ACCURATE SUBTYPING OF LUNG CANCERS BY MODELLING CLASS DEPENDENCIES

George Batchkala[1] *Bin Li*[1] *Mengran Fan*[1] *Mark McCole*[2] *Cecilia Brambilla*[3]
Fergus Gleeson[4] *Jens Rittscher*[1]
[1]IBME/BDI, Dept. of Engineering Science, University of Oxford, Oxford, UK,
[2]Dept. of Cellular Pathology, Oxford University Hospitals NHS Trust, UK
[3]Dept. of Histopathology, Royal Brompton & Harefield Hospitals, Guy's and St Thomas' NHS Trust, UK
[4]NCIMI/BDI, Department of Oncology, University of Oxford, Oxford, UK

ABSTRACT

Identifying subtypes and histological patterns is crucial for lung cancer diagnosis and treatment. Nevertheless, datasets with complete subtyping information are scarce, and most existing work has primarily focused on categorising lung cancers into fundamental types, omitting the distinction of adenocarcinoma patterns. We present a computational approach for a more comprehensive lung cancer subtyping from histology by modelling the dependencies between cancer subtypes and histological patterns in a multi-label setting. Our approach utilises slide-level labels that indicate cancer subtypes as well as the presence of cancer-associated patterns, thereby alleviating the need for labour-intensive region-based annotations. A new dataset with cancer-associated pattern labels is constructed and combined with datasets from publicly available repositories. We evaluate our model's ability to simultaneously differentiate cancer subtypes and cancer-associated patterns. The result demonstrates that our modules enabled conventional weakly-supervised classification models on multi-label problems, achieving subset accuracy of 84% when differentiating lung cancer subtypes and cancer-associated histological patterns.

Index Terms— lung cancer, computational pathology, multi-label classification, multiple-instance learning

1. INTRODUCTION

Lung cancer constitutes the primary cause of oncological mortality worldwide [1]. Squamous Cell Carcinomas (LUSC) and Adenocarcinomas (LUAD) account for more than 80% of all lung cancer cases [2, 3]. Adenocarcinomas, representing around 50% of all cases [4], are categorised into sub-classes by the presence of adenocarcinoma-specific morphological patterns: acinar, lepidic, micropapillary, papillary, solid.

Histopathological subclassification of lung adenocarcinoma patterns is particularly challenging since these patterns are often found in combination within the same tumour. Existing publicly available datasets [5] classify LUAD according to the pattern most represented in the cross-sectional area of the histological section (so-called predominant pattern). However, several studies have revealed that secondary histological patterns can be useful in the management of multiple LUAD, as their morphological features can be applied to support the diagnosis of multiple synchronous or asynchronous cancers [6]. Therefore, in addition to identifying the subtypes of cancer, this work aims to determine all the histological patterns associated with LUAD.

Progress has been made towards computationally subtyping lung cancers from H&E whole slide images (WSIs). Due to the difficulty of obtaining clinical samples with annotations and fine-grained pattern labels, many works focused on classifying weakly-labelled (slide-level labelled) slides from public databases such as TCGA and TCIA-CPTAC [7] lung cancer cohorts into LUAD, LUSC, and benign tissue [8, 9, 10]. Other groups, based on extensive region-based annotations, trained models to predict the predominant morphological pattern of LUAD from patches [5, 11]. Alsubaie *et al.* [11] validated their approach on patches from LUAD slides, while Wei *et al.* [5] used a sliding-window-based method and heuristically aggregated patch-level predictions into slide-level predictions on test dataset from Dartmouth Hitchcock Medical Center (DHMC). Qiao *et al.* [12] were first to consider both lung cancer subtyping and adenocarcinoma pattern prediction on the TCGA and TCIA-CPTAC [7] patches (not slide-level).

Unlike previous works [5, 11, 12], we attempt to jointly learn subtype and presence of LUAD patterns in a weakly-supervised fashion relying on only slide-level labels that indicate the cancer subtypes or the presence of adenocarcinoma patterns in the slide. Additionally, we formulate the problem as a multi-label problem where the sample labels are a series of binary indicators for the presence of cancer subtypes and/or adenocarcinoma patterns. This is motivated by the fact that 1) pathologists classify LUAD based on the presence of adenocarcinoma patterns, and benign and LUSC regions are free from any adenocarcinoma patterns. 2) Multiple tumour subtypes and/or adenocarcinoma patterns can present in the same slide, and some indicators in the labels can be missing due to partial observation during pathological examination. We summarise our contributions as follows.

First, we construct a multi-label dataset combining TCGA,

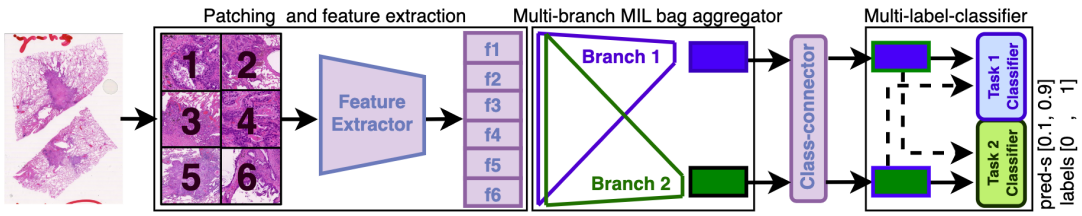

Fig. 1. **The proposed class-dependency injection framework.** Standard feature extraction (from patches 1-6 to feature vectors f1-f6) + multi-branch MIL pipeline is followed by the class-connector module, which reweighs every bag embedding from a specific branch using bag embeddings from other branches. Updated embeddings are passed to the multi-label classifier. Dashed lines show that the classifier can also be used to pass information between different tasks. Linear classifier achieves it by sharing the same weights for all tasks while communicating convolutional classifier accepts all class embeddings as input.

TCIA-CPTAC, and DHMC [5] datasets with only slide-level labels. The LUAD pattern labels are either parsed from the TCIA-CPTAC cohort-information document [7] or available alongside the DHMC dataset. We further incorporate samples from the DART lung health programme [13] with slide-level labels specifying the presence of cancer subtypes and LUAD patterns. Second, we propose a class-dependency injection method that allows the learning of robust bag representation suitable for multi-label problems under weakly-supervised conditions (see Figure 1). Popular MIL-based methods learning from weak labels (ABMIL [14], DSMIL [9], CLAM [10]) show the effectiveness of using separate attention branches for different classes but do not model the dependencies between the classes. We address this gap by injecting class dependencies through our proposed class-connector and classifier modules.

2. JOINTLY PREDICT SUBTYPES AND PATTERNS

Our goal is to train multi-label classification models to simultaneously classify lung cancer tissue into LUAD, LUSC, and benign tissue, as well as identify the presence of adenocarcinoma patterns. Our network consists of 1) a pre-trained feature extractor, 2) a multi-branch MIL bag aggregator, 3) a class-communicator module, and 4) a multi-label classifier module. We denote our whole slide dataset as $\{\mathcal{S}_k^K\}$ where K is the number of slides and $\mathcal{S}_k = \{x_n^{N_k}\}$ is a slide consisting of a patch set where N_k is the number of patches in the kth slide and x_n is the nth patch in the slide.

2.1. Feature Extraction

We followed the feature extraction process described in [9]. Patches of size 224x224 are extracted from whole slide images (WSIs) at 10x magnification ($\approx$ 0.5 microns per pixel) with no overlap. Background patches are filtered. We utilised a ResNet-18 [15] feature extractor pre-trained by [9] on the TCGA lung cohort using SimCLR [16] to extract a $F = 512$-dimensional feature embedding $h_n = f_{resnet}(x_n), h_n \in \mathbb{R}^{F\times 1}$ for each patch. The weights of f_{resnet} are frozen.

2.2. Multi-branch MIL bag aggregator

The focus of this work is to explicitly model subtype dependencies. To investigate the ability of our proposed modules for multi-label weakly-supervised whole slide image classification, we use two popular MIL backbones for training histopathology classifiers from slide-level labels: ABMIL [14] and DSMIL [9]. Both networks accept a bag of features with shape $N_k \times F$. We deploy a separate aggregation branch for each class and produce a bag embedding matrix $B = f_{mil}(\{h_1, ..., h_{N_k}\}), B \in \mathbb{R}^{C\times E}$, where E is the size of embedding corresponding to each binary class and C is the number of binary classes. In the baseline setting, a classification head is applied to the bag embedding, where each entry of the embedding matrix is scored and a prediction for the corresponding binary class is obtained.

2.3. Class Communicator

Our aim is to obtain a more accurate and holistic multi-label prediction by modelling the dependencies between the binary classes. Hence, we pass the bag embedding matrix B through a module that allows communications between the entries B_i in the bag embedding matrix, and therefore the following classification head can operate on an updated entry conditioned additionally on the entries of other classes $B_j, j \neq i$ (the context embedding). For this purpose, we adopt two common architectures, the original self-attention mechanism introduced by Bahdanau *et al.* [17] and the transformer multi-head self-attention mechanism by Vaswani *et al.* [18]. We denote the transformed bag embedding matrix as $\hat{B} = f_{cc}(B), \hat{B} \in \mathbb{R}^{C\times E}$. In our experiments, we compare the two attention methods and the baseline where an identity operation is used ($B = \hat{B}$). We use the original attention mechanism with a single alignment matrix W [17] and a single-head transformer self-attention mechanism.

1) Bahdanau self-attention on input matrix $B \in \mathbb{R}^{C\times E}$. **Alignment Scores**: For each class embedding B_i and each context embedding B_j: $A_{ij} = \mathbf{v}^T \cdot \tanh(\mathbf{W} \cdot B_i + \mathbf{W} \cdot B_j)$ where $\mathbf{v}$ is a learnable parameter vector and $\mathbf{W}$ is a learn-

	LUAD	LUSC	Benign	acinar	lepidic	micropapillary	papillary	solid
TCGA	428 / 106	403 / 109	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
CPTAC	290 / 301	283 / 236	352 / 354	91 / 101	7 / 5	8 / 10	36 / 41	29 / 13
DHMC	0 / 143	0 / 0	0 / 0	0 / 59	0 / 19	0 / 9	0 / 5	0 / 51
DART	137 / 0	24 / 0	4 / 0	130 / 0	122 / 0	97 / 0	38 / 0	39 / 0
Total	855 / 550	709 / 345	356 / 354	221 / 160	129 / 24	105 / 19	74 / 46	68 / 64

Table 1. **Train-validation (1920 slides) / Test (1249 slides).** Columns 1-3: Data Distribution of adenocarcinoma (LUAD), normal/benign, and squamous cell (LUSC) slides. Columns 4-8: Presence of five main adenocarcinoma patterns on the adenocarcinoma slides. The DHMC dataset is fully put into the test set. The DART dataset is fully in the train set.

able weight matrix applied to both B_i and B_j. **Context vectors**: $\hat{B}_i = \sum_j \text{softmax}(A)_{ij} \cdot B_j$ The output for each class is a weighted sum of all class embeddings, with the softmax-normalized alignment scores as weights.

2) Transformer multi-head self-attention on an input matrix $B \in \mathbb{R}^{C \times E}$operates as follows. First, **queries, keys, and values** are computed $Q = BW_Q$, $K = BW_K$, $V = BW_V$, where W_Q, W_K, W_V - learnable matrices. Second, **attention scores** A are computed for all pairs of classes: $A_{ij} = (Q_i \cdot K_j^T)/\sqrt{d_k} = (Q_i \cdot K_j^T)/\sqrt{E}$, where d_k is the dimension of the keys and is equal to $E = E/1$ (the number of self-attention heads is 1). Finally, reweighed class embeddings $\hat{B}_i$ are computed: $\hat{B}_i = \sum_{j=1}^{C} \text{softmax}(A)_{ij} \cdot V_j$ where V_j is the value vector for class j, and $\text{softmax}(A)_{ij}$ is the softmax-normalized attention weight.

2.4. Multi-label Classifier

The classification head takes the transformed bag embedding matrix $\hat{B} \in \mathbb{R}^{C \times E}$ as input and produces prediction $Y = f_{mc}(\hat{B}), Y \in \mathbb{R}^{C \times 1}$.

Communicating convolutional layer. Prediction for each of the C classes is computed by applying C distinct convolutional filters $W_c \in \mathbb{R}^{C \times E}$ to $\hat{B}$. This layer is implemented using a 1D convolution with kernel size E (length of an embedding) and the number of input and output channels equal to C - the number of classes (i.e., weights matrix W of shape $C \times C \times E$ and C scalar biases).

Depthwise-separable convolutional layer. Prediction for each class is computed using its own embedding $\hat{B}_c$. For each class, a filter of shape $1 \times E$ is learnt. Together, the layer is represented by a $C \times E$ matrix and C scalar biases.

Linear layer. A linear layer, containing weights of shape $E \times 1$ and a scalar bias, is shared among classes.

3. DATASET

We used three public lung cancer datasets: TCGA, TCIA-CPTAC [7], and DHMC [5]. We also collected 164 samples from the DART lung health programme [13]. Table 1 shows the distribution of known labels in the combined dataset and how we split the datasets into train-validation and test sets.

TCGA slides are annotated as either LUAD or LUSC. Hence we marked all LUAD patterns as unknown on LUAD slides and as absent on LUSC slides. We followed the same exclusion (10 slides) and split (80/20) criteria as Li *et al.* [9].

DHMC slides are LUAD with predominant pattern labels. We marked all predominant patterns as present and all the other patterns as unknown. Since it is the first time this dataset is used for pattern presence prediction and not predominant pattern classification, we fully put it into the test set to enable others to compare their algorithms with ours.

TCIA-CPTAC slides come from patients that had either LUAD or LUSC, however, many of these patients also have slides with Benign/Normal Tissue. For all LUSC and Benign/Normal slides, LUAD patterns were marked as absent (negative). For LUAD slides, the cohort information document was parsed and the pattern was marked present if it was mentioned as present, otherwise, the presence was marked as unknown. TCIA-CPTAC was the only part of our datasets that contained a considerable number of samples with benign tissue, hence we split its patients equally into train-validation and test sets stratifying on the LUAD vs LUSC label.

DART lung health dataset had the LUAD patterns explicitly annotated as present or absent for all LUAD slides by two expert thoracic pathologists. This is important since secondary LUAD patterns are also significant for clinical prognosis and treatment but are rarely reported in public datasets.

4. EXPERIMENTS AND RESULTS

Implementation and Training Details. To ignore predictions with unknown labels a masked cross-entropy loss was used. The models were trained for 10 epochs with a batch size of 1 using Adam optimiser [19] with learning rates 1e-4 and 2e-4 for ABMIL and DSMIL respectively, weight decay 0.005, beta's (0.5, 0.9). We used a Cosine Annealing [20] scheduler to vary the learning rate with T_{max}=num epochs and η_{min}=5e-6. During training, the parameters of f_{mil}, f_{cc}, and f_{mc} were optimized (see Section 2).

Results and Discussion. We evaluated the proposed modules using two popular weakly-supervised WSI classification backbones: ABMIL [14] and DSMIL [9]. Ablation studies were performed to demonstrate the performance gain of the proposed class communicator modules and multi-label

Architecture	Acc.	LUAD	LUSC	Benign	acinar	lepidic	m-papillary	papillary	solid
abmil + identity + linear	13.69	81.48	64.31	45.52	76.39	53.32	73.30	81.16	67.03
abmil + transformer + linear	17.61	84.49	26.50	31.28	86.72	88.11	98.04	83.65	87.95
abmil + transformer + c_conv	82.63	95.06	96.16	99.70	96.58	98.59	99.95	98.81	95.28
abmil + transformer + ds_conv	83.19	94.90	96.06	99.73	95.99	98.73	99.93	98.34	95.36
dsmil + identity + linear	30.26	83.50	82.71	67.15	80.72	81.80	96.98	79.88	85.37
dsmil + transformer + linear	32.83	84.69	85.99	78.55	94.85	80.27	98.72	80.55	88.03
dsmil + transformer + c_conv	69.58	88.75	94.68	97.85	92.60	92.13	98.44	86.66	89.95
dsmil + transformer + ds_conv	84.39	93.77	95.85	99.30	95.51	98.70	99.91	99.00	94.68

Table 2. **Test performance summary on comb-8 data. Column 2: Subset accuracy** calculated as the proportion of samples with fully correct predictions for all considered labels. **Columns 3-10: ROC AUC** are calculated separately for each task on the test set. For both metrics (subset accuracy, ROC AUC), predictions with unknown labels are ignored. Proportions of samples with known labels: LUAD 1, LUSC 1, Benign 1, acinar 0.68, lepidic 0.58, micropapillary 0.57, papillary 0.60, solid 0.61.

classification heads. Subset accuracy on the test set, defined as the proportion of samples for which all labels (including subtypes and patterns) were correctly predicted, are reported. The performance of individual tasks was quantified using the standard metric ROC AUC.

We tested the model in 4 scenarios: 1) comb-2: LUAD and LUSC; 2) comb-3: LUAD, LUSC, benign; 3) comb-8: LUAD, LUSC, benign, and 5 LUAD patterns; 4) comb-5: only 5 LUAD patterns. We present the subset accuracies and ROC AUC scores for the full multi-label classification task (comb-8) in Table 2 and summarise the subset accuracies for the above scenarios in a heatmap (Fig. 2). The results demonstrate that the baseline weakly-supervised classification models can be deficient in multi-label settings and achieve unsatisfactory performance (subset accuracy $< 60\%$, first column of Figure 2). After applying the class communication module, the classifier successfully learns to differentiate the 5 adenocarcinoma patterns (comb-5 in Figure 2), but still shows unsatisfactory performance for simultaneously predicting cancer subtypes and adenocarcinoma patterns (comb-8 in Figure 2). After incorporating the proposed multi-label classification head, a large performance gain is also obtained in the 8-label task where both subtypes and patterns are classified.

A plausible explanation is that the baseline models assume that the learned bag embeddings for different classes are independent, and thus each entry of the bag embedding encodes all information needed to infer the label for that class. Additionally, the baseline classification head is also shared across the classes, which assumes that the classes are lateral (no hierarchical dependency). The proposed class communication module allows the information to flow between the entries of the bag embedding, enabling each entry for each class to incorporate information from other classes. This is beneficial for capturing the dependencies between classes and yielding an enriched and more robust bag embedding. Finally, because the adenocarcinoma patterns are essentially sub-classes within LUAD, the underlying structure among the labels is not entirely lateral but rather partially hierarchical. Our multi-label classification head, which offers more flexibility over a standard linear classification head, is therefore capable of accommodating this complexity to a certain extent.

5. CONCLUSION

We proposed a novel class-dependency modelling method that can be readily incorporated into weakly-supervised whole slide classification models for multi-label problems. Our model allows information to be shared between the classification branches. Incorporating the class dependency into the model architecture resulted in a more accurate joint prediction of broad and fine lung cancer subtypes. In addition, we present a new dataset combining UK lung screening (DART) and public datasets for joint learning of lung cancer classes and adenocarcinoma patterns. The limitations of this work are 1) less than 10% of LUAD slides in the test set have positive labels for lepidic or micropapillary pattern presence, 2) most adenocarcinoma patterns reported in DHMC and TCIA-CPTAC datasets are predominant, which makes their identification easier than for secondary patterns.

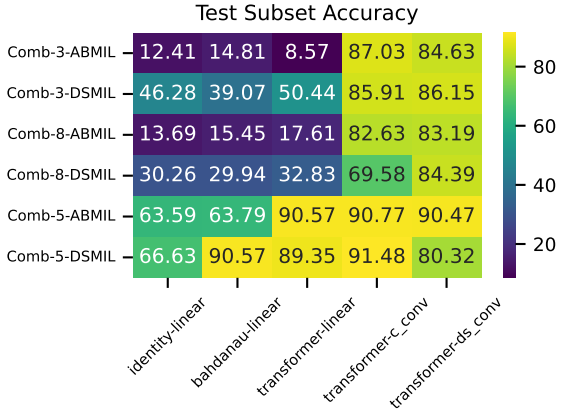

Fig. 2. Subset accuracies of data-model combinations. Y-labels show the tasks (comb-3, comb-5, or comb-8) and the MIL-aggregator architecture (ABMIL or DSMIL). X-labels show the combinations of the class-communicator (identity vs transformer) and the multi-label classification head (linear, communicating convolution - "c_conv", and depthwise-separable convolution - "ds_conv".)

6. DECLARATION

The results published here are in part based upon data generated by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) [7] and by the TCGA Research Network (https://www.cancer.gov/tcga). Ethical approval was not required as confirmed by the license attached with the open access data.

Approval for the DART lung health project was granted by the Clinical Trials and Research Governance Committee of the University of Oxford (Number: PID15885-A003-SP001, Date 24/02/2022). DART lung health project has received the following approvals. REC: 21/WM/0278 - 22/12/2021. CAG: 22/CAG/0010 - 22/12/2022. IRAS: 301420 - 24/2/2022.

The authors of this study are funded by the Innovate UK grant 40255. GB is funded by FG's A2 Research funds and supported by Health Data Science CDT (EP/S02428X/1).

7. REFERENCES

[1] L A Torre, R L Siegel, and A Jemal, "Lung Cancer Statistics," in *Lung Cancer and Personalized Medicine: Current Knowledge and Therapies*, pp. 1–19. Springer International Publishing, Cham, 2016.

[2] MR Davidson, AF Gazdar, and BE Clarke, "The pivotal role of pathology in the management of lung cancer," *J Thorac Dis*, vol. 5 Suppl 5, pp. S463–478, Oct. 2013.

[3] T Huang, J Li, C Zhang, Q Hong, D Jiang, M Ye, and S Duan, "Distinguishing Lung Adenocarcinoma from Lung Squamous Cell Carcinoma by Two Hypomethylated and Three Hypermethylated Genes: A Meta-Analysis," *PLOS ONE*, vol. 11, no. 2, pp. e0149088, Feb. 2016.

[4] R Meza, C Meernik, J Jeon, and ML Cote, "Lung Cancer Incidence Trends by Gender, Race and Histology in the United States, 1973–2010," *PLOS ONE*, vol. 10, no. 3, pp. 1–14, Jan. 2015.

[5] JW Wei, L Tafe, YA Linnik, LJ Vaickus, N Tomita, and S Hassanpour, "Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks," *Sci Rep*, vol. 9, no. 1, pp. 3358, Mar. 2019.

[6] E Kuhn, P Morbini, A Cancellieri, S Damiani, A Cavazza, and CE Comin, "Adenocarcinoma classification: patterns and prognosis," *Pathologica-Journal of the Italian Society of Anatomic Pathology and Diagnostic Cytopathology*, vol. 110, no. 1, pp. 5–11, 2018.

[7] "CPTAC-LUAD, CPTAC-LSCC. The Clinical Proteomic Tumor Analysis Consortium Lung Adenocarcinoma Collection (Version12) [dataset]. The Clinical Proteomic Tumor Analysis Consortium Lung Squamous Cell Carcinoma Collection (Version14) [dataset]," 2018.

[8] N Coudray, PS Ocampo, T Sakellaropoulos, N Narula, M Snuderl, D Fenyö, AL Moreira, N Razavian, and A Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nat Med*, vol. 24, no. 10, pp. 1559–1567, Oct. 2018.

[9] B Li, Y Li, and K W Eliceiri, "Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning," *CVPR*, pp. 14318–14328, June 2021.

[10] MY Lu, DFK Williamson, TY Chen, RJ Chen, M Barbieri, and F Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed Eng 2021 5:6*, pp. 555–570, Mar. 2021.

[11] N Alsubaie, M Shaban, DRJ Snead, SA Khurram, and NM Rajpoot, "A Multi-resolution Deep Learning Framework for Lung Adenocarcinoma Growth Pattern Classification," in *MIUA, Southampton, UK, July 9-11, Proceedings*. 2018, vol. 894, p. 311, Springer.

[12] B Qiao, K Jumai, J Ainiwaer, M Niyaz, Y Zhang, Y Ma, L Zhang, W Luh, and I Sheyhidin, "A novel transfer-learning based physician-level general and subtype classifier for non-small cell lung cancer," *Heliyon*, vol. 8, no. 12, pp. e11981, Dec. 2022.

[13] FV Gleeson and A Powell, "DART: The Integration and Analysis of Data using Artificial Intelligence to Improve Patient Outcomes with Thoracic Diseases (https://dartlunghealth.co.uk/)," 2020.

[14] M Ilse, JM Tomczak, and M Welling, "Attention-based Deep Multiple Instance Learning," in *ICML*, 2018.

[15] K He, X Zhang, S Ren, and J Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.

[16] T Chen, S Kornblith, M Norouzi, and G Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *ICML*, Nov. 2020, pp. 1597–1607.

[17] D Bahdanau, K Cho, and Y Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *ICLR*, 2015.

[18] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin, "Attention is All you Need," in *NeurIPS*, 2017, pp. 5998–6008.

[19] DP Kingma and J Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, 2015.

[20] I Loshchilov and F Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," in *ICLR*, 2017.