagnosis systems. In recent years, we have witnessed some achievements in assisted disease diagnosis using deep learning models [3–9]. Unlike natural images, WSIs possess extremely high resolutions (typically 100,000 × 100,000 pixels) [10], which poses a labor-intensive task for pixel-wise annotation, making it a significant challenge for constructing accurate classification models. Furthermore, the complexity of WSIs is also reflected in the potential low contrast in localized areas. This phenomenon can reduce the distinction between lesion areas and normal tissues, increasing the risk of misdiagnosis or missed diagnosis. Low contrast not only affects the judgment of human pathologists but also adversely impacts the feature extraction and classification accuracy

of deep learning models [11]. Particularly during the instance feature aggregation phase, the information from positive instances with low contrast can be more easily diluted by other features, leading to potential misclassification at the bag level.

To tackle these challenges, the current approach commonly adopts multiple instance learning (MIL) for WSI classification modeling [12–15]. MIL treats each WSI as a bag containing multiple instances, which are fixed-sized (e.g.,  $256 \times 256$ ) image patches segmented from the tissue regions of the WSI. If there exist positive (abnormal) instances within the bag, it is labeled as positive; otherwise, it is labeled as negative (normal). Subsequently, feature extraction and aggregation of all instances are performed to generate predictions at the bag level.

The inherent complexity of WSIs poses challenges for developing classification models that are not as straightforward as traditional computer vision. Due to the small proportion of positive instances within a positive WSI bag (less than 10%), the most critical features of positive instances may be diluted or even discarded during the stage of instance feature aggregation, leading to misjudgments in the final results. Additionally, with thousands or even tens of thousands of instances in a WSI, but only label at the bag level, the weak supervision signals make it easy for the model to get stuck in local optima during the training process, resulting in poor generalization of the trained model.

In recent years, the modeling approaches based on MIL have shown some achievements in WSI classification. However, there is still much room for improvement in constructing robust models for accurate WSI classification. Due to the inherent characteristics of WSI, most existing MIL models are trained based on bag-level labels, which makes it challenging for the models to learn rich feature representations [16,17]. Furthermore, considering computational costs, many models only sample a few high-weight instance features as bag representatives for subsequent bag prediction [4,16–18]. However, this approach discards the potential semantic relationships among different instances, ultimately leading to suboptimal solutions for the classification models, as these latent semantic relationships are crucial in WSI classification. To address the aforementioned issues, we need to explore how to enable the models to learn the latent semantic relationships among different instances in order to construct robust WSI classification models.

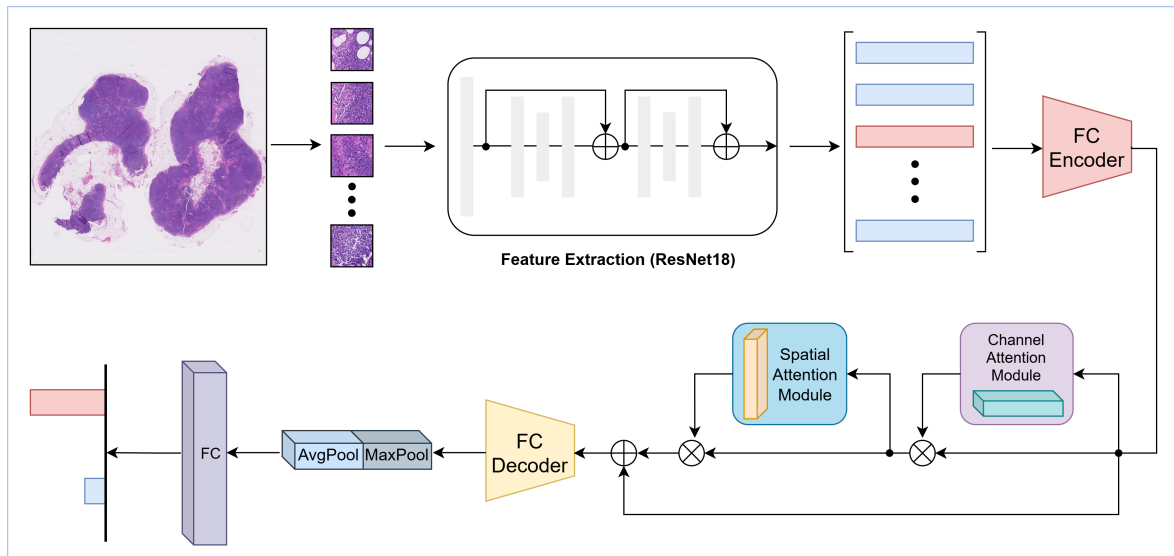
In order to capture the potential semantic relationships between different instances more effectively, we propose a novel dual-attention MIL framework (Figure 1). Specifically, DAMIL utilizes a dual-attention architecture to train the model. The first attention mechanism acquires the channel feature weights of all instances, while the second attention mechanism acquires the weights of all instances in the spatial dimension. Distinct from previous MIL approaches (Figure 2), our method simultaneously takes into account both the spatial and channel information of the instance features within each bag, enabling a more refined representation of bag-level features.

Our main contributions can be summarized as follows:

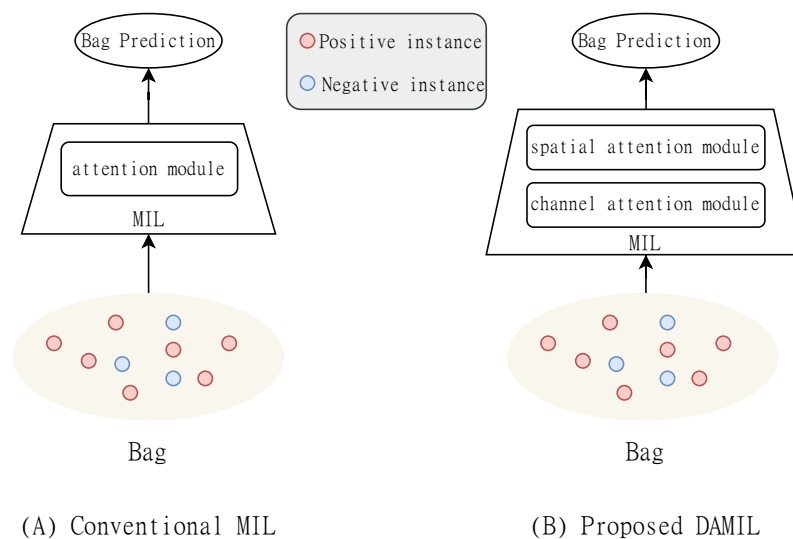
1. We propose a novel dual-attention MIL framework that can capture richer latent connections among instances within a bag, both in the spatial and channel domains, ultimately generating more discriminative bag-level representations for classification prediction.
2. The idea of dual-attention module, which was originally used for 2D image classification, is skillfully applied to our proposed framework, which makes an attempt to provide a new way of thinking for the later migration of models between different learning tasks.
3. We evaluated DAMIL based on the benchmark WSI dataset Camelyon16, and the experimental results show that our method outperforms the current state-of-the-art method in terms of ACC, AUC, and F1.
4. The effectiveness of the proposed module is demonstrated through rich ablation experiments, and the interpretable heatmaps predicted by the model are obtained using Grad-CAM [19], which proves that the ROIs generated by the model are in good agreement with the pixel-level annotations.

The rest of this paper is organized as follows: Section 2 introduces an overview of the concept of MIL, the current state of research on MIL in the classification of whole-slide

images (WSIs), and the application of attention mechanisms in MIL. We delve into the paradigm of multiple instance learning (MIL) and provide a comprehensive exposure to the component modules of the proposed Deep Attention-based Multiple Instance Learning (DAMIL) framework in Section 3. Section 4 presents the experimental results and offers an in-depth discussion of these findings. The conclusion of this paper described in Section 5.



**Figure 1.** Overview of the proposed DAMIL. The WSI is first cropped into a number of patches and then feature extraction is performed with the pre-trained Resnet18. The generated feature vector matrix is passed sequentially through the encoder, channel attention module, spatial attention module, decoder, pooling layer, and fully connected layer to generate the finally prediction.



**Figure 2.** Illustration of the difference between the attention-based conventional MIL model and the proposed dual-attention MIL model.

## 2. Related Work

### 2.1. Multiple Instance Learning (MIL)

MIL is a machine learning framework developed from traditional supervised learning, and is specifically designed for scenarios where annotating at the instance level is difficult or impractical. In supervised learning, we typically work with a large amount of labeled data, where each data instance (sample) has a clear label indicating its class. This learning paradigm assumes that each sample is independent and individually labeled, and cannot

be further decomposed into smaller parts. In contrast, MIL adopts a more relaxed form of supervision. In this framework, data are organized into a series of bags, with each bag containing multiple instances. The key difference from traditional supervised learning is that we only know the label for the entire bag, not for each individual instance within it. For example, a bag may be labeled as positive because it contains at least one positive instance, but we do not know which specific instances are positive. Similarly, a negative bag implies that all its instances are negative. This setup allows the model to work with incomplete labeling information, making it suitable for problems with high annotation costs or technical infeasibilities.

MIL opens up new application domains in machine learning by redefining the organization of the sample set. In the context of MIL, the label of an instance can be seen as latent, as they are not directly observable in the learning process. This requires learning algorithms to reason at a higher level and identify valuable patterns from incomplete or indirect supervision. This learning approach has demonstrated powerful potential and flexibility in domains such as image processing, text classification, and medical diagnosis, especially when dealing with large-scale datasets with only coarse-grained annotations. By focusing attention at the bag level rather than individual instances, MIL provides an effective strategy for handling complex and diverse datasets.

## *2.2. Application of MIL in WSI Classification*

In the field of digital pathology, the classification of WSIs is a challenging task. These high-resolution images contain tens of thousands of cellular structures, including various heterogeneous factors that may affect diagnostic outcomes. Due to the high-dimensional nature and large number of samples in WSIs, direct instance-level annotation is both impractical and resource-consuming. MIL provides an effective modeling strategy for such data [20]. Within the MIL framework, a WSI can be regarded as a bag that contains multiple instances, which may have diagnostic significance (such as specific cells or tissue structures). As MIL allows for bag-level rather than instance-level annotations, it aligns well with the characteristics of WSI data. In practice, pathologists often provide a diagnosis for the entire slide rather than annotating each individual cell or tissue.

Several studies have demonstrated the effectiveness of MIL-based network models for WSI classification applications [4,21–25]. These models can be broadly categorized into two types: instance-level methods [4,25–28] and embedding-level methods [13,18,29–31]. Instance-level methods typically assign bag-level labels to each instance, resulting in pseudo-labels for individual instances. This approach may selectively aggregate instance features, such as selecting the top K most representative instances to represent the entire bag's features, thus forming a bag-level representation. This strategy is suitable for scenarios that involve critical instances that play a decisive role in the final classification decision. In contrast, embedding-level methods map individual instances into a feature space and perform feature aggregation through specific operators to generate bag-level representations. These operators can be simple statistical operations, such as taking the mean or maximum, or more complex techniques like attention mechanisms and neural network layers. This approach captures richer information in WSIs as it considers not only the inter-instance relationships but also adapts to the unique distribution of instances within each bag.

According to existing literature, embedding-level methods often outperform instance-level methods in terms of performance [30,32]. On the one hand, embedding-level methods better preserve the inter-instance relationships and structural information during feature aggregation. On the other hand, instance-level methods may lose the capture of other valuable information due to over-reliance on selected instances. Therefore, when selecting the appropriate MIL method for WSI classification, researchers need to consider their ability to preserve image heterogeneity and extract key features, in order to optimize classification performance. With the continuous advancement of deep learning techniques,

these methods are being further improved and optimized to better address the complexity of WSI data.

### 2.3. Application of Attention Mechanism in MIL

In the early applications of MIL, traditional models often employed simple feature aggregation techniques such as mean pooling or max-pooling [33,34] to integrate information from individual instances within a bag and obtain a bag-level representation. The core idea behind these methods is that pooling operations can capture the most salient features within a bag or average features across the entire bag, thereby enabling bag-level classification. While these simple aggregators achieved some success in the initial MIL research, they have limitations when dealing with data that have complex structures and subtle differences, such as WSIs, as these methods struggle to capture important information at a fine-grained level.

In recent years, embedding-based MIL methods have made positive advancements in WSI classification tasks, particularly with the introduction of attention-based models [29,31,35–37]. The incorporation of attention mechanisms allows models to more flexibly learn the relationships between instances and the importance of each instance for the bag-level classification task. The model proposed in [29] was the first attempt to combine attention mechanisms with MIL, marking a significant advancement in the field. In this model, an auxiliary side branch network is utilized to learn the weight values, i.e., attention scores, for each instance, enabling the model to emphasize instance features that are more relevant for classification while suppressing irrelevant information. The weight values of instances are then used to obtain a weighted bag-level representation, significantly enhancing the model's understanding of complex structures within WSIs. Following [29], many research works have adopted and improved upon this attention-based MIL framework [6,18,30,31]. For example, the DSMIL [18] determines weight values by calculating the cosine distance between instances and the most representative instance. This approach potentially identifies other instances that are similar to the most representative instance and assigns them higher weights. Another example is TransMIL [30], which calculates attention scores based on mutual information between instances, highlighting the interaction and interdependence among instances. Leveraging this mutual information helps the model better capture spatial contextual information within WSIs, thus providing richer and more detailed representations at the bag level.

These attention-based models have not only achieved promising results in WSI classification but have also spurred the application of MIL in the field of medical image analysis. Each model places emphasis on different aspects of instance weight calculation, showcasing adaptability to diverse data distributions and characteristics. These innovative approaches offer new perspectives and tools for addressing complex biomedical problems, contributing to the accuracy and interpretability of automated pathology diagnosis systems. With ongoing advancements in deep learning and attention mechanisms, it is anticipated that these attention-based MIL models will evolve to become even more sophisticated and specialized, addressing a broader array of challenging biomedical image analysis tasks.

## 3. Method

We now present our method for weakly supervised WSI classification. In this section, we shall introduce the formulation of MIL and present our model: DAMIL.

### 3.1. Background: MIL Formulation

Taking the binary classification problem of MIL as an example, given a bag of instances  $X = \{x_1, x_2, x_3 \dots x_n\}$ , where  $n$  is the number of instances in the bag, there is a potential semantic relationship between the instances of  $x_1, x_2, x_3 \dots x_n$ , but their respective labels  $y_1, y_2, y_3 \dots y_n$  are unknown. We need to predict the label  $Y \in \{0, 1\}$  of bag  $X$ , which can be defined as



$$Y(X) = \begin{cases} 0, & \text{iff } \sum_{i=1}^n y_i = 0 \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

i.e., the label of a bag is positive if there is at least one positive instance in the bag and negative otherwise. The model for predicting the label  $Y$  of a bag  $X$  can be described as

$$Y(X) = c(a(f(x_1), f(x_2), \dots, f(x_n))) \quad (2)$$

where  $c$  is the classifier that achieves the final output, and  $a$  and  $f$  have different meanings depending on how the MIL is modeled. If the modeling is based on an instance-level approach, then  $f$  is the rater that classifies and scores all instances in the bag, while  $a$  is the aggregator that aggregates features for instances previously selected based on the scoring. If the modeling is based on an embedding-level approach, then  $f$  is the feature extractor that performs feature extraction on all instances, while  $a$  is the aggregator that further generates the bag representation based on all instance features.

Currently, most MIL models used for WSI classification employ attention models. In this type of MIL model, the feature aggregator  $a$  is mainly composed of two modules: the attention module and the classifier. The attention module is responsible for calculating the weights of all instance features and aggregating them to obtain a bag representation  $b(x)$ . The classifier then utilizes the bag representation obtained in the previous step to output the prediction results. The entire process of learning and prediction can be described as follows [38]:

$$Y(X) = c(b(x)) \quad (3)$$

$$b(x) = a(f(x_1), f(x_2), \dots, f(x_n)) = \sum_{i=1}^n \alpha_i f(x_i) \quad (4)$$

$$\alpha_i = \text{softmax}(\mathcal{O}_a(x_i)) \quad (5)$$

where  $\alpha_i$  is the weight coefficient of the instance feature and  $\mathcal{O}_a$  is a neural network with a nonlinear activation function.

### 3.2. DAMIL Proposed

The entire pipeline of the classification framework consists of the following parts:

1. Image segmentation. The WSI is segmented into patches according to a set size, these patches can be overlapping or non-overlapping, and the patches with too little organization regions are discarded according to a set threshold.

2. Feature extraction. The patches are embedded into the feature vectors space of [1,512] by pre-trained Resnet18, all the feature vectors from each WSI form a matrix; the number of rows of the matrix varies because the number of patches in each WSI segmentation is not consistent.

3. Classification based on DAMIL. The feature vector matrix corresponding to each WSI is fed into DAMIL for classification prediction.

Our key innovation is to consider the whole feature vector matrix as a 1D feature map from the dimension of instance quantity, i.e., to consider the feature dimension of each feature vector matrix as a channel dimension and the dimension of the number of instances of each feature vector matrix as a one-dimensional spatial dimension (figuratively speaking, to flatten out a 2D feature map into a 1D feature map, as shown in Figure 3), and, thus, to design a novel module for capturing potential semantic relationships between instances. Specifically, we propose DAMIL, which consists of two essential components: a module for extracting channel attention weights and a module for extracting spatial attention weights. Figure 1 demonstrates the architecture of DAMIL. Given a feature matrix  $F \in \mathbb{R}^{N \times C}$  as input, DAMIL will first encode the features in equal dimensions, and then sequentially extract the channel-based attention weights  $H_c \in \mathbb{R}^{1 \times C}$  and the

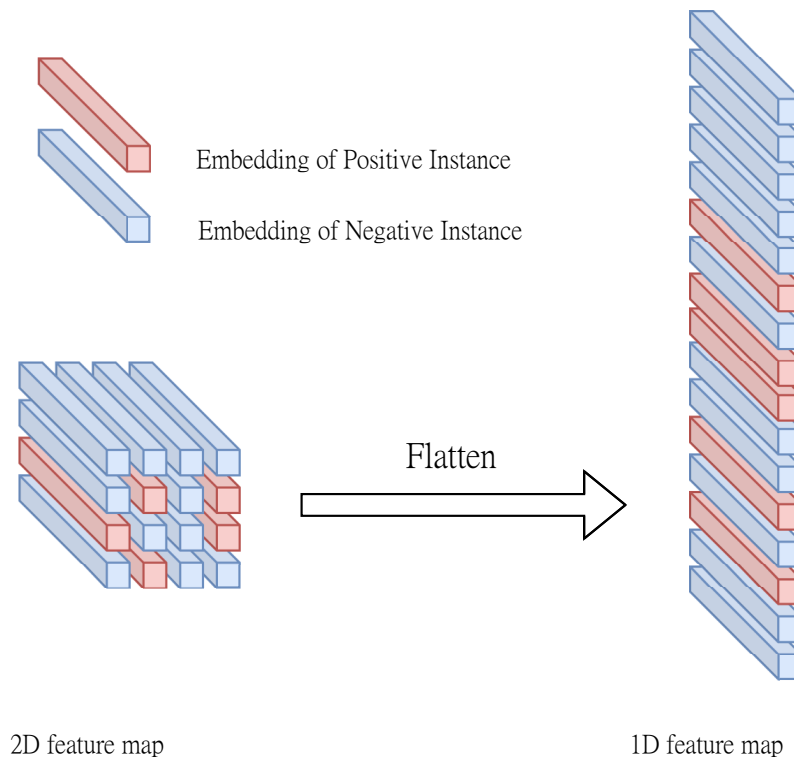
space-based attention weight  $H_s \in \mathbb{R}^{N \times 1}$ ; the complete dual-attention extraction process can be described as follows:

$$F_1 = H_c(F) \otimes F \quad (6)$$

$$F_2 = H_s(F_1) \otimes F_1 \quad (7)$$

$$F_3 = F_2 + F \quad (8)$$

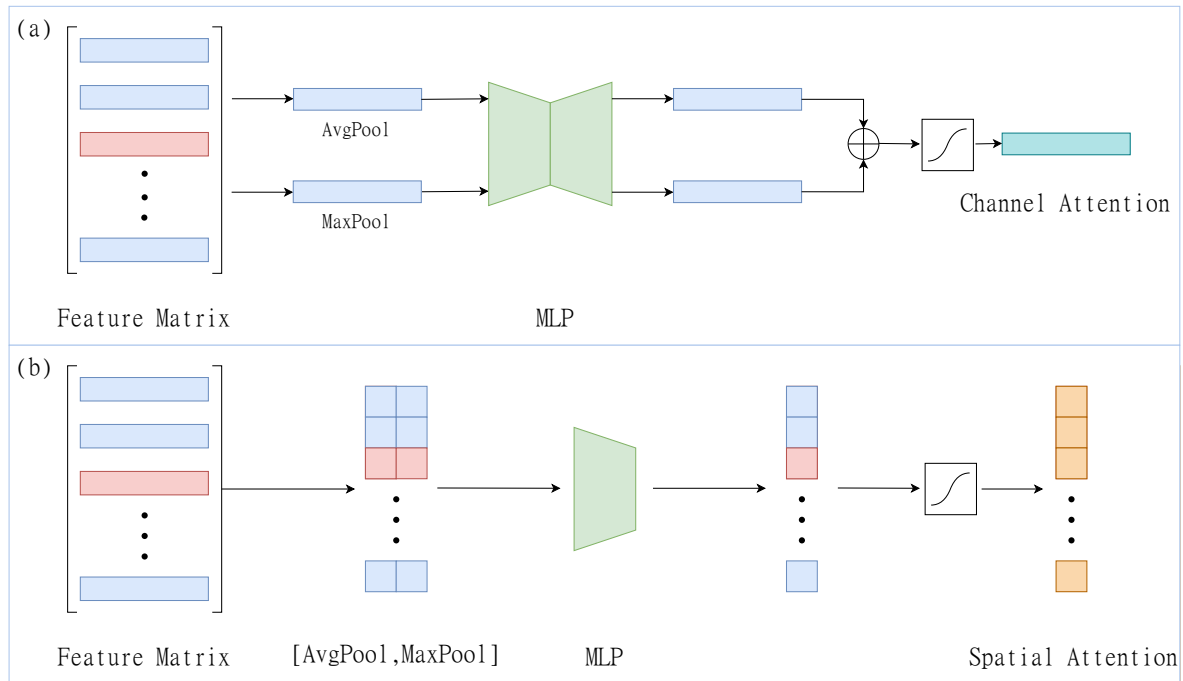
where  $\otimes$  represents the element-wise multiplication of the matrix based on the broadcasting mechanism, and  $F_3$  is the final output of the model based on the residual linkage. Figure 4 describes the computational process of each attention module, and each attention module is explained in detail below.



**Figure 3.** Graphical representation of the transformation from a 2D feature map to a 1D feature map.

### 3.3. Channel Attention Module

We utilize the inter-channel relationships of all instances in the bag to compute the channel attention weights. Since each channel of the feature vector matrix describes a particular aspect of the instance, the channel attention weights mainly describe which features are the most meaningful among all the instance features. In terms of aggregation of spatial information, most of the work uses average pooling [39,40]. However, ref. [41] argues that max-pooling can capture another important aspect of feature information, which can be utilized to calculate more refined channel weights. Therefore, we adopt a strategy that combines max-pooling and average pooling followed by concatenation. Subsequent experimental results demonstrate that taking both max-pooling and average pooling does outperform taking a single pooling.



**Figure 4.** Illustration of each attention submodule. As depicted in the diagram, (a) illustrates the channel attention module, while (b) illustrates the spatial attention module. Both attention modules utilize max-pooling and average pooling for their outputs. Channel attention compresses the dimension of instance quantity for pooling operations, while spatial attention compresses the channel dimension for pooling operations.

We first compressed the spatial dimensions of the feature vector matrix using average pooling and max-pooling to obtain two [1, 512] feature vectors representing the average and maximum pooled features, respectively. Subsequently, these two feature vectors are nonlinearly mapped using an multilayer perceptron (MLP) with a nonlinear activation function, and then the outputs were element-wise summed and converted to the final channel attention weights using the sigmoid function. The whole process is shown in Figure 4a, which can be described as follows:

$$H_c(F) = \text{sigmoid}(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \tag{9}$$

### 3.4. Spatial Attention Module

Similar to computing spatial attention weights based on 2D feature maps, we compute spatial attention weights for different instances based on the dimension of instance quantity of the feature vector matrix. Compared to channel attention, spatial attention can effectively highlight the more important instances [42], which is just complementary to channel attention. Similar to the computation of channel attention, we first compress the channel dimension using max-pooling and average pooling to generate two feature vectors, which represent the maximum and average pooled features of the whole channel, respectively. We concatenate these two feature vectors to form a dual-channel 1D feature map. This feature map is then nonlinearly mapped and downscaled by an MLP with a nonlinear activation function, and, finally, sigmoid-transformed into the final spatial attention weights. The specific operation is shown in Figure 4b, which can be described as follows:

$$H_s(F) = \text{sigmoid}(\text{MLP}(\text{AvgPool}(F) \parallel \text{MaxPool}(F))) \tag{10}$$

where  $\parallel$  denotes the concatenation operation of two feature vectors.



### 3.5. Combination Mode of Dual-Attention Modules

For the feature vector matrix obtained from the WSI, the two attention modules mentioned above can work individually or collaborate, highlighting the most meaningful features and important instances, respectively. When the two modules cooperate, there are two types of connections: serial and parallel. Through experimental verification, we found that the serial connection performs better than the parallel connection. Specifically, the connection sequence of channel first and then spatial attention achieves the optimal experimental results, which will be discussed in detail in the subsequent section. After being processed by the dual-attention modules, a feature vector matrix with the same dimension as the original input is obtained. This feature vector matrix undergoes max-pooling and average pooling based on the dimension of instance quantity. The resulting two feature vectors are concatenated and fed into an MLP with a nonlinear activation function for nonlinear mapping and dimensionality reduction. Finally, the predicted probabilities for package categories are obtained.

## 4. Experiments and Results

In this section, we conducted a comprehensive performance comparison between our approach and the current state-of-the-art methods. Additionally, we further validated the contributions of various modules in our framework to the prediction results through a series of ablation experiments.

### 4.1. Dataset

We present our experimental findings based on the publicly available dataset Camelyon16 and TCGA Lung Cancer. Camelyon16 is the most widely utilized dataset in WSI classification research, comprising 399 WSI images of breast cancer screenings classified into two categories: normal and tumor. Due to the detailed pixel-level annotations provided for positive WSIs, this dataset is not only used for classification but also extensively employed in the task of ROI localization [18]. The TCGA Lung Cancer dataset includes two subtypes of lung cancer: lung adenocarcinoma and lung squamous cell carcinoma. It comprises a total of 1054 diagnostic digital slides available for download from the National Cancer Institute's data portal. We randomly divided the whole-slide images (WSIs) into 839 training slides and 210 testing slides, discarding five slides due to poor quality or damage. During the data preprocessing stage, we initially segmented each WSI into image patches with a size of  $256 \times 256$  pixels, at both a  $20\times$  and  $10\times$  magnification level. This segmentation approach ensured that each image patch contained sufficient details, facilitating subsequent feature extraction and classification. Throughout this process, we specifically excluded image patches with less than 35% tissue region coverage, ensuring input data quality and reducing interference from irrelevant background in subsequent analysis. Similar to the methods employed in references [18,31], we utilized a pre-trained Resnet18 neural network to map each image patch to a 512-dimensional feature space. Resnet18 was chosen as our feature extractor due to its depth and ability to capture complex features. This approach allowed us to extract a set of expressive feature vectors from the original image patches, providing a strong foundation for model learning.

### 4.2. Implementation Details and Evaluation Metrics

During the training process of the model, we employed the Adam optimizer with an initial learning rate of 0.0001. The weight decay was set to  $1 \times 10^{-5}$ . To adjust the learning rate, we used the cosine annealing schedule. Furthermore, in order to prevent prolonged ineffective training and promptly capture the optimal model during the training process, we incorporated an early stopping mechanism into our training strategy. The batch size for training the MIL model was set to 1 (i.e., each batch contained only one bag). We conducted the experiments on hardware equipped with an RTX 4090 GPU.

Considering the imbalanced nature of the dataset, we assessed the performance metrics in all experiments using not only accuracy (ACC) and F1-Scores but also prioritized the

area under the curve (AUC) as the primary measure. AUC exhibits lower sensitivity to class imbalance and provides a more comprehensive reflection of the model's classification ability. ACC was calculated using a threshold of 0.5 across all experiments. To ensure the robustness and reliability of the evaluations, we employed a 10-fold crossvalidation method, which ensured that each data point had a chance to be used for validation, allowing for a more objective assessment of the model's performance. Finally, we reported the average performance metrics obtained on the 10 validation subsets to provide a comprehensive evaluation of the overall model performance.

We present the performance comparison results between DAMIL and the state-of-the-art MIL methods on the CAMELYON-16 and the TCGA Lung Cancer dataset. To provide a comprehensive evaluation, we selected several representative MIL methods for comparison: (1) MeanPooling and max-pooling, instance-level methods that directly aggregate instances to obtain bag-level representations; (2) ABMIL, a classical MIL model based on attention mechanism that assigns weights to different instances; (3) DSMIL, utilizing non-local attention pooling technique to enhance the model's recognition ability for key instances; (4) TransMIL, a transformer-based architecture that leverages the powerful global dependency modeling capability of transformers; and (5) CLAM, incorporating a clustering constraint mechanism during the multiple instance learning process to enhance the model's selectivity and diversity of instance features, thereby improving its discriminative power. The experimental results of all these models are from their official implementations and adopt the same hyperparameter settings as our proposed model. This approach ensures the accuracy and fairness of our evaluation and comparison results, enabling us to accurately measure the performance improvement of DAMIL compared to existing models.

The experimental results, as shown in Tables 1 and 2, reveal that the MIL approach utilizing both max-pooling and average pooling techniques falls short of achieving an accuracy rate (ACC) and area under the curve (AUC) higher than 65% across different magnification levels. This further emphasizes the superiority of embedding-based methods over instance-based ones. When analyzing the CAMELYON-16 dataset, a significant reality that cannot be overlooked is the relatively small proportion (less than 10%) of tumor regions in the majority of positive whole-slide pathology images (WSIs). This characteristic poses a challenge to many MIL models in effectively learning representative patch-level representations for accurate classification. Consequently, the experimental results on this dataset fully demonstrate a model's competence in addressing this issue. In this challenging context, our DAMIL model showcases its exceptional adaptability and classification performance. In particular, at 20x magnification, DAMIL outperforms the state-of-the-art models by 2.25% in ACC, 3.76% in AUC, 6.24% in recall rate, and 4.1% in F1-Score, respectively. Of noteworthy mention is DAMIL's minimal standard deviation in AUC, recall rate, and F1-Score, showcasing the model's stability and robustness. Similar trends are observed in the experiments conducted at 10x magnification, where DAMIL also achieves the highest ACC, AUC, and F1-Score, surpassing other optimal models by 2.51%, 1.68%, and 3.42% respectively. By comparing the results obtained at the two different magnification levels, the superior performance of the DAMIL model in handling highly imbalanced datasets becomes apparent. These findings not only highlight DAMIL's efficacy in slide-level labels but also demonstrate the potential for its application in the field of pathological image analysis.

Notably, although the new model incurs increased computational overhead, this investment largely translates into significant performance gains. The added computational complexity primarily stems from the dual-attention mechanism and more sophisticated feature extraction steps. These enhancements greatly improve the model's accuracy and stability, especially when handling complex pathological images. Specifically, the dual-attention mechanism enhances the model's ability to identify key regions and adapt to varied pathological environments. While this requires additional computational resources, experimental results indicate that this investment is worthwhile. To further enhance the model's scalabil-

ity, we plan to employ model compression and parallel computing strategies in the future, which will help reduce computational costs while maintaining performance.

**Table 1.** Experimental results of the CAMELYON-16 dataset (magnified 20×). The corresponding standard deviations are indicated by ±. The best experimental results are highlighted in bold.

Method	Accuracy	AUC	Recall	F1-Score	FLOPs	Model Size
Mean pooling	0.6391 ± 0.0291	0.6227 ± 0.0559	0.1500 ± 0.1070	0.2342 ± 0.1375	62.1 M	521.7 K
Max-pooling	0.6341 ± 0.0334	0.5344 ± 0.0886	0.2062 ± 0.1769	0.2735 ± 0.1965	62.1 M	521.7 K
ABMIL [29]	0.8547 ± 0.0384	0.8438 ± 0.0718	0.7188 ± 0.1112	0.7951 ± 0.0652	77.8 M	652.1 K
DSMIL [18]	0.8420 ± 0.0594	0.8107 ± 0.1146	0.6750 ± 0.1553	0.7652 ± 0.1098	116.9 M	851.6 K
CLAM [31]	0.8601 ± 0.0411	0.8736 ± 0.0418	0.7500 ± 0.0977	0.8151 ± 0.0644	94.4 M	787.3 K
TransMIL [30]	0.8472 ± 0.0379	0.8642 ± 0.0504	0.6750 ± 0.0874	0.7777 ± 0.0622	610.8 M	2.64 M
DAMIL	<b>0.8772 ± 0.0397</b>	<b>0.9018 ± 0.0387</b>	<b>0.7812 ± 0.0675</b>	<b>0.8361 ± 0.0509</b>	189.4 M	1.43 M

**Table 2.** Experimental results of the CAMELYON-16 dataset (magnified 10×). The corresponding standard deviations are indicated by ±. The best experimental results are highlighted in bold.

Method	Accuracy	AUC	Recall	F1-Score
Mean pooling	0.6367 ± 0.0197	0.6074 ± 0.0495	0.1062 ± 0.0662	0.1833 ± 0.1013
Max-pooling	0.6440 ± 0.0357	0.5783 ± 0.0836	0.2938 ± 0.2021	0.3600 ± 0.1930
ABMIL [29]	0.8346 ± 0.0393	0.8302 ± 0.0565	0.6625 ± 0.1388	0.7561 ± 0.0787
DSMIL [18]	0.7920 ± 0.0842	0.7863 ± 0.1121	0.5188 ± 0.2146	0.6390 ± 0.2177
CLAM [31]	0.8413 ± 0.0351	0.8392 ± 0.0646	0.6812 ± 0.0622	0.7737 ± 0.0510
TransMIL [30]	0.8019 ± 0.0514	0.8128 ± 0.0715	<b>0.6938 ± 0.0906</b>	0.7372 ± 0.0627
DAMIL	<b>0.8597 ± 0.0375</b>	<b>0.8470 ± 0.0690</b>	0.6750 ± 0.1012	<b>0.7903 ± 0.0696</b>

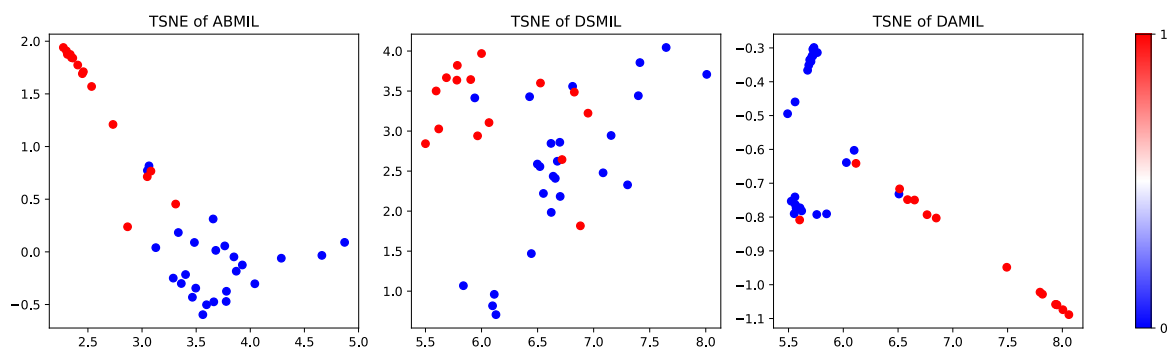
Additionally, we conducted supplementary experiments on the TCGA Lung Cancer dataset, as shown in Table 3. This dataset is renowned for its high heterogeneity and diverse sample sources, featuring significant genetic and epigenetic differences across various subtypes such as lung adenocarcinoma and lung squamous cell carcinoma. Furthermore, sample imbalance and potential technical noise adds complexity to the clinical setting. In this challenging context, the DAMIL model demonstrated exceptional competitiveness and broad adaptability. Consistent with the results on the CAMELYON-16 dataset, DAMIL also excelled in this experiment, particularly in terms of stability, with leading ACC, AUC, recall, and F1-Scores, further underscoring its robustness. In this new experimental environment, DAMIL not only maintained outstanding performance but also exhibited excellent adaptability to diverse pathological image data, reinforcing our conclusion that DAMIL is a powerful tool for handling imbalanced and complex pathological image data. These results not only highlight DAMIL’s efficiency with slide-level labels but also underscore its broad potential for application in the field of pathological image analysis.

**Table 3.** Experimental results of the TCGA Lung Cancer dataset (magnified 20×). The corresponding standard deviations are indicated by ±. The best experimental results are highlighted in bold.

Method	Accuracy	AUC	Recall	F1-Score
Mean pooling	0.7218 ± 0.0235	0.7423 ± 0.0515	0.2135 ± 0.0382	0.3146 ± 0.0265
Max-pooling	0.7449 ± 0.0331	0.7903 ± 0.0721	0.3622 ± 0.0651	0.4261 ± 0.1032
ABMIL [29]	0.8541 ± 0.0160	0.8463 ± 0.1012	0.6825 ± 0.1301	0.7726 ± 0.0412
DSMIL [18]	0.8332 ± 0.0275	0.8595 ± 0.1306	0.6203 ± 0.0147	0.7581 ± 0.0674
CLAM [31]	0.8615 ± 0.0191	0.8761 ± 0.0754	0.6964 ± 0.0521	0.8192 ± 0.0486
TransMIL [30]	0.8472 ± 0.0177	0.8624 ± 0.0972	0.7143 ± 0.1002	0.7653 ± 0.0615
DAMIL	<b>0.8703 ± 0.0207</b>	<b>0.9123 ± 0.0146</b>	<b>0.7692 ± 0.0514</b>	<b>0.8436 ± 0.0713</b>

### 4.3. Experimental Analysis

In order to delve into the sophistication of our DAMIL framework in the field of MIL, we have employed T-SNE, an exceptional nonlinear dimensionality reduction technique, to conduct clustering analysis and visualization processing on the bag-level feature vectors generated by several state-of-the-art models on the test set. As revealed in Figure 5, our visualization results transform data points in the high-dimensional feature space into interpretable two-dimensional charts, where the red and blue dots symbolize positive and negative WSIs, respectively. In Figure 5, it is evident that feature vectors with the same labels cluster together well, forming distinct partitions. Moreover, there are clear boundaries between clusters of different labels, showcasing a high degree of discrimination among the feature vectors. This not only unveils the powerful capability of the DAMIL model in feature learning, but also emphasizes its noticeable advantage in distinguishing different categories. The experimental results strongly validate the effectiveness of DAMIL in accurately identifying and categorizing bag-level features from complex biomedical image datasets. These clustering tendencies and clear classification boundaries directly reflect the potential of the model in practical applications, particularly in clinical settings requiring fine discrimination. Through these intuitive visualizations, the performance of the DAMIL model is thoroughly showcased, thereby providing a reliable foundation for our research in the field of MIL.



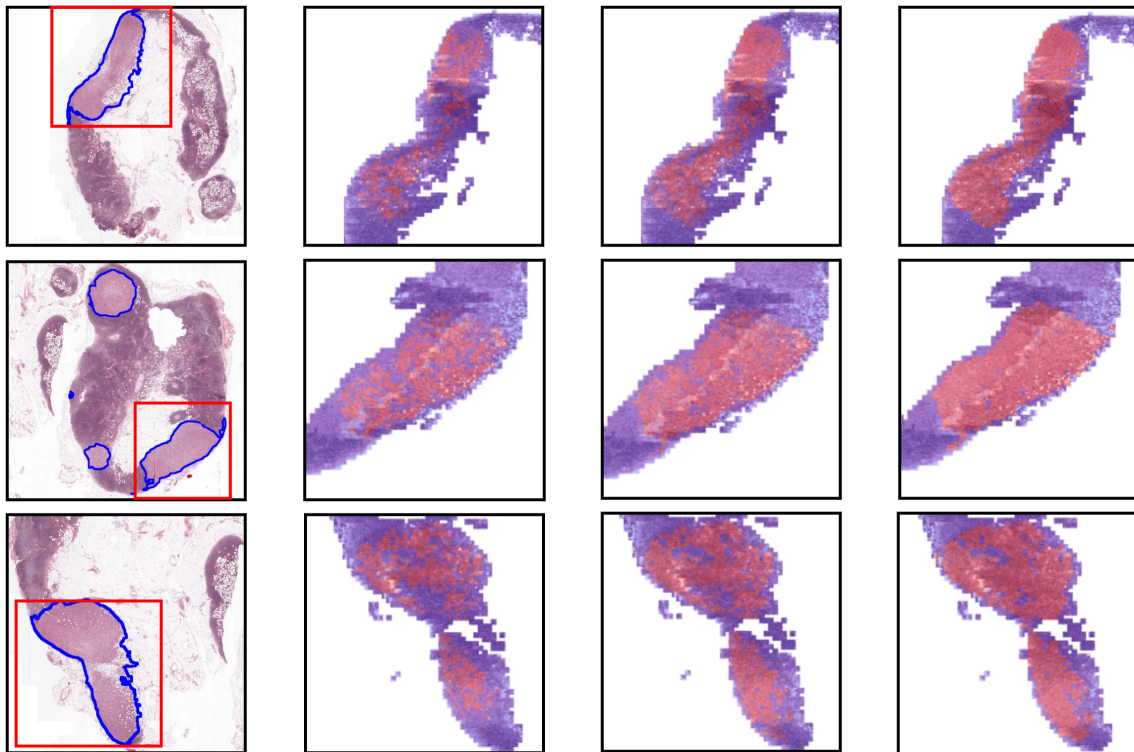
**Figure 5.** Visualization of the clustering of package representations generated by the model using T-SNE. From left to right, the clustering results for ABMIL [29], DSMIL [18], and DAMIL.

The core of the DAMIL framework lies in its advanced processing pipeline, which first encodes the input feature vector matrix with equal dimensions. It then sequentially passes through the channel attention module and spatial attention module to highlight the most influential features and crucial instances in the data, respectively. This dual-attention strategy effectively captures the complex interactions between deep features and instances. Ultimately, with the help of a decoder, the model re-maps the encoded features for accurate prediction and classification.

To further enhance the interpretability of the results, we employed the Grad-CAM technique. This technique reveals the contribution of each instance in the network's predictions by assigning an activation value to each instance. We normalize these activation values and precisely map them back to the corresponding spatial positions of the original WSI, generating intuitive visualizations. As shown in Figure 6, this figure includes the results of comparative analysis, where the first column displays pixel-level annotations of the WSI, and the subsequent three columns represent the interpretable heatmaps generated by the ABMIL, DSMIL, and DAMIL models at the corresponding positions in the WSI.

Upon careful observation of Figure 6, we can clearly see that the DAMIL model not only identifies more positive instances than the ABMIL and DSMIL models but also exhibits higher accuracy in depicting the boundaries between tumors and normal tissues. Particularly commendable is the high consistency between the interpretable heatmaps generated by DAMIL and the pixel-level annotations. This not only demonstrates DAMIL's

exceptional performance in capturing details but also reinforces its practical value in medical image processing, especially in scenarios that require precise determination of tumor edges. Through this in-depth visual analysis, we not only elucidate the effectiveness of the DAMIL model but also provide the medical community with a powerful tool to improve diagnostic accuracy and enhance the ability to interpret model decisions.



**Figure 6.** Interpretable heatmap of a WSI. The initial column displays pixel-level annotations of lymph node metastasis in a WSI, while the subsequent columns showcase the interpretable heatmaps corresponding to the red-boxed regions of the WSI acquired via ABMIL [29], DSMIL [18], and DAMIL, respectively.

#### 4.4. Ablation Study

We conducted a series of ablation experiments based on the CAMELYON-16 dataset, including comparative experiments of different methods for computing spatial and channel attention, as well as comparative experiments for different combination patterns of the two attention modules. The details of the experiments are described in detail below.

The channel attention mechanism is devoted to highlighting key features in the network and achieving this goal by performing pooling operations in the dimension of instance quantity. This study delved into various approaches to effectively capture channel attention, including the separate use of average pooling, the separate use of max-pooling, as well as the combination of these two pooling methods. In the final approach, we employed an MLP with shared parameters to parallelly process the results of both pooling methods, ultimately obtaining a complete representation of channel attention through element-wise summation. According to the comparative results shown in Table 4, the strategy of combining average pooling and max-pooling significantly outperforms individual pooling methods in terms of performance. The reason for this significant performance improvement may be that average pooling and max-pooling represent two complementary ways of extracting information. Average pooling preserves the overall background information in channels by extracting the average value of features, while max-pooling highlights key signals by focusing on the strongest responses. This fusion approach allows the network to integrate global and local information, forming a more comprehensive and balanced feature representation. Moreover, the parameter sharing mechanism of MLP further enhances learning efficiency,



enabling the model to finely adjust attention allocation to features. This comprehensive approach not only fully utilizes the advantages of the two pooling techniques but also provides an effective way to accurately capture and leverage features that contribute to the task, thereby significantly improving the overall performance of the model.

In the realm of advanced feature representation, spatial attention plays an integral role, aiming to emphasize crucial instances, which is achieved through pooling operations along the channel dimension. In order to extract effective spatial attention, this study explored different strategies, namely employing average pooling, max-pooling, and a combination of both. Unlike the generation of channel attention, for spatial attention, we experimented with two integration techniques: one through element-wise addition and the other through concatenation followed by processing with a fully connected layer. As depicted in Table 5, experimental results demonstrate that generating spatial attention by concatenating the results of average and maximum pooling, followed by a fully connected layer, yields significantly superior effects compared to other experimental settings. The underlying reason behind this outcome can be attributed to the concatenation strategy, which offers a richer and more detailed combination of features. Unlike element-wise addition, this approach avoids losing any information from individual pooling operations. Concatenation preserves both the global features induced by average pooling and the salient features extracted by max-pooling, which can be comprehensively optimized and integrated during subsequent processing with a fully connected layer. As a result, the model gains a more precise ability to identify and emphasize crucial instances, thereby crucially enhancing its predictive performance. This signifies that, when designing attention mechanisms, the rational fusion of different information types is pivotal in enhancing the efficacy of spatial attention.

To validate the performance of DAMIL, we have devised a series of experiments aimed at assessing the individual usage of the modules as well as their combined applications. Furthermore, we have conducted comparative analyses of various approaches for integrating these modules. The results indicate a significant improvement in experimental outcomes when the channel attention and spatial attention modules are fused together, as depicted in Table 6. Further analysis reveals that integrating the channel attention module before the spatial attention module yields superior performance. The effectiveness of this combination can be examined from several perspectives. First, the introduction of the channel attention module plays a crucial role in feature selection. In deep learning model processing, different feature channels often carry semantic information at varying levels and domains. The channel attention mechanism can dynamically score the importance of each channel based on its contribution to the task. This early filtering allows the model to discard irrelevant or redundant features from the outset, providing a more precise and optimized foundation for subsequent module operations. Secondly, with pre-processing by channel attention, the highlighted key information in the feature maps offers a clearer and more defined target for spatial attention. Spatial attention can then focus on detailed analysis of local areas and the exploration of feature relationships using the refined features. This processing order not only enhances spatial attention's sensitivity to local patterns but also increases the model's overall discriminative power by reducing interference from secondary information. Moreover, this strategy of selection before localization fully exploits the potential of attention mechanisms to refine information progressively. Throughout this process, the flow of information is optimized step by step, transforming from vague to clear, from scattered to focused. This hierarchical application of attention equips the model with a layered ability to analyze complex scenarios, maintaining high accuracy and robustness when handling various input types. From an information theory perspective, this configuration can be viewed as a process of progressive entropy reduction, thereby improving the model's information utilization efficiency and learning robustness. By mitigating interference from irrelevant information, this approach aids in developing more efficient and universally applicable deep learning architectures. In conclusion, these findings expand the understanding of leveraging attention mechanisms to enhance model performance in



the field of deep learning, providing strong theoretical and empirical support for future research directions and performance optimization strategies.

We have not only explored the effects of two attention modules on predictive performance, but also compared and analyzed the individual impacts of the encoder, decoder, and residual connections in the framework. As shown in Table 7, when the encoder, decoder, and residual connections are removed separately, there is a decrease in predictive performance. Specifically, the absence of the encoder has the least impact on the model's performance, while the absence of residual connections has the most significant negative impact, as evidenced by both the AUC value and recall rate evaluation metrics. The encoder plays a vital role in extracting and transforming the input data features, so the removal of the encoder may indicate that the model cannot fully capture the relevant information in the input data. However, the subsequent modules may compensate for this loss to some extent, given their inherent strength. On the other hand, the decoder is responsible for reconstructing the target output based on the output of the dual-attention model. Its absence may directly result in a deficiency in predictive functionality. As for residual connections, they are commonly used to alleviate the issue of gradient vanishing during network training, ensuring the effective flow of information through network layers. Without residual connections, network training may become more challenging, which explains why their absence has the greatest impact on predictive results.

**Table 4.** Comparison of different implementations of channel attention. The corresponding standard deviations are indicated by  $\pm$ . The best experimental results are highlighted in bold.

Method	Accuracy	AUC	Recall	F1-Score
AvgPool	0.8747 $\pm$ 0.0310	0.8758 $\pm$ 0.0579	0.7312 $\pm$ 0.0725	0.8226 $\pm$ 0.0478
MaxPool	<b>0.8772 <math>\pm</math> 0.0320</b>	0.8652 $\pm$ 0.0490	0.7375 $\pm$ 0.0968	0.8253 $\pm$ 0.0542
AvgPool + MaxPool	0.8772 $\pm$ 0.0397	<b>0.9018 <math>\pm</math> 0.0387</b>	<b>0.7812 <math>\pm</math> 0.0675</b>	<b>0.8361 <math>\pm</math> 0.0509</b>

**Table 5.** Comparison of different implementations of spatial attention. The corresponding standard deviations are indicated by  $\pm$ . The best experimental results are highlighted in bold.

Method	Accuracy	AUC	Recall	F1-Score
AvgPool	<b>0.8798 <math>\pm</math> 0.0466</b>	0.8744 $\pm$ 0.0516	0.7625 $\pm$ 0.0823	0.8343 $\pm$ 0.0679
MaxPool	0.8747 $\pm$ 0.0389	0.8717 $\pm$ 0.0507	0.7438 $\pm$ 0.1158	0.8230 $\pm$ 0.0645
AvgPool + MaxPool	0.8747 $\pm$ 0.0389	0.8740 $\pm$ 0.0654	0.7312 $\pm$ 0.0934	0.8215 $\pm$ 0.0613
[ AvgPool, MaxPool]	0.8772 $\pm$ 0.0397	<b>0.9018 <math>\pm</math> 0.0387</b>	<b>0.7812 <math>\pm</math> 0.0675</b>	<b>0.8361 <math>\pm</math> 0.0509</b>

**Table 6.** Comparison of channel attention and spatial attention used alone and in different combinations. + represents the parallel connection;  $\rightarrow$  represents the series direction. The corresponding standard deviations are indicated by  $\pm$ . The best experimental results are highlighted in bold.

Method	Accuracy	AUC	Recall	F1-Score
Only Channel	<b>0.8798 <math>\pm</math> 0.0347</b>	0.8686 $\pm$ 0.0581	0.7312 $\pm$ 0.1104	0.8255 $\pm$ 0.064
Only Spatial	0.8723 $\pm$ 0.0413	0.8683 $\pm$ 0.0687	0.7312 $\pm$ 0.1319	0.8155 $\pm$ 0.0730
Channel + Spatial	0.8748 $\pm$ 0.0406	0.8863 $\pm$ 0.0537	0.7688 $\pm$ 0.1104	0.8294 $\pm$ 0.0596
Spatial $\rightarrow$ Channel	0.8772 $\pm$ 0.0462	0.8740 $\pm$ 0.0378	0.7500 $\pm$ 0.1021	0.8284 $\pm$ 0.0679
Channel $\rightarrow$ Spatial	0.8772 $\pm$ 0.0397	<b>0.9018 <math>\pm</math> 0.0387</b>	<b>0.7812 <math>\pm</math> 0.0675</b>	<b>0.8361 <math>\pm</math> 0.0509</b>

**Table 7.** Comparison of experiments with and without codec module and residual connection. The corresponding standard deviations are indicated by  $\pm$ . The best experimental results are highlighted in bold.

Method	Accuracy	AUC	Recall	F1-Score
DAMIL w/o Encoder	0.8722 $\pm$ 0.0462	0.8821 $\pm$ 0.0389	0.7812 $\pm$ 0.0793	0.8302 $\pm$ 0.0620
DAMIL w/o Decoder	0.8672 $\pm$ 0.0354	0.8796 $\pm$ 0.0540	0.7375 $\pm$ 0.0823	0.8148 $\pm$ 0.0555
DAMIL w/o ResConn	0.8748 $\pm$ 0.0388	0.8503 $\pm$ 0.0612	0.7062 $\pm$ 0.0979	0.8156 $\pm$ 0.0662
DAMIL	<b>0.8772 <math>\pm</math> 0.0397</b>	<b>0.9018 <math>\pm</math> 0.0387</b>	<b>0.7812 <math>\pm</math> 0.0675</b>	<b>0.8361 <math>\pm</math> 0.0509</b>

## 5. Conclusions and Future Work

In this study, we introduced a novel dual-attention-based framework for multiple instance learning, devised to bolster the representational prowess of MIL paradigms. Our method demonstrated marked superiority over extant methodologies upon assessment with the prevalently employed public dataset, CAMELYON-16. The cornerstone of our technological breakthrough is the adept integration of a dual-attention mechanism within the milieu of multiple instance learning. When juxtaposed with antecedent models that are reliant on a singular attention mechanism, our proposed model exhibited an enhanced capacity for discerning more distinctive representations of whole-slide images (WSIs), which is pivotal for precise label prediction. The developed model proficiently underscored the paramount instances and their intrinsic salient features. To substantiate the model's efficacy, rigorous experimental comparisons against a gamut of leading-edge models were carried out, with DAMIL demonstrating preeminence over all the current state-of-the-art models. To augment model interpretability, we elucidated the model's predictive process through visualization techniques, which revealed its proficiency in more accurately identifying positive instances and delineating clear demarcations between tumorous and normal tissues, in concordance with pixel-level annotations rendered by pathologists. We are confident that our contributions represent a significant advancement in computational pathology for tumor diagnosis assistance and anticipate that DAMIL will catalyze progressive developments in MIL-based WSI categorization.

Future research endeavors will be directed towards the conception of more sophisticated channel and spatial attention models, as well as feature aggregation frameworks, with the aim to meticulously capture salient macroscopic or microscopic features within WSIs. These attributes are hypothesized to be of paramount importance for the prediction of WSI classification, with the potential to further elevate the benchmark for experimental findings. To enhance the scalability of DAMIL and its application in practical clinical settings, we plan to explore optimization strategies for the model. This will involve employing model compression techniques to reduce computational burden and implementing parallel computing strategies to increase processing efficiency. These optimization measures will help maintain high performance while enabling the model's real-time application in clinical environments. Through these approaches, we aim to further advance DAMIL's practicality in the field of pathological image analysis, facilitating its seamless integration into clinical workflows.

**Author Contributions:** Conceptualization, D.L. and X.H.; methodology, D.L.; software, D.L. and C.L.; validation, D.L. and C.L.; formal analysis, D.L. and X.H.; data curation, D.L.; writing—original draft preparation, D.L.; writing—review and editing, C.L., X.H., and B.H.; visualization, D.L.; supervision, X.H. and B.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data used in this study are publicly available and can be accessed at <https://camelyon16.grand-challenge.org> for the CAMELYON16 dataset, accessed on 10 October 2023. The TCGA Lung Cancer dataset can be downloaded from <https://portal.gdc.cancer.gov>, accessed on 3 November 2023.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Yao, X.H.; He, Z.C.; Li, T.Y.; Zhang, H.R.; Wang, Y.; Mou, H.; Guo, Q.; Yu, S.C.; Ding, Y.; Liu, X.; et al. Pathological evidence for residual SARS-CoV-2 in pulmonary tissues of a ready-for-discharge patient. *Cell Res.* **2020**, *30*, 541–543. [[CrossRef](#)] [[PubMed](#)]
2. Xing, F.; Xie, Y.; Su, H.; Liu, F.; Yang, L. Deep learning in microscopy image analysis: A survey. *IEEE Trans. Neural Networks Learn. Syst.* **2017**, *29*, 4550–4568. [[CrossRef](#)] [[PubMed](#)]
3. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med Image Anal.* **2017**, *42*, 60–88. [[CrossRef](#)] [[PubMed](#)]
4. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Mirafior, A.; Werneck Krauss Silva, V.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **2019**, *25*, 1301–1309. [[CrossRef](#)]

5. Chen, P.H.C.; Gadepalli, K.; MacDonald, R.; Liu, Y.; Kadowaki, S.; Nagpal, K.; Kohlberger, T.; Dean, J.; Corrado, G.S.; Hipp, J.D.; et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **2019**, *25*, 1453–1457. [[CrossRef](#)]
6. Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S.E.; Zheng, Y. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18802–18812.
7. Li, B.; Keikhosravi, A.; Loeffler, A.G.; Eliceiri, K.W. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. *Med Image Anal.* **2021**, *68*, 101938. [[CrossRef](#)]
8. Sirinukunwattana, K.; Pluim, J.P.; Chen, H.; Qi, X.; Heng, P.A.; Guo, Y.B.; Wang, L.Y.; Matuszewski, B.J.; Bruni, E.; Sanchez, U.; et al. Gland segmentation in colon histology images: The glas challenge contest. *Med Image Anal.* **2017**, *35*, 489–502. [[CrossRef](#)]
9. Shoaib, M.S.; Suhail, Z. COVID-19 lungs ct scan lesion segmentation. *Found. Univ. J. Eng. Appl. Sci. HEC Recognized Y Categ.* **2024**, *4*, 21–35. [[CrossRef](#)]
10. Bazargani, R.; Fazli, L.; Goldenberg, L.; Gleave, M.; Bashashati, A.; Salcudean, S. Multi-Scale Relational Graph Convolutional Network for Multiple Instance Learning in Histopathology Images. *arXiv* **2022**, arXiv:2212.08781. [[CrossRef](#)]
11. Muslim, H.S.M.; Khan, S.A.; Hussain, S.; Jamal, A.; Qasim, H.S.A. A knowledge-based image enhancement and denoising approach. *Comput. Math. Organ. Theory* **2019**, *25*, 108–121. [[CrossRef](#)]
12. Yang, J.; Chen, H.; Zhao, Y.; Yang, F.; Zhang, Y.; He, L.; Yao, J. Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 35–45.
13. Wibawa, M.S.; Lo, K.W.; Young, L.S.; Rajpoot, N. Multi-scale Attention-Based Multiple Instance Learning for Classification of Multi-gigapixel Histology Images. In Proceedings of the European Conference on Computer Vision. Springer, Tel Aviv, Israel, 23–27 October 2022; pp. 635–647.
14. Su, Z.; Rezapour, M.; Sajjad, U.; Gurcan, M.N.; Niazi, M.K.K. Attention2Minority: A salient instance inference-based multiple instance learning for classifying small lesions in whole slide images. *arXiv* **2023**, arXiv:2301.07700. [[CrossRef](#)]
15. Qu, L.; Ma, Y.; Luo, X.; Wang, M.; Song, Z. Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. *arXiv* **2023**, arXiv:2307.02249. [[CrossRef](#)]
16. Dehaene, O.; Camara, A.; Moindrot, O.; de Lavergne, A.; Courtiol, P. Self-supervision closes the gap between weak and strong supervision in histology. *arXiv* **2020**, arXiv:2012.03583.
17. Lu, M.Y.; Chen, R.J.; Wang, J.; Dillon, D.; Mahmood, F. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv* **2019**, arXiv:1910.10825.
18. Li, B.; Li, Y.; Eliceiri, K.W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14318–14328.
19. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
20. Carbonneau, M.A.; Cheplygina, V.; Granger, E.; Gagnon, G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.* **2018**, *77*, 329–353. [[CrossRef](#)]
21. Xu, Y.; Zhu, J.Y.; Eric, I.; Chang, C.; Lai, M.; Tu, Z. Weakly supervised histopathology cancer image segmentation and classification. *Med Image Anal.* **2014**, *18*, 591–604. [[CrossRef](#)]
22. Hou, L.; Samaras, D.; Kurc, T.M.; Gao, Y.; Davis, J.E.; Saltz, J.H. Patch-based convolutional neural network for whole slide tissue image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2424–2433.
23. Quéllec, G.; Cazuguel, G.; Cochener, B.; Lamard, M. Multiple-instance learning for medical image and video analysis. *IEEE Rev. Biomed. Eng.* **2017**, *10*, 213–234. [[CrossRef](#)]
24. Kandemir, M.; Hamprecht, F.A. Computer-aided diagnosis from weak supervision: A benchmarking study. *Comput. Med Imaging Graph.* **2015**, *42*, 44–50. [[CrossRef](#)]
25. Chikontwe, P.; Kim, M.; Nam, S.J.; Go, H.; Park, S.H. Multiple instance learning with center embeddings for histopathology classification. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020, Proceedings, Part V 23; Springer: Berlin/Heidelberg, Germany, 2020; pp. 519–528.
26. Kanavati, F.; Toyokawa, G.; Momosaki, S.; Rambeau, M.; Kozuma, Y.; Shoji, F.; Yamazaki, K.; Takeo, S.; Iizuka, O.; Tsuneki, M. Weakly-supervised learning for lung carcinoma classification using deep learning. *Sci. Rep.* **2020**, *10*, 9297. [[CrossRef](#)]
27. Xu, G.; Song, Z.; Sun, Z.; Ku, C.; Yang, Z.; Liu, C.; Wang, S.; Ma, J.; Xu, W. Camel: A weakly supervised learning framework for histopathology image segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10682–10691.
28. Lrousseau, M.; Vakalopoulou, M.; Classe, M.; Adam, J.; Battistella, E.; Carré, A.; Estienne, T.; Henry, T.; Deutsch, E.; Paragios, N. Weakly supervised multiple instance learning histopathological tumor segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020, Proceedings, Part V 23; Springer: Berlin/Heidelberg, Germany, 2020; pp. 470–479.

29. Ilse, M.; Tomczak, J.; Welling, M. Attention-based deep multiple instance learning. In Proceedings of the International conference on machine learning. PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2127–2136.
30. 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. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2136–2147.
31. 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. *Nat. Biomed. Eng.* **2021**, *5*, 555–570. [[CrossRef](#)] [[PubMed](#)]
32. Wang, X.; Yan, Y.; Tang, P.; Bai, X.; Liu, W. Revisiting multiple instance neural networks. *Pattern Recognit.* **2018**, *74*, 15–24. [[CrossRef](#)]
33. Feng, J.; Zhou, Z.H. Deep MIML network. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
34. Pinheiro, P.O.; Collobert, R. From image-level to pixel-level labeling with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1713–1721.
35. Tomita, N.; Abdollahi, B.; Wei, J.; Ren, B.; Suriawinata, A.; Hassanpour, S. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA Netw. Open* **2019**, *2*, e1914645. [[CrossRef](#)] [[PubMed](#)]
36. Hashimoto, N.; Fukushima, D.; Koga, R.; Takagi, Y.; Ko, K.; Kohno, K.; Nakaguro, M.; Nakamura, S.; Hontani, H.; Takeuchi, I. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3852–3861.
37. Naik, N.; Madani, A.; Esteva, A.; Keskar, N.S.; Press, M.F.; Ruderman, D.; Agus, D.B.; Socher, R. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. *Nat. Commun.* **2020**, *11*, 5727.
38. Juyal, D.; Shingi, S.; Javed, S.A.; Padigela, H.; Shah, C.; Sampat, A.; Khosla, A.; Abel, J.; Taylor-Weiner, A. SC-MIL: Supervised Contrastive Multiple Instance Learning for Imbalanced Classification in Pathology. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 7946–7955.
39. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
41. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
42. Komodakis, N.; Zagoruyko, S. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.