Medical Image Analysis 96 (2024) 103203

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Unsupervised mutual transformer learning for multi-gigapixel Whole Slide Image classification

Sajid Javed [a,*], Arif Mahmood [b], Talha Qaiser [c], Naoufel Werghi [a], Nasir Rajpoot [c,d,e]

[a] Department of Computer Science, Khalifa University of Science and Technology, Abu Dhabi, P.O. Box 127788, United Arab Emirates
[b] Department of Computer Science, Information Technology University, Lahore, Pakistan
[c] Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK
[d] Department of Pathology, University Hospitals Coventry and Warwickshire, Walsgrave, Coventry, CV2 2DX, UK
[e] The Alan Turing Institute, London, NW1 2DB, UK

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Computational pathology
Cancer imaging
Multi-gigapixel Whole Slide Images
Unsupervised learning
Vision transformer

ABSTRACT

The classification of gigapixel Whole Slide Images (WSIs) is an important task in the emerging area of computational pathology. There has been a surge of interest in deep learning models for WSI classification with clinical applications such as cancer detection or prediction of cellular mutations. Most supervised methods require expensive and labor-intensive manual annotations by expert pathologists. Weakly supervised Multiple Instance Learning (MIL) methods have recently demonstrated excellent performance; however, they still require large-scale slide-level labeled training datasets that require a careful inspection of each slide by an expert pathologist. In this work, we propose a fully unsupervised WSI classification algorithm based on mutual transformer learning. The instances (i.e., patches) from gigapixel WSIs are transformed into a latent space and then inverse-transformed to the original space. Using the transformation loss, pseudo labels are generated and cleaned using a transformer label cleaner. The proposed transformer-based pseudo-label generator and cleaner modules mutually train each other iteratively in an unsupervised manner. A discriminative learning mechanism is introduced to improve normal versus cancerous instance labeling. In addition to the unsupervised learning, we demonstrate the effectiveness of the proposed framework for weakly supervised learning and cancer subtype classification as downstream analysis. Extensive experiments on four publicly available datasets show better performance of the proposed algorithm compared to the existing state-of-the-art methods.

1. Introduction

Despite considerable advancements in cancer diagnosis and treatment, it continues to be a primary contributor to global mortality (Fitzgerald et al., 2022; He et al., 2019). Each year, approximately 20 million new cancer cases are reported, imposing a substantial burden on the healthcare system (Sung et al., 2021). Visual examination of tissue slides, often stained with Hematoxylin and Eosin (H&E) dyes, has been considered the *gold standard* for cancer diagnosis in clinical practice (Rindi et al., 2018; Srinidhi et al., 2021; Lu et al., 2021a; Lipkova et al., 2022; Hosseini et al., 2019). Modern day digital slide scanners can digitize tissue slides into high-resolution multi-gigapixel Whole-Slide Images (WSIs) at 250*nm* per pixel, with each image containing several billions of pixels and making the direct applications of machine learning methods a challenge (Srinidhi et al., 2021; Jaume et al., 2021; Li et al., 2021a; Guan et al., 2022; Wu et al., 2022; Zhang et al., 2022a; Chen et al., 2021; Di et al., 2022; Chen et al., 2022a). Computational pathology has recently emerged as an essential area that deals with the research and development of novel machine learning methods for gigapixel WSIs with applications to early cancer detection (Esteva et al., 2019; Ardila et al., 2019) and precision medicine (Cui and Zhang, 2021; Fuchs and Buhmann, 2011; Tizhoosh and Pantanowitz, 2018; Srinidhi et al., 2021). Recent developments in this area have demonstrated excellent performance in various clinical tasks for analyzing tumor micro-environment, survival prediction, and response to therapy (Bilal et al., 2021; Chen et al., 2022b; Lu et al., 2021a; Lipkova et al., 2022; Lu et al., 2021b; Chen et al., 2020; Ge et al., 2022).

Due to their huge size, annotating WSIs at the region level for fully supervised training (Fig. 1(a)) is a costly and time-consuming task for pathologists (Zheng et al., 2022). To address this challenge, Multiple Instance Learning (MIL) based weakly-supervised methods have recently been proposed that require only WSI-level labels (Fig. 1(b)) for WSI classification (Li et al., 2021a; Guan et al., 2022; Di et al., 2022; Srinidhi et al., 2021; Lu et al., 2021b). Although MIL methods have

* Correspondence to: Khalifa University of Science and Technology, Abu Dhabi, P.O. Box 127788, United Arab Emirates.
E-mail addresses: sajid.javed@ku.ac.ae (S. Javed), arif.mahmood@itu.edu.pk (A. Mahmood).

https://doi.org/10.1016/j.media.2024.103203
Received 7 September 2023; Received in revised form 30 March 2024; Accepted 13 May 2024
Available online 21 May 2024

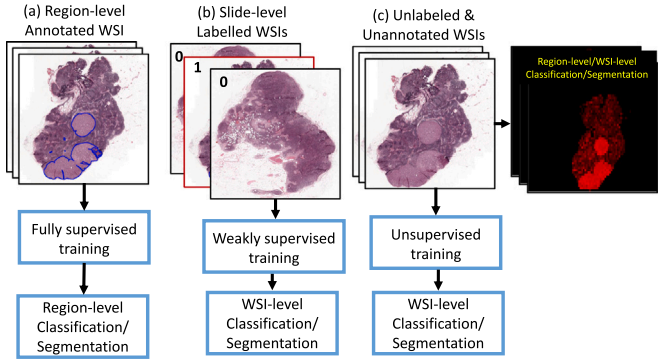

Fig. 1. Comparison of different types of supervision for WSI classification task. (a) Fully-supervised training requires region-level normal or tumor annotation (Zhang et al., 2022b; Srinidhi et al., 2021). (b) Weakly supervised training requires slide-level labels (Lu et al., 2021b; Li et al., 2021a; Guan et al., 2022). (c) The proposed unsupervised training requires neither region-level annotations nor slide-level labels for WSI classification. The red region in the detection map shows the predicted tumor regions using our proposed unsupervised algorithm.

reduced the cost compared to the region-level annotation, an expert pathologist still has to exhaustively inspect all regions consisting of several hundreds of thousands of cells within each WSI and assign a label to each slide (Chen et al., 2021a; Chuang et al., 2021; Bejnordi et al., 2017; Tomczak et al., 2015). Such inspection is still expensive and time-consuming and may limit the size of labeled WSIs dataset resulting in overfitting of MIL methods, poorly learned features, and degraded performance. In the current work, we move one step forward by proposing a fully unsupervised WSI classification algorithm that requires unlabeled WSIs as input and learns to predict instance-level disease positive/negative prediction (Fig. 1(c)). This problem is challenging yet rewarding as it may completely eradicate the cost of obtaining laborious region-level annotations and slide-level labels from pathologists and enable classification systems to be deployed without human intervention.

Unsupervised learning methods have often been considered not using any human supervision, such as different clustering methods including K-means, TSNE, and spectral clustering (Ezugwu et al., 2022). A closely related set of methods include self-supervised learning techniques which aim to produce robust representations invariant to data augmentation and different types of noises (Jing and Tian, 2020; Liu et al., 2022). Wang et al. incorporated contrastive learning in transformer models to improve the performance of self-supervised learning for WSI classification (Wang et al., 2022). Vu et al. proposed H2T representation which is learned from unsupervised clustering techniques applied to WSI patches (Vu et al., 2023). Chen et al. recently proposed the HIPT method by leveraging the natural hierarchical structure in WSI using self-supervised learning (Chen et al., 2022a). These approaches provide robust representations which are then utilized for WSI classification as a downstream analysis task.

In the current work, we propose an unsupervised WSI classification algorithm that generates slide-level labels without human intervention. We exploit the fact that the number of disease-negative instances (WSI patches) is significantly larger than the number of disease-positive instances within WSI training datasets. For instance, in the CAMELYON-16 dataset (Bejnordi et al., 2017), there are 0.85 m positive patches and 1.38 m negative patches. Therefore, if a learning mechanism such as an auto-encoder is trained without using positive or negative labels, it will better learn to represent the negative patches. Our algorithm is inspired by observing the behavior of the auto-encoder reconstruction error for the negative and the positive patches in the WSIs. We found that this error is often more significant for the positive patches when compared with their negative counterparts. Our interpretation is that negative patches are more homogeneous than positive ones, which exhibit larger variations in terms of texture and patterns (Turashvili and Brogi, 2017; Alizadeh et al., 2015; Marusyk and Polyak, 2010). We verified this fact by measuring the average within-patch entropy of the DCT transform of all positive patches in the CAMELYON-16 dataset as 0.881 compared to 0.556 for the negative instances. The same fact is also verified by measuring the similarity between small local windows within each patch. Using the Pearson Correlation Coefficient (PCC), we found the average within-patch PCC to be 0.357 among local windows of the positive instances as compared to the average PCC of 0.771 for negative patches.

Based on the above observations, we advocate that the reconstruction error can be leveraged to discriminate between the positive and negative patches. To that end, we proposed investigating this hypothesis using a transformer-based architecture. In the proposed algorithm, we transform input features to a latent space and then inverse transform to the original space. The latent space is learned such that the transformation error is low for the disease-negative instances and high for the disease-positive ones, acting thus as an indicator of the patch type (i.e. positive or negative). Our proposed algorithm consists of two main modules including a pseudo label generator and a label cleaner. Both modules mutually learn from each other in consecutive iterations and are trained in an end-to-end fully unsupervised manner. A discriminative learning mechanism is also proposed to enhance the discriminative capability of the transformer pseudo-label generator aimed at producing better quality labels. Our algorithm is evaluated on four publicly available WSI classification datasets, including CAMELYON-16 (Bejnordi et al., 2017) for breast cancer, The Cancer Genome Atlas (TCGA) lung cancer, TCGA for renal cell carcinoma and TCGA breast cancer (Tomczak et al., 2015). Rigorous experimental evaluations demonstrate better performance of the proposed unsupervised algorithm for WSI classification. We have also performed experiments using a weakly supervised variant of our proposed algorithm resulting in performance enhancement. We also fine-tuned our proposed unsupervised pre-trained model to perform downstream analysis tasks such as cancer subtype classification. In these experiments, the proposed algorithm consistently outperformed the existing State-Of-The-Art (SOTA) methods. We summarize our main contributions as follows:

1. We propose a fully unsupervised mutual transformer learning algorithm for WSI classification using instance-level prediction. The proposed architecture consists of two modules including a transformer pseudo label generator and transformer label cleaner, with both modules learning mutually from each other and improving the performance for instance-level classification. To the best of our knowledge, it is the first rigorous attempt to tackle the WSI classification problem in a fully unsupervised manner.
2. The transformer pseudo label generator is based on the novel idea of learning a latent space via discriminative learning such that disease-negative instances can be inverse transformed with small errors while disease-positive instances observe large transformation errors.
3. We perform rigorous experimental evaluations on four different WSI classification datasets. Cancer subtype classification is also evaluated as a downstream analysis task with weak supervision. Our results demonstrate better performance of the proposed algorithm compared to several SOTA methods.

The rest of this work is organized as follows: Section 2 presents a literature review on WSI classification methods. Section 3 describes our proposed methodology in detail. Section 4 presents the experimental evaluation while Section 6 draws the conclusion and describes the future directions of the current work.

2. Literature review

Deep learning has advanced computational pathology applications, however, the evolution has been hampered by the need for large-scale manually annotated WSI datasets. To address this problem, MIL-based weakly supervised methods have been proposed, thereby avoiding expensive and time-consuming pixel-wise annotations (Shao et al., 2021; Zhang et al., 2022a; Li et al., 2021a; Lu et al., 2021b). It has been empirically observed that a fully supervised classifier trained on a small pixel-level manually annotated dataset may overfit while a weakly-supervised classifier trained on a larger WSI-level labeled dataset may generalize better (Campanella et al., 2019). In the literature, MIL-based weakly-supervised methods have recently gained much popularity towards WSI classification (Srinidhi et al., 2021). These methods can be broadly categorized into local and global representation methods (Hou et al., 2016; Kanavati et al., 2020; Ilse et al., 2018; Lu et al., 2021b). In the local methods, the label of each tissue instance is independently estimated and all labels are aggregated to estimate the WSIs labels by averaging or max-pooling operation. In the global methods, representations of all instances within a bag are aggregated to obtain a global bag representation which is then used for the WSI classification.

Local Methods: Hou et al. proposed a patch-based CNN model to differentiate between different cancer sub-types (Hou et al., 2016). The patch-level classification results are aggregated by using a decision-based fusion model. Kanavati et al. proposed instance-level fully supervised and weakly supervised learning to predict lung cancer from WSIs (Kanavati et al., 2020). Tellez et al. proposed a neural image compression method to analyze the WSI using a patch-level encoder (Tellez et al., 2019). Lerousseau et al. proposed a weakly-supervised MIL method for tumor segmentation in WSIs using region-level annotations (Lerousseau et al., 2020). Xu et al. proposed instance-level label prediction and WSI segmentation method using slide-level labels (Xu et al., 2019). Chikontwe et al. proposed the weakly supervised WSI-level segmentation methods leveraging patch-level compression (Chikontwe et al., 2022). In these methods, only a small number of instances in each WSI contributes to the training therefore a large number of WSIs are required.

Global Methods: Ilse et al. proposed a neural network-based permutation-invariant aggregation operator to obtain global representation from histology images (Ilse et al., 2018). Lu et al. proposed a clustering-based attention method to be applied to the MIL problem for improving WSI classification performance (Lu et al., 2021b). Sharma et al. proposed an end-to-end network for clustering the WSI instances into different groups (Sharma et al., 2021). From each group, a few instances are sampled for training and an attention method is used for WSI classification. These methods assume the instance to be generated from an independent and identically distributed process however, the spatially adjacent instances within WSI are highly correlated with each other. Therefore, Shao et al. proposed transformer-based correlation, as well as both morphological and spatial information for WSI classification (Shao et al., 2021). Several other MIL-based variants are proposed for improved performance in medical imaging (Wang et al., 2018; Zhu et al., 2017; Srinidhi et al., 2021). Although, global-methods are better than local methods, however, for highly imbalanced classification problems the information of rare classes may get lost within the majority class during the features aggregation process.

Self-supervised Learning Methods: These methods aim to produce rich feature representations using a formulated supervision by the data itself. The learned representations are then employed to improve the performance of the downstream analysis tasks. These techniques can be broadly categorized into contrastive learning-based and pre-text-based methods.

The contrastive learning-based methods extract augmentation invariant information and instance discriminating features by pulling closer similar samples and pushing away dissimilar ones (Le-Khac et al., 2020). The pre-text-based methods include magnification prediction, stain channel prediction, cross-stain prediction, color reconstruction, and neighborhood image-related transformation. Several contrastive learning-based methods have been recently proposed in computational pathology. Li et al. proposed a contrastive learning framework to extract good representations to be used in MIL methods (Li et al., 2021a). Ciga et al. proposed a self-supervised learning method on large-scale histopathology datasets across multiple organs with different types of stains and resolutions (Ciga et al., 2022). The learned features are then used to train a linear classifier in a supervised manner for the downstream task. Huang et al. also proposed a self-supervised learning model which aggregates feature representations based on spatial information and correlation between different patches (Huang et al., 2021). These features are then used for survival analysis as a downstream task. Li et al. also proposed a contrastive learning-based features extraction method using self-invariance, inter-invariance, and intra-invariance between WSI patches (Li et al., 2021b). The features are then used for a linear classifier for cancer subtype classification. Abbet et al. proposed a self-supervised learning method that simultaneously learns the tissue region representation as well as the clustering metric (Abbet et al., 2020). The learned representations are then used to predict survival using colorectal cancer WSIs. Vu et al. learned holistic WSI-level representation using a handcrafted framework based on deep CNN (Vu et al., 2023). The learned representations are then utilized for distinct cancer subtype classification as a downstream analysis task. Their proposed handcrafted histological transformer (H2T) is reported to be faster an order of magnitude faster than the state-of-the-art transformers. More self-supervised learning methods can be seen in Wang et al. (2022), Chen et al. (2022a).

Although self-supervision can also be employed in our proposed framework to further improve performance, currently our method is different from the existing self-supervised learning methods. We do not propose any pre-text task nor we employ contrastive learning for unsupervised WSI classification. In contrast to existing methods which learn features using self-supervision and then employ them in supervised downstream analysis tasks, we propose a fully unsupervised WSI classification algorithm. Our proposed algorithm without using slide-level labels or region-level annotations learns to identify cancerous patches in a large repository of WSIs. Similar to the existing self-supervised learning methods, we also extend our work for downstream analysis tasks using supervised and semi-supervised settings. To the best of our knowledge, no rigorous fully unsupervised WSI classification algorithm has been found in the literature.

3. Proposed methodology

The proposed algorithm consists of a transformer pseudo label generator that assigns positive/negative labels to patches based on the transformation error and a label cleaning network. The first module consists of a transformer projector and an inverse projector module which are trained to minimize the transformation error between the original and the inverse-transformed feature vectors. The label cleaning network is also a transformer model trained to clean the noisy pseudo labels using a transformer label cleaner. The cleaned labels are then used to improve the transformer pseudo label generator in the next iteration using the discriminative learning mechanism. Both transformer pseudo label generator and pseudo label cleaner modules mutually learn from each other, improving each other iteratively for instance-level classification. For improved WSI classification, a graph smoothing mechanism is proposed as a post-processing step to suppress isolated spatially sparse positive labels.

The schematic illustration of our proposed algorithm dubbed as Unsupervised Mutual Transformer Learning (UMTL) for WSI classification is depicted in Fig. 2. The UMTL consists of five main steps including feature extraction heads, Transformer Pseudo Label Generator (TPLG),

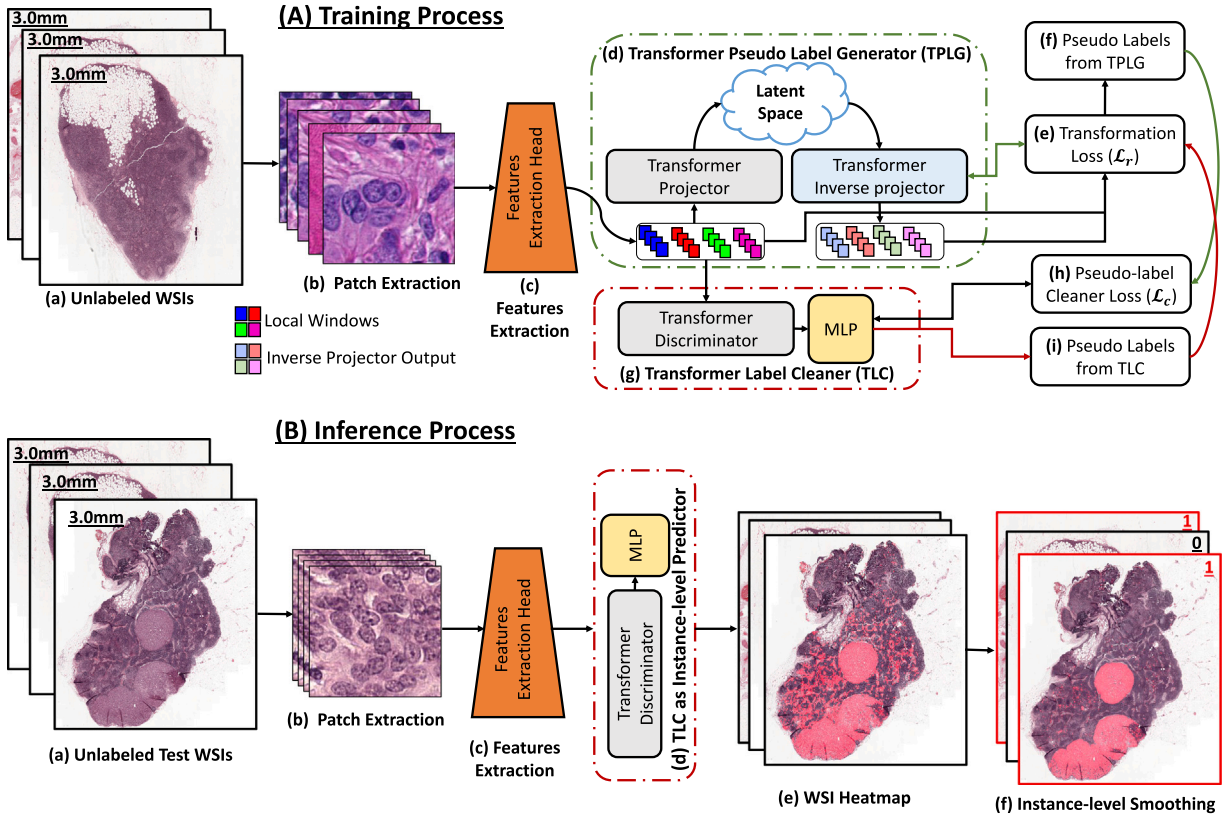

Fig. 2. System diagram of the proposed UMTL algorithm for WSI classification. (A) Training process: (a) shows the unlabeled WSIs, (b) the instances of size 224 × 224 × 3 pixels are extracted; (c) feature extraction head; (d) Transformer Pseudo Label Generator (TPLG); (e) TPLG transformation loss (Eq. (4)); (f) Pseudo labels from TPLG (Eq. (5)); (g) Transformer pseudo-Label Cleaner (TLC); (h) Pseudo label cleaner loss (Eq. (6)); and (i) Pseudo labels from TLC (Eq. (7)). (B) Inference process: (a) Unlabeled test WSIs, (b) the instances of size 224 × 224 × 3 pixels, (c) features extraction head, (d) trained TLC network employed as an instance-level label predictor, (e) predicted WSI map where red region shows the positive instances, and (f) instance-level label smoothing process and slide-level label prediction.

Transformer pseudo Label Cleaner (TLC), discriminative learning mechanism, and instance-level label smoothing for WSI classification during the inference stage. We first formulate the problem and then we explain each step in detail.

Problem Formulation: In the unsupervised WSI classification problem context, we consider each WSI as a bag consisting of multiple instances (a.k.a patches). Specifically, let $W_j = \{p_{i,j}\}_{i=1}^{n}$ be the jth WSI consisting of n instances and $p_{i,j} \in \mathbb{R}^{m\times m\times 3}$ denotes the ith instance, $1 \le j \le b$, where b is the number of WSIs. In unsupervised settings, neither the slide-level labels nor the region-level annotations are used for training. Our main goal is to estimate the slide-level label $Y_j \in \{0, 1\}$ using instance-level pseudo labels $\ell_{i,j} \in \{0, 1\}$:

$$Y_j = \begin{cases} 1 \text{ if } \sum_{i=1}^{n} \ell_{i,j} \ge \beta_{WSI} \\ 0 \text{ otherwise} , \end{cases} \tag{1}$$

where β_{WSI} is the minimum number of disease-positive instances for a WSI to be considered as positive-label. In this work, our main aim is to estimate the number of disease-positive instances within the WSIs to predict slide-level labels using our proposed UMTL algorithm.

3.1. Feature extraction and pre-processing

3.1.1. Feature extraction head

Each instance $p_{i,j}$ is input to a feature extraction head consisting of five convolutional layers which are learned such that overall loss is minimized in an end-to-end manner. The output of the feature extraction head is $f_{i,j} = F_h(p_{i,j}) \in \mathbb{R}^{m\times m\times c}$ which preserves the input instance size except for the number of channels c which are increased to $c \ge 3$. The learned features $f_{i,j}$ are re-arranged as a sequence of local windows $w_{i,j,k} \in \mathbb{R}^{a\times a\times c}$ considered as words, where $1 \le k \le n_k$, $n_k = m^2/a^2$. We also employ learnable positional encoding $u_{i,j,k} \in \mathbb{R}^{a\times a\times c}$ for each local window $w_{i,j,k}$ (Dosovitskiy et al., 2020; Carion et al., 2020). A position-aware representation $g_{i,j,k} = u_{i,j,k} + w_{i,j,k}$ is then computed and used for further processing.

3.1.2. Instance clustering as a pre-processing step

The unsupervised training of the UMTL algorithm is under-constraint due to the lack of ground-truth labels. To improve the performance of the UMTL, we propose an instance clustering-based pre-processing step to clean the training data by reducing tissue heterogeneity.

In most WSIs, the tumor region is relatively sparse while the normal region is more dominant. To discriminate between the instances belonging to these regions, we employed a simple K-means clustering method. The training data is grouped into k_o clusters using the representations obtained from the features extraction head. The k_l largest clusters are considered normal instances and used for the training in the first iteration. This pre-processing step does not completely separate the two types of instances; however, it relatively cleans the input data for better training of the proposed network in the first iteration. In the later iterations, such a pre-processing step is not required because we start getting pseudo labels from the label cleaning network which is then used for discriminative learning of the label-generating network. In our ablation study (Table 2) without instance clustering, the model has shown degraded performance. Thus, we observed empirically the need for a pre-processing step in our proposed algorithm.

3.2. Transformer Pseudo Label Generator (TPLG)

Transformers are the powerful frameworks for many tasks including image classification, object detection, and representation learning (Shamshad et al., 2022; Vaswani et al., 2017; Chen et al., 2021b; Khan et al., 2021; Carion et al., 2020). In this work, we employ a similar transformer architecture proposed by Vaswani et al. (2017). The instances are projected to latent space by using a transformer-based projector and then inverse-transformed to the original space using a transformer inverse projector. The transformation loss is then used to assign pseudo labels to each instance of the WSI.

3.2.1. Transformer projector

Our transformer projector consists of a Multi-head Self Attention (MSA) layer followed by a Multi-layer Perceptron (MLP) containing two fully connected layers. Each WSI instance is re-arranged as a sequence of position-aware word representation, $g_{i,j,k}$ which is input to the transformer projector. The projector transforms it to a learnable latent space such that $q_{i,j,k}$ be the latent representation of $g_{i,j,k}$. The input to the $1st$ layer of the projector is $p_0 = [g_{i,j,1}, g_{i,j,2}, \dots, g_{i,j,n_k}]$ and the subsequent projection steps are formulated as follows:

$$\begin{aligned} q_x &= k_x = v_x = \mathbf{LN}(p_{x-1}), \\ \hat{p}_x &= \mathbf{MSA}(q_x, k_x, v_x) + p_{x-1}, \\ p_x &= \mathbf{MLP}(\mathbf{LN}(\hat{p}_x)) + \hat{p}_x, \\ p_L &= [q_{(i,j,1)}, q_{(i,j,2)}, \dots, q_{(i,j,n_k)}], \end{aligned} \tag{2}$$

where q_x, k_x, and v_x are the query, key, and value, output by the $x-1$ projector layer and p_{x-1} is the input to that layer, $\hat{p}_x$ is the output of the MSA, p_x is the output of the MLP, and p_L is the output of the last layer of the transformer projector. Also, $x = 1, 2, \dots, L$ denotes the number of projector layers and LN represents the Layer Normalization (Ba et al., 2016).

3.2.2. Transformer inverse projector

The inverse projector assumes an opposite role to that of the projector. More specifically, the inverse projector learns an inverse mapping from the latent space to that of the original feature space. Therefore, the architecture of the inverse projector is similar to that of the transformer projector consisting of two MSA layers followed by MLP. The difference to that transformer projector is we employ an inverse projection embedding as an additional input to the inverse projector. This inverse projection embedding $b_{i,j,k} \in \mathbb{R}^{a^2 \times c}$ is learned to facilitate the inverse projection of features to the original space. The computation of the transformer inverse projector is then formulated for the xth layer where $1 \le x \le L$ and L are the total number of layers in the inverse projector.

$$\begin{aligned} z_0 &= p_L, q_x = k_x = \mathbf{LN}(z_{x-1}) + b_{i,j,k}, v_x = \mathbf{LN}(z_{x-1}), \\ \hat{z}_x &= \mathbf{MSA}(q_x, k_x, v_x) + z_{x-1}, \hat{q}_x = \mathbf{LN}(\hat{z}_x) + b_{i,j,k}, \\ \hat{k}_x &= \hat{v}_x = \mathbf{LN}(z_0), \tilde{z}_x = \mathbf{MSA}(\hat{q}_x, \hat{k}_x, \hat{v}_x) + \hat{z}_x, \\ z_x &= \mathbf{MLP}(\mathbf{LN}(\tilde{z}_x)) + \tilde{z}_x, \end{aligned} \tag{3}$$

where z_0 is the input to the $1st$ layer of the inverse projector, z_{x-1} is the input to the x-1 layer, q_x and k_x are the query and key output by the LN layer, v_x is the output of LN, $\hat{z}_x$ is the output of the MSA, $\hat{q}_x$ is the output of the LN layer, and z_x is the output of the $x-1$ layer of the inverse projector. The output of the Lth layer of the transformer inverse projector is $z_L = [\hat{g}_{i,j,1}, \hat{g}_{i,j,2}, \dots, \hat{g}_{i,j,n_k}]$. The transformation loss $\mathscr{L}_1^w(i,j,k)$ at window (i,j,k) is defined as:

$$\begin{aligned} &\mathscr{L}_1^w(i,j,k) = \|g_{i,j,k} - \hat{g}_{i,j,k}\|_1, \ \mathscr{L}_1^p(i,j) = \sum_{k=1}^{n_k} \mathscr{L}_1^w(i,j,k), \\ &\mathscr{L}_1^{WSI}(j) = \sum_{i=1}^{n} \mathscr{L}_1^P(i,j), \ \mathscr{L}_r = \sum_{j=1}^{b_t} \mathscr{L}_1^{WSI}(j), \end{aligned} \tag{4}$$

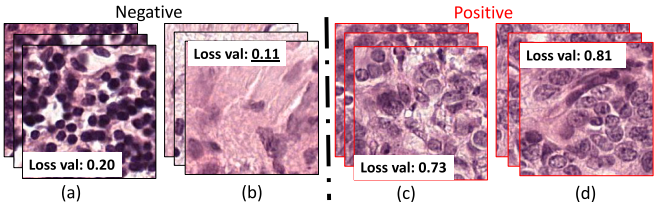

Fig. 3. Exemplar instances from positive and negative labels along with instance-level loss. (a) & (b) show instances of lymphocytes and stromal while (c) & (d) show tumor instances. Transformation loss is low for negative-labeled instances and high for positive ones.

$\mathscr{L}_1^p(i,j)$ is the loss at instance-level, $\mathscr{L}_1^{WSI}(j)$ is the loss at the WSI-level, and $\mathscr{L}_r$ is the loss of overall training data having b_t number of WSIs. During the training of the transformer projector and inverse projector, $\mathscr{L}_r$ loss is minimized. For the purpose of pseudo label generation for the ith instance in the jth WSI, a simple thresholding approach is used as:

$$\ell_{i,j} = \begin{cases} 1 \text{ if } \frac{\mathscr{L}_1^p(i,j) - \min_{Batch}(\mathscr{L}_1^p(i,j))}{\max_{Batch}(\mathscr{L}_1^p(i,j))} \ge \beta_r \\ 0 \text{ otherwise }, \end{cases} \tag{5}$$

where β_r is an instance-level threshold computed using the training data as discussed in the ablation study (see Fig. 8). In the following sub-sections, a pseudo label cleaner is proposed to further refine the pseudo labels generated by TPLG.

3.3. Transformer pseudo Label Cleaner (TLC)

To clean the noise in the pseudo labels, we propose to train a Transformer-based pseudo Label Cleaner (TLC) module as shown in Fig. 2.

The TLC is trained for the classification task in an end-to-end manner. The input to the TLC is the patch features obtained using the features extraction head and their associated pseudo labels obtained from TPLG using (5). TLC consists of a transformer discriminator followed by an MLP consisting of two fully connected layers. The transformer discriminator is similar to that of the transformer projector which projects features to a latent space which are then fed to the MLP for classification.

Once, TLC is trained using these pseudo labels it is then used to generate relatively clean pseudo labels based on the probabilities $\phi_{i,j}$ using the cross-entropy loss:

$$\mathscr{L}_c = \frac{-1}{b_t} \sum_{j=1}^{b_t} \sum_{i=1}^{n} \ell_{i,j} \phi_{i,j} + (1 - \ell_{i,j}) \ln(1 - \phi_{i,j}), \tag{6}$$

The clean pseudo labels $\ell_{i,j}^c$ are predicted using:

$$\ell_{i,j}^c = \begin{cases} 1 \text{ if } \phi_{i,j} \ge \beta_c, \\ 0 \text{ otherwise }, \end{cases} \tag{7}$$

where β_c is a threshold used to decide positive or negative label and it is estimated using the training data (see Fig. 8). These labels $\ell_{i,j}^c$ will then be utilized for the training of the TPLG in the next iteration. Both TPLG and TLC modules iteratively refine each other by mutual learning in an end-to-end manner. As a result, the performance of the proposed UMTL improves over consecutive iterations.

3.3.1. Discriminative learning of TPLG

We enhance the discrimination between positive and negative patches by proposing a discriminative learning mechanism. After the initial few iterations, the transformation target is replaced with a Gaussian random noise for the patches getting pseudo-positive labels. We investigated that this strategy improves the discrimination between

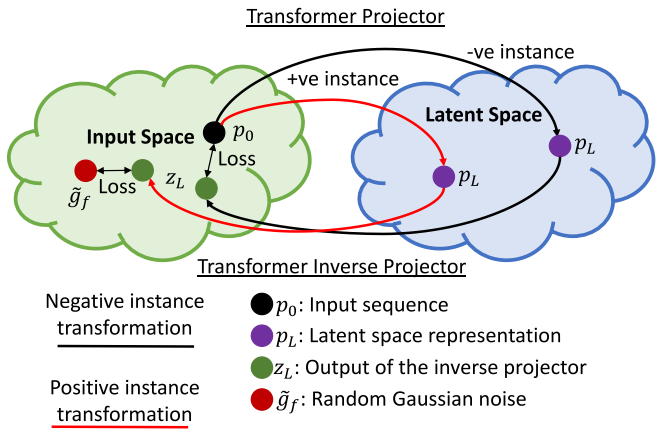

Fig. 4. A latent space is learned by Transformer Pseudo Label Generator (TPLG). The transformation error for pseudo-negative labeled instances is ensured to be low, while for pseudo-positive labeled instances, the error is aimed to be high using discriminative learning.

the transformation of the positive and the negative instances. In our experiments, during the second and onward iterations of TPLG, the pseudo labels from the TLC module are available for further training. For the negative labels, the transformation loss is measured between the original and the inverse projected features while for the positive labeled instances, the transformation loss is measured between the inverse projected features and a fixed random Gaussian noise vector as shown in Fig. 4. Such a approach will result in increased transformation error for positive-labeled instances and decreased loss for negative-labeled instances resulting in improved discriminative ability of the UMTL algorithm (see Fig. 3). Our strategy will prevent the TPLG module from learning the transformation of positive labeled instances while it will expedite learning to transform the negative labeled instances resulting in better discrimination. This is a type of negative learning where positive instances are unlearned while negative instances are better reconstructed. To make TPLG more discriminative, Eq. 4 is employed only for the negative-labeled instances, while for the positive ones, the following formulation is used for the transformation loss minimization:

$$\mathscr{L}_1^w(i,j,k) = \|\tilde{g}_f - \hat{g}_{i,j,k}\|_1, \tag{8}$$

where $\tilde{g}_f \in \mathbb{R}^{a\times a\times c}$ is a random Gaussian noise having normal distribution $N(0,1)$. We empirically observed that having a fixed noise matrix as a target better deludes TPLG than using a varying target for each positive instance.

3.4. Overall training of UMTL

The overall training of the proposed algorithm is shown in Algorithm 1. The Instance Clustering (IC) is the first step, followed by the training of TPLG by minimizing the transformation loss (Eq. (4)). Initial pseudo-labels are generated using Eq. (5), which are then used to train the TLC component using Eq. (6). The refined pseudo-labels obtained by the TLC component are used to fine-tune the TPLG component using a discriminative learning approach, resulting in improved pseudo-labels by TPLG. In the subsequent iterations, both TPLG and TLC components are iteratively trained until the convergence of the pseudo-labels.

3.5. Inference using instance-label smoothing

To infer the label of a test WSI, it is first divided into instances of size 224 × 224 similar to the size of training instances. For each instance, the transformation loss $\mathscr{L}_1^p$ is obtained from the trained TPLG. For each WSI, an instance-based spatial graph is constructed such that

Algorithm 1 Pseudo-code of the UMTL algorithm

Require: Unlabeled WSI patches, k_0, k_l, β_r, and β_c.
Ensure: Instance level labels (positive, negative).
1. Apply Instance Clustering to make k_0 clusters.
2. Select the largest clusters (k_l).
3. Train TPLG using Eq. (4).
4. Generate pseudo labels using Eq. (5).
while convergence not obtained **do**
5. Train TLC using pseudo labels.
6. Train TPLG using discriminative learning.
7. Use trained TLC for inference.

each instance is connected to its spatial neighbors according to its location within the WSI. Let A be the adjacency matrix of this graph and $\mathscr{L}_1^p$ be the attribute of node p. Assuming a smooth distribution of $\mathscr{L}_1^p$ across the spatial neighbors in WSI, we employ graph convolutions to smooth the estimated losses (Kipf and Welling, 2016).

A node attribute vector is formed using the $\mathscr{L}_1^p$ from each node in the spatial graph. In order to smooth the loss function ($\mathscr{L}_1^p$), the node attribute vector ℓ_s is multiplied by graph adjacency matrix A. The n_m such multiplications will smooth loss over n_m hop neighbors resulting in loss spatial smoothing across full WSI. This formulation ensures the smoothing of abrupt loss variations and the removal of isolated peaks within the WSI.

The resulting losses are given by $\hat{\ell}_s = \sigma(A^n \ell_s)$, where σ is an activation function, ReLU in our case. All nodes having smoothed loss greater than a threshold β_r in Eq. (5) are considered positive nodes. Based on the connectivity of the positive-labeled nodes overall label of the WSI is then inferred. A WSI is inferred as disease-positive if the size of the largest positive-labeled connected component within the spatial graph is larger than a β_{WSI} threshold value.

3.6. Weakly supervised UMTL algorithm

Most existing methods for WSI classification are trained in a weakly-supervised fashion. Therefore, we also incorporate weak supervision in our proposed unsupervised UMTL algorithm and dubbed it W-UMTL. In the first setting, we train UMTL with weak supervision for cancer vs. normal WSI classification. For more details of this setting, please refer to Section 4.3.

The second problem relates to the cancer subtype classification which requires further classification beyond just cancer vs. normal binary classification. For this purpose, we perform downstream analysis by first differentiating cancer vs. normal instances using the proposed UMTL algorithm trained in fully unsupervised settings. Then, only a TLC module is fine-tuned for cancer subtype classification of only positive instances using inherited WSI-level labels. Therefore, we dub our downstream algorithm in this setting as Downstream UMTL (D-UMTL) which is also a weakly supervised algorithm. At test time, the normal vs. cancer instances are first differentiated using UMTL and then only positive instances are further classified for a particular cancer subtype using D-UMTL. Cancer subtyping at the WSI level is performed using the same instance-level smoothing process as described in Section 3.5.

4. Experimental evaluations

We compare the performance of the UMTL algorithm with its different variants and SOTA weakly-supervised MIL-based methods on four different WSI classification datasets. To validate the effectiveness of UMTL, we use different experimental protocols including fully unsupervised, limited weakly supervised, and training for downstream analysis tasks. We have also performed ablation studies to demonstrate the contribution of each component of the proposed algorithm. In the below sub-sections, we first discuss the datasets, performance evaluation

metrics, and implementation details. Then, we discuss the evaluations of the UMTL and W-UMTL algorithms followed by the downstream analysis task.

4.1. Datasets and experimental settings

We have evaluated our proposed unsupervised WSI classification algorithm on four publicly available datasets including CAMELYON-16 (Bejnordi et al., 2017) for breast cancer, TCGA for Lung Cancer (TCGA-LC), TCGA Renal Cell Carcinoma (TCGA-RCC), and TCGA BReast CAncer (TCGA-BRCA) for predicting HER2 status (Tomczak et al., 2015). The details of each of these datasets are given in the below subsections.

4.1.1. CAMELYON-16 dataset

It contains 400 WSIs with a split of 270/130 for training/testing purposes. The training dataset consists of 159 normal slides or negative cases and 111 WSIs containing tumor regions of breast cancer metastasis considered as positive cases. Tumor regions are annotated at pixel-level and labels at slide-level are assigned by an expert pathologist. However, for training in our fully unsupervised UMTL algorithm, neither region-level annotations nor slide-level labels are used. For testing purposes, slide-level labels are used to evaluate the performance of the compared methods. The main challenge in this dataset is that the positive slides contain only small portions of the tumor.

4.1.2. TCGA lung cancer dataset

TCGA-LC dataset consists of 1046 slides of two cancer subtypes including LUng Squamous cell Carcinoma (LUSC) (Ilya et al., 2012) and LUng ADenocarcinoma (LUAD) (Network et al., 2014) and 589 normal WSIs. Within 1046 WSIs, this dataset contains 534 LUAD and 512 LUSC slides, respectively. Compared to CAMELYON-16, tumor regions are significantly larger and only slide-level labels are available in this dataset. On this dataset, two different types of experiments are performed including fully unsupervised WSI classification for cancer vs. normal using UMTL and downstream analysis task for LUSC vs. LUAD WSI classification using D-UMTL with only slide-level labels.

For UMTL, we randomly split the 1635 WSIs into 80% training and 20% testing splits while ensuring patient-level separation. For D-UMTL, we randomly split the WSIs into 836 training slides and 210 testing slides for LUAD vs. LUSC classification while ensuring patient-level separation. In both evaluations, we performed five-fold cross-validation experiments by randomly selecting the training and testing splits each time and average results are reported.

4.1.3. TCGA Renal Cell Carcinoma (RCC) dataset

This dataset contains 477 normal WSIs and 726 WSIs with three cancer subtypes including KIdney Renal Papillary cell carcinoma (KIRP) (218 WSIs) (Network, 2016), KIdney Renal clear cell Carcinoma (KIRC) (390 WSIs) (Chapel Hill Kimryn Rathmell et al., 2013), and KIdney Chromophobe renal Cell Carcinoma (KICH) (118 WSIs) (Davis et al., 2014). Similar to the TCGA-LC, random 80% & 20% training/testing splits are made while ensuring patient-level separation, and then 5-fold cross-validation experiments are performed.

Similar to TCGA-LC, experiments are performed in two different settings: fully unsupervised for cancer vs. normal WIS classification using UMTL, and cancer subtype classification (KIRP vs. KIRC vs. KICH) as a downstream analysis task using D-UMTL with only slide-level labels.

4.1.4. TCGA BReast CAncer (TCGA-BRCA) dataset

TCGA-BRCA dataset is used for the prediction of Human Epidermal growth factor Receptor 2 (HER2) status which is a critical task in clinical practice for cancer treatment and prognostication (Ilya et al., 2012). This dataset contains 608 WSIs with slide-level labels of HER2- status and 101 HER2+ status. For training and validation, 80% data with patient-level separation is used while the remaining 20% is used for testing. We employed 5-fold cross-validation for comparison with other SOTA methods. On this dataset, first, cancer vs. normal patch-level classification is performed using fully unsupervised UMTL. Then, only using the positive patches, HER2 +ve vs. -ve downstream classification is performed using D-UMTL. However, results are only reported for cancer subtype classification because normal WSIs are unavailable in this dataset.

4.1.5. Evaluation metrics

All experiments are evaluated using well-known measures including Accuracy (Acc), Area Under the Curve (AUC), and F_1 measures as reported by recent SOTA methods (Li et al., 2021a; Zhang et al., 2022a; Guan et al., 2022; Shao et al., 2021). Since region-level annotations are also available in CAMELYON-16, therefore, we also performed lesion-based evaluation using the Free-response Receiver Operating Characteristic (FROC) measure. It is defined as the average sensitivity at predefined six false positive rates: 1/4, 1/2, 1, 2, 4, and 8 FPs per WSI.

4.1.6. Implementation details

For patch extraction from WSIs, we first employed the OTSU thresholding method to separate the tissue region from the background. The tissue region is then divided into non-overlapping patches of size 224×224 at 20× magnification level. In CAMELYON-16, the number of extracted patches is around 3.7 Million (M), in TCGA-LC 12.6M, in TCGA-RCC 8.9M, and in TCGA-BRCA 5.8M. In the pre-processing step (Section 3.1.2), the instances are clustered with $k_o = 10$, and $k_l = 3$ largest clusters are retained in all experiments.

The overall architecture consists of features head and transformer layers. Our features extraction head consists of one convolutional layer followed by two ResBlocks each consisting of two convolutional layers. The first convolutional layer contains 3 input channels, 64 feature maps, and 3×3 size of kernel window. The convolutional layers in each ResBlock contain 64 input channels, 64 output channels, and 5×5 kernel size. Each transformer projector and the inverse projector contain 12 layers. The feature extraction head and transformers are pre-trained on the large-scale ImageNet dataset. The images in the IamgeNet dataset have high diversity, consisting of more than 1M natural images from 1 K distinct classes. These images contain a wide variety of texture and color information leading to effective pre-training of our transformers and features extraction head. We conducted our experiments on a DGX NVIDIA workstation with 256 GB of RAM and 4 T V100 GPUs. We trained both networks in an end-to-end manner using the Adam optimizer with 120 epochs. The initial learning rate was set as $5e^{-5}$ with a batch size of 256. Thresholds for TPLG and TLC are data-driven and found to be $\beta_r = 0.50$ in Eq. (5), and $\beta_c = 0.50$ in Eq. (7). In Eq. (1), $\beta_{WSI} = 10\%$ of the number of instances is used in all our experiments. The ablation study of these values is discussed in the ablation study Section 4.2.2.

Our TPLG and TLC contain 114M and 44M parameters, respectively. During inference, only the TLC module is used to predict the instance-level labels. The time to predict the labels of 1,000 instances of size $224 \times 224 \times 3$ takes 92 s (using a single GPU) while using multiple GPUs it reduces to 65 s. Using WSI at a 20x magnification level of size $(96K \times 56K \times 3)$, the inference time is 5.56 min on multiple GPUs.

4.2. Unsupervised WSI classification results

4.2.1. Results and discussion

Cancer vs. normal WSI classification is performed in a fully unsupervised manner using our proposed UMTL algorithm on three independent datasets including CAMELYON-16, TCGA-LC, and TCGA-RCC.

Table 1
Performance of the proposed algorithm for cancer vs. normal WSI classification in two different settings including fully unsupervised algorithm UMTL and Weakly-supervised UMTL (W-UMTL) on three datasets. For UMTL, 0% WSI labels are used for both FROC and AUC. For the W-UMTL variant, different percentages of WSI labels are used and AUC is reported using the testing splits of each dataset. The lesion-based evaluation is also performed in the CAMELYON-16 dataset in a fully unsupervised manner and FROC is reported.

Datasets	0% FROC	0% AUC	30% AUC	40% AUC	50% AUC	60% AUC	70% AUC	80% AUC	90% AUC	100% AUC
CAMELYON-16	0.388	0.844	0.867	0.901	0.922	0.941	0.949	0.951	0.961	0.966
TCGA-LC	–	0.856	0.835	0.865	0.894	0.918	0.935	0.951	0.971	0.975
TCGA-RCC	–	0.822	0.881	0.902	0.922	0.941	0.966	0.977	0.985	0.991

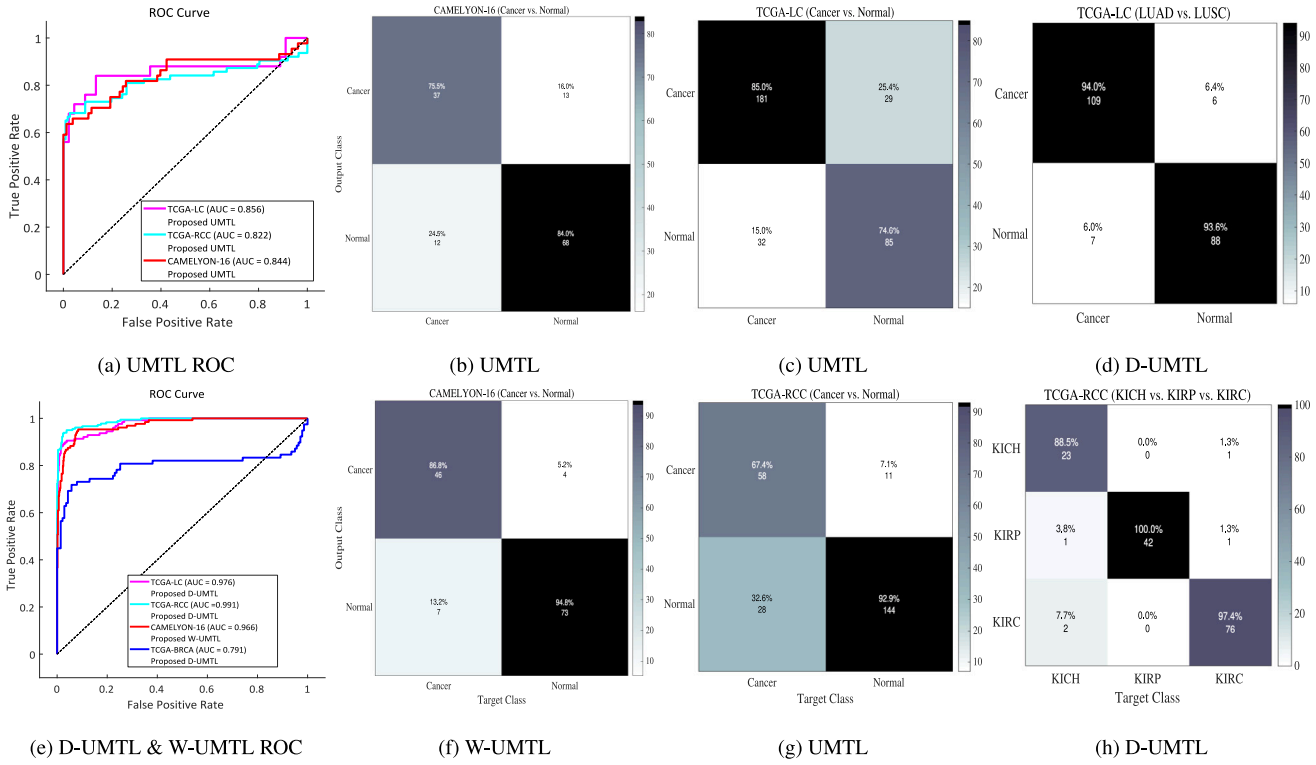

Fig. 5. Performance of the proposed UMTL, D-UMTL, and W-UMTL algorithms in terms of ROC and confusion matrices on the test sets of four independent datasets including CAMELYON-16, TCGA-LC, TCGA-RCC, and TCGA-BRCA. (a)-(c) and (g) show the performance of the proposed UMTL in terms of ROC and confusion matrices on three datasets for cancer vs. normal classification. (d) shows the performance of D-UMTL on the TCGA-LC dataset for LUAD vs. LUSC classification. (e) shows the performance of the proposed D-UMTL and W-UMTL in terms of ROC on four datasets. (f) shows the performance of the proposed W-UMTL on the CAMELYON-16 dataset for cancer vs. normal classification. (h) shows the performance of the D-UMTL on the TCGA-RCC dataset for KICH vs. KIRP vs. KIRC classification.

No existing fully unsupervised methods could be found in the literature therefore, we have to make comparisons with weakly-supervised methods where necessary.

CAMELYON-16 dataset: For this dataset, two experiments are performed in a fully unsupervised manner including lesion segmentation and WSI classification.

For the case of lesion segmentation, using 0% labels or annotations, cancerous lesions are segmented using our proposed UMTL algorithm. In this experiment, we obtained 38.8% performance as reported in Table 1. Our performance is better than some existing weakly-supervised methods including Mean Pooling, Max-Pooling, and RNN-MIL, and comparable with classic AB-MIL as shown in Table 4. The unsupervised lesion segmentation obtained by the UMTL algorithm is shown in Fig. 6. A visual comparison with region-level ground-truth annotation reveals the effectiveness of the unsupervised lesion segmentation estimated by the proposed UMTL algorithm.

For unsupervised WSI classification results, we obtained 84.40% performance in terms of AUC as shown in Table 1, ROC curve of UMTL in Fig. 5(a), and confusion matrix in Fig. 5(b). Among the existing weakly-supervised methods, our proposed UMTL algorithm is comparable with PT-MTA, classic AB-MIL, and Max-Pooling while better than the Mean Pooling method (see Table 4).

TCGA-LC dataset: For this dataset, fully unsupervised WSI classification is performed achieving 85.60% performance using the proposed UMTL algorithm as shown in Table 1.The ROC curve and confusion matrix are also presented in Figs. 5(a) and (c). Our performance is better than the weakly supervised PT-MTA method.

TCGA-RCC dataset: For this dataset, in a fully unsupervised settings our proposed UMTL algorithm obtained 82.20% AUC performance for WSI classification as shown in Table 1. The ROC curve and confusion matrix are also presented in Figs. 5(a) and (g).

4.2.2. Ablation studies and analysis

Since there are no existing fully unsupervised WSI classification methods, therefore we use several variants of our proposed UMTL algorithm for detailed performance comparisons. Some of these variants are designed by exclusion or inclusion of different components as mentioned in Table 2. Therefore, the performance variations reflect the relative contribution of each component while the UMTL has

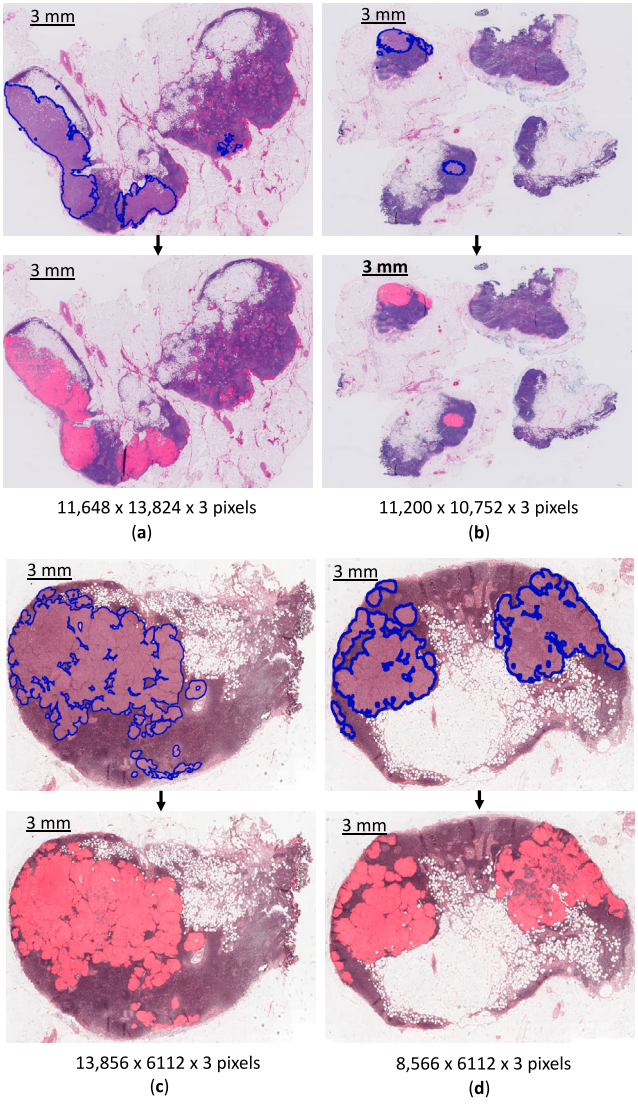

Fig. 6. Visualization of instance labels obtained by UMTL algorithm. (a)–(b) show two different WSIs, (c)–(d) Show two subfields of the same WSI selected from CAMELYON-16 test set. The top row shows the ground-truth region-level tumor annotation with blue boundaries. The bottom row shows the positive instances with pink color while the remaining region shows the negative instances.

Table 2
Ablation (Abl) studies on the UMTL using CAMELYON-16 test-set. Instance Clustering IC is a pre-processing step, Transformer Psuedo Label Generator (TPLG), Auto-encoder Psuedo Label Generator (APLG), Discriminative Learning (DL), MLP Label Cleaner (MLC), Transformer Label-Cleaner (TLC), and Instance-Label Smoothing (ILS) components are evaluated.

Variant	IC	TPLG	DL	TLC	ILS	F_1	Acc	AUC
UMTL	✓	✓	✓	✓	✓	0.751	0.832	0.844
$UMTL_{v1}$		✓	✓	✓	✓	0.729	0.813	0.822
$UMTL_{v2}$	✓	✓			✓	0.712	0.791	0.803
$UMTL_{v3}$	✓	✓		✓	✓	0.733	0.810	0.822
TLC_C	✓			✓	✓	0.677	0.751	0.772
$UMTL_{v4}$	✓	✓	✓	MLC	✓	0.728	0.813	0.831
Auto-MLP	✓	APLG	✓	MLC	✓	0.662	0.713	0.731
IC only	✓				✓	0	0	0
$UMTL_{v5}$	✓	✓	✓	✓		0.737	0.807	0.822

demonstrated the best performance compared to all variants. These experiments are performed using the CAMELYON-16 test set under fully unsupervised settings.

1. **Significance of Instance Clustering (IC) Pre-processing Step:** In this experiment, the pre-processing Instance Clustering (IC) step (Section 3.1.2) is removed from the proposed UMTL algorithm to evaluate its significance. The resulting algorithm is dubbed as $UMTL_{v1}$. The overall F_1 performance of $UMTL_{v1}$ is degraded by 2.20% compared to UMTL which shows the contribution of the pre-processing IC step.
2. **Performance of Transformer Pseudo Label Generator (TPLG):** In this experiment, only the TPLG module is employed, while the Transformer-based Label Cleaner (TLC) module is excluded as a result, the DL step is also removed. This version of the proposed UMTL algorithm is dubbed as $UMTL_{v2}$. Compared to UMTL, the performance of $UMTL_{v2}$ is degraded by 3.90% which demonstrates that the TPLG in itself can also be used for fully unsupervised WSI classification. However, the best combination is having both TPLG and TLC modules.
3. **Significance of Discriminative Learning (DL) Step:** In this experiment, the TPLG module is modified by the exclusion of the DL step. This version of the proposed UMTL algorithm is dubbed as $UMTL_{v3}$. As a result, we only train the TPLG module using the transformation loss on both +ve and -ve instances. Since the DL step enabled iterative refinement of TPLG this refinement is also not possible in consecutive iterations. Compared to the proposed UMTL algorithm, the $UMTL_{v3}$ demonstrated 1.80% performance degradation. Therefore, the iterative refinement of UMTL using the DL step positively contributes to the performance of the overall learning algorithm.
4. **Clustering-based Pseudo Labels:** In this experiment, TPLG is removed, and the pseudo labels are generated by using IC such that the labels of the largest $k_l = 3$ clusters are used as negative and the remaining cluster labels are used as positive. Label cleaning is then performed using the TLC module. This version is dubbed as TLC_C. In this version, the Instance Label Smoothing (ILS) component is employed similarly to the proposed UMTL algorithm. Compared to the UMTL algorithm, the performance of TLC_C is reduced by 7.40%. The significant reduction in performance may be attributed to the noise in the clustering-based pseudo-labels. Compared to that the pseudo labels generated by our proposed TPLG module have reduced noise and improved the overall performance.
5. **MLP Label Cleaner:** In this experiment, a simple MLP is used as a label cleaner module. This version is dubbed as $UMTL_{v4}$. The input to the MLP is latent space features (shown in Fig. 4) and MLP is trained using the cross-entropy loss function. The performance of $UMTL_{v4}$ is 72.80% which is 2.30% less than the proposed UMTL algorithm demonstrating the relative importance of the transformer-based label cleaner.
6. **Using Autoencoder and MLP:** In this experiment, we employed a simple Auto-encoder Pseudo Label Generator (APLG) using ResNet50 instance-level features. This version is dubbed as Auto-MLP. APLG consists of five fully connected layers [1024, 512, 256, 512, 1024] and MLP is used as a Label Cleaner (MLC). The performance of Auto-MLP is 66.20% which is 8.90% less than the proposed UMTL algorithm showing the importance of transformer-based architecture both in TPLG and TLC.
7. **Significance of Instance Level Smoothing (ILS):** In this experiment, the ILS is removed from the proposed UMTL algorithm as described in Section 3.5. Instead of ILS, Eq. (1) is used for WSI-level classification with $\beta_{WSI} = 10\%$. This version is dubbed as $UMTL_{v5}$. Compared to UMTL, the performance of $UMTL_{v5}$ degraded by 1.40% in F_1 score and a 2.20% in AUC. The reduction in performance demonstrates the significance of the ILS step.
   In a different experiment, we evaluated the importance of the ILS step for WSI-level segmentation using the CAMELYON16 dataset. With the ILS step, FROC is 38.80% compared to 34.15% without the ILS step.

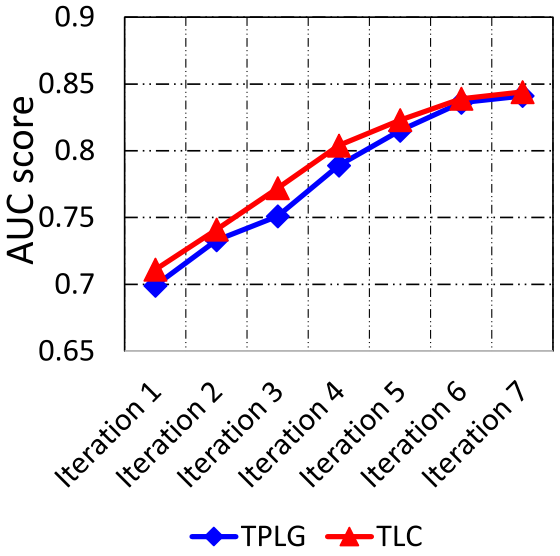

Fig. 7. Convergence analysis of the UMTL algorithm. The AUC obtained by the TPLG and TLC components increases with the increasing number of iterations.

8. **Convergence of UMTL:** In this ablation, we show the convergence of UMTL in consecutive iterations. As discussed in 3.4, in each iteration, the AUC obtained for TPLG and TLC components are plotted in Fig. 7. As the iterations proceed, both TPLG and TLC converge by achieving better AUC scores on the CAMELYON16 testing split.

4.2.3. Ablation studies on parameters tuning

1. **Selection of TPLG Threshold:** In the TPLG module, a threshold on the transformation loss is required to decide whether an instance is positive or negative. For this purpose, a threshold β_r is introduced in Eq. (5). To empirically select the value of β_r, the distribution of transformation loss is plotted over the training data as shown in Fig. 8(a). The transformation loss is scaled from 0 to 1 by dividing by the maximum loss on any instance. It is assumed that the instances with close to 0 errors are negative, while those with close to 1 are positive. In Fig. 8(a), we observed a dip in the percentage of the instances at 0.50 transformation error; therefore, we select $\beta_r = 0.50$. For a more precise selection of β_r, such a probability distribution needs to be estimated at the end of each epoch, and the value of β_r will correspond to the loss having the minimum probability.
   In a different experiment, the value of β_r is increased to 0.75 while β_c is kept at 0.50. As a result, we observed an AUC of 80.40% compared to 84.40% using $\beta_r = 0.50$ on the CAMELYON16 testing split. This demonstrates the suitability of using $\beta_r = 0.50$.
2. **Selection of TLC Threshold:** In the TLC module, a probability is generated for an instance to be positive or negative. The distribution of this probability over the training dataset is plotted in Fig. 8(b). We observed a dip in the distribution at the probability of 0.50 therefore, we select $\beta_c = 0.50$ in Eq. (7). Moreover, in this plot, we also observe a higher percentage of the instances towards 0 and 1 probabilities compared to the transformation loss plot. It demonstrates the performance of the label cleaner for pushing the tumor instances towards the probability of 1 and normal instances towards the probability of 0.
   In a different experiment, the value of β_c is increased to 0.75 while β_r is kept at 0.50. As a result, we observed an AUC of 82.40% compared to 84.40% using $\beta_c = 0.50$ on the CAMELYON16 testing split. This demonstrates the suitability of using $\beta_c = 0.50$.

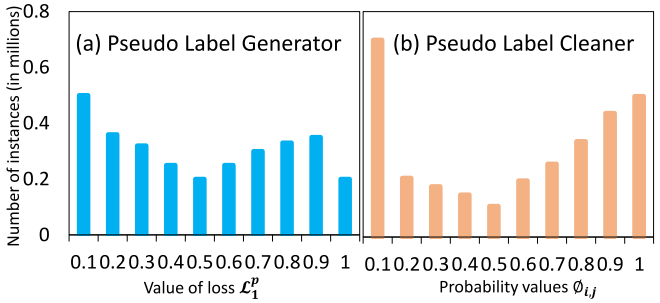

Fig. 8. (a) Distribution of transformation loss and (b) classification probabilities for CAMELYON-16 training split. The value of β_r and β_c in Eqs. (5) and (7) is set to be 0.5.

3. **Ablation on the Number of Clusters Parameter:** For the Instance Clustering (IC) pre-processing step, input instance data is grouped into k_o clusters, and k_l larger clusters are considered negative while the remaining clusters are considered as positive. In the first experiment, $k_l = 3$ is fixed and k_o is varied as 5,10,15,20, and 25. The best AUC is observed at $k_o = 10$ shown in Table 3.
   In the second experiment, $k_o = 10$ is fixed and k_l is varied as 1,1–3,1-5,1–7, and 1–9. The best AUC is observed at $k_l = 1 - 3$ as shown in Table 3. This means that the three largest clusters out of a total of 10 clusters produced the best performance.

4.3. Weakly supervised WSI classification results and comparison

The main focus of the current work is fully unsupervised WSI classification, however, currently, no such methods have been found in the literature for comparison. The nearest methods we observed are weakly-supervised MIL-based methods including Mean-Pooling and Max-Pooling as used by SOTA (Zhang et al., 2022a), RNN-MIL (Campanella et al., 2019), classic AB-MIL (Ilse et al., 2018), DS-MIL (Li et al., 2021a), CLAM-SB (Lu et al., 2021b), CLAM-MB (Lu et al., 2021b), PT-MTA (Li et al., 2019), Trans-MIL (Shao et al., 2021), DTFD-MIL (Zhang et al., 2022a), MS-ABMIL (Hashimoto et al., 2020), C2C (Sharma et al., 2021), ZoomMIL (Thandiackal et al., 2022), NIC (Tellez et al., 2019), SRCL (Wang et al., 2022), and NAGCN (Guan et al., 2022). For these methods, the results are reported by the original authors.

For a fair comparison, we trained the proposed UMTL algorithm with supervision and dubbed it as W-UMTL. For this purpose, the number of labeled WSIs in the training data is gradually increased from 10% to 100% as shown in Table 1. Since instance-level labels are not available, therefore, we make each instance inherit the label from its parent WSI. A label cleaning mechanism is then employed based on the TPLG loss which is used as a pre-trained model. The instances with a loss >0.50 from normal WSIs and those having a loss < 0.50 from positive WSIs are discarded to clean the inherited labels for an improved training process. The remaining instances are used in the end-to-end training of the TLC module. We reported the performance of the proposed weakly-supervised learning algorithm (W-UMTL) for cancer vs. normal WSI classification on three datasets including CAMELYON-16, TCGA-LC, and TCGA-RCC in Table 1. The performance of the proposed algorithm improves with increasing the level of supervision while the best performance is observed with 100% weak supervision.

Table 4 shows the weakly-supervised classification results on the CAMELYON-16 test set and its comparison with existing SOTA methods. We report W-UMTL results with 100% slide-level labels for cancer vs. normal WSI classification. For the weakly-supervised setting, we obtained an AUC of 96.60% which is better than the SOTA methods including Trans-MIL and DTFD-MIL. The ROC curve and confusion

Table 3
AUC on CAMELYON-16 by varying k_o&k_l after **1st** Epoch.

Fixed $k_l = 3$	$k_o = 5$ 0.771	$k_o = 10$ **0.798**	$k_o = 15$ 0.797	$k_o = 20$ 0.784	$k_o = 25$ 0.781
Fixed $k_o = 10$	$k_l = 1$ 0.751	$k_l = 1$ *to* 3 **0.798**	$k_l = 1$ *to* 5 0.781	$k_l = 1$ *to* 7 0.772	$k_l = 1$ *to* 9 0.766

Table 4
Performance comparison of the proposed W-UMTL algorithm with SOTA methods on the testing splits of CAMELYON-16 for cancer vs. normal WSI classification.

Methods	F_1	Acc	AUC	FROC
Mean Pooling	0.355	0.626	0.528	0.116
Max-Pooling	0.754	0.826	0.854	0.331
RNN-MIL (Campanella et al., 2019)	0.798	0.844	0.875	0.304
Classic AB-MIL(Ilse et al., 2018)	0.780	0.845	0.854	0.405
DS-MIL(Li et al., 2021a)	0.815	0.899	0.916	**0.437**
CLAM-SB (Lu et al., 2021b)	0.775	0.837	0.871	–
CLAM-MB (Lu et al., 2021b)	0.774	0.823	0.878	–
PT-MTA (Li et al., 2019)	–	0.827	0.845	–
Trans-MIL (Shao et al., 2021)	0.797	0.883	0.930	–
DTFD-MIL (Zhang et al., 2022a)	**0.882**	0.908	**0.946**	–
MS-ABMIL (Hashimoto et al., 2020)	–	0.876	0.887	–
NIC (Tellez et al., 2019)	0.766	0.802	0.714	0.299
SRCL (Wang et al., 2022)	0.866	0.922	0.942	0.455
H2T (Vu et al., 2023)	0.791	0.815	0.866	
Proposed W-UMTL	**0.895**	0.911	**0.966**	**0.476**

matrix of W-UMTL are also presented in Figs. 5(e)–(f). The weakly-supervised lesion-based evaluation resulted in 47.60% FROC which is better than the compared SOTA methods (Table 4).

Cancer vs. normal WSI classification experiments are also performed on TCGA-LC and TCGA-RCC datasets by varying the slide-level labels from 10% to 100% (Table 1). Unfortunately, on these datasets, such a classification has not been found in the literature therefore, we are not able to compare these results with any existing SOTA methods.

4.4. Evaluations on downstream analysis tasks

To compare the proposed UMTL algorithm with existing weakly-supervised methods for downstream analysis tasks we extend our method by the inclusion of weak supervision and dubbed it as D-UMTL. More details can be found in Section 3.6. We compared the proposed D-UMTL algorithm with weakly-supervised methods as well as self-supervised methods. Both of these categories of methods use weak supervision for downstream analysis tasks.

4.4.1. Comparison with weakly-supervised methods

These comparisons are performed on three distinct datasets including TCGA-LC, TCGA-RCC, and TCGA-BRCA.

Experiment on TCGA-LC dataset is performed for LUAD vs. LUSC cancer subtype classification task and the results are reported in Table 5. The proposed D-UMTL algorithm with weak supervision obtained a 97.60% AUC score outperforming all SOTA methods. The ROC curve and confusion matrix of D-UMTL are also presented in Figs. 5(e) and (d). The closest competitor is DTFD-MIL obtaining 96.10% AUC.

Similar to the TCGA-LC dataset, an **experiment on TCGA-RCC** is performed for KICH vs. KIRP vs. KIRC cancer subtype WSI classification. The results are reported in Table 5. The ROC curve and confusion matrix of D-UMTL are also presented in Figs. 5(e) and (h). The proposed D-UMTL algorithm with weak supervision obtained 97.20% Acc outperforming existing SOTA methods while obtaining comparable AUC (99.10%). The closest competitor is NAGCN obtaining 95.40% Acc and 99.20% AUC.

Table 6 shows the results of predicting **HER2 status** (either HER2+ or HER2-) on the TCGA-BRCA dataset. For the weakly supervised setting, D-UMTL obtained an AUC of 79.10% better than the SOTA approaches, including the recently proposed SlideGraph (Lu et al., 2022)

Table 5
Performance comparison of the proposed D-UMTL algorithm with SOTA methods for cancer subtype classification on TCGA-LC (LUAD vs. LUSC) and TCGA-RCC (KICH vs. KIRP vs. KIRC) datasets.

Methods	TCGA-LC			TCGA-RCC	
	F_1	Acc	AUC	Acc	AUC
Mean Pooling	0.809	0.833	0.901	0.905	0.978
Max-Pooling	0.833	0.846	0.901	0.937	0.987
RNN-MIL (Campanella et al., 2019)	0.831	0.845	0.894	–	–
Classic AB-MIL (Ilse et al., 2018))	0.866	0.869	0.941	0.893	0.970
DS-MIL (Li et al., 2021a)	0.876	0.888	0.939	0.929	0.984
CLAM-SB (Lu et al., 2021b)	0.864	0.875	0.944	0.881	0.972
CLAM-MB (Lu et al., 2021b)	0.874	0.878	0.949	0.896	0.979
C2C (Sharma et al., 2021)	–	0.873	0.938	0.919	0.987
PT-MTA (Li et al., 2019)	–	0.737	0.829	0.905	0.970
Trans-MIL (Shao et al., 2021)	0.876	0.883	0.960	0.946	0.988
DTFD-MIL (Zhang et al., 2022a)	**0.891**	0.894	**0.961**	–	–
MS-ABMIL (Hashimoto et al., 2020)	–	0.900	0.955	–	–
NAGCN (Guan et al., 2022)	–	**0.902**	0.952	**0.954**	**0.992**
NIC (Tellez et al., 2019)	0.795	0.821	0.744	0.733	0.705
HIPT (Chen et al., 2022a)	–	0.895	0.952	0.923	0.980
Prop. D-UMTL	**0.911**	**0.933**	**0.976**	**0.972**	**0.991**

Table 6
Performance of the proposed D-UMTL algorithm for HER2 status prediction on TCGA-BRCA. The AUC is reported using the test split.

Methods	AUC
RNN-MIL (Campanella et al., 2019)	0.670
Kather *et al.* (Kather et al., 2020)	0.620
Kather *et al.* (Kather et al., 2019)	0.680
Rawat *et al.* (Rawat et al., 2020)	0.710
CLAM (Lu et al., 2021b)	0.650
SlideGraph (Lu et al., 2022)	0.750
NIC (Tellez et al., 2019)	0.633
SRCL (Wang et al., 2022)	**0.766**
H2T (Vu et al., 2023)	0.702
Proposed D-UMTL	**0.791**

method. These results show the effectiveness of our transformer-based architecture for downstream analysis tasks using weak supervision.

4.4.2. MIL-based methods using UMTL as features extractor

The proposed UMTL algorithm is a fully end-to-end deep learning algorithm that is proposed for the WSI classification task. It can also be utilized as a feature extractor, similar to other methods that employ unsupervised features and MIL-based classifiers to perform downstream analysis tasks. We extract features using our trained transformer discriminator (Fig. 2(B)). The MLP is replaced with the classic ABMIL and CLAM-SB methods. On the CAMELYON16 dataset, the ABMIL method obtained 85.40% AUC, while UMTL+ABMIL obtained 88.70% AUC. Similarly, the CLAM-SB method obtained 87.10% AUC while UMTL+CLAM obtained 87.50% AUC. Thus, the UMTL features have resulted in improved performance of these MIL-based methods.

4.4.3. Comparison with self-supervised learning methods

In self-supervised learning-based methods, first, a data representation is learned in unsupervised manners then a classifier is trained for downstream analysis tasks such as cancer subtype classification. We compare the results of our proposed algorithm D-UMTL with three very recent self-supervised learning-based methods including HIPT (Chen et al., 2022a), H2T (Vu et al., 2023) and SRCL (Wang et al., 2022) shown in Table 7.

For LUAD vs. LUSC WSI classification in the TCGA-LC dataset, the proposed D-UMTL algorithm obtained best performance of 97.60% while for KICH vs. KIRP vs. KIRC subtype classification in TCGA-RCC dataset, D-UMTL performance is comparable with H2T and SRCL methods. It should be noted that self-supervised learning can also be used in our algorithm to improve performance.

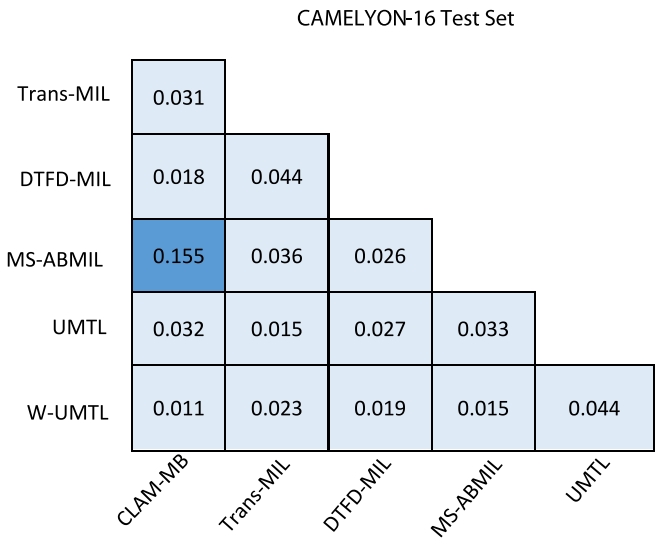

Fig. 9. Statistical analysis of the SOTA methods on CAMELYON-16 test set. p values are reported when doing right-tailed pairwise t-tests on results of the best-performing methods which are reported in Table 4. The p values are corrected using the Benjamini/Hochberg method.

4.5. Statistical analysis

We have also performed a statistical comparison of the proposed algorithm with existing SOTA methods to evaluate their closeness. For this purpose, we performed right-tailed pairwise t-tests for the results of CLAM-MB, Trans-MIL, DTFD-MIL, MS-ABMIL, UMTL, and W-UMTL. We used the CAMELYON-16 test set for cancer vs. normal classification and estimated p values using the results of the compared SOTA methods reported in Table 4 and UMTL results reported in Table 1. The p values are reported in Fig. 9. The Benjamini/Hochberg method was used to adjust the p values for multiple hypothesis tests.

As shown in Fig. 9, when using the official test set of the CAMELYON-16, all compared methods are statistically different e.g., $p < 0.05$. While only two methods including MS-ABMIL and CLAM-MB are not statistically different ($p > 0.05$). The proposed W-UMTL AUC is statistically better than that of DTFD-MIL and Trans-MIL (0.966 vs 0.946 and 0.930) with p values of 0.019 and 0.023, respectively. Overall, statistically speaking, our proposed algorithm performed better than the compared methods.

5. Limitations of the UMTL algorithm

The main limitation of the proposed UMTL algorithm is the requirement of class imbalance. Particularly, positive instances must be smaller in number compared to the negative instances in the training data. It will help the proposed algorithm to better understand the negative instances reconstructing them with reduced error. In practical applications, most of the time, WSI is normal while only a small portion could be the positive cancerous region. Therefore, this limitation does not limit the applicability of the proposed algorithm.

6. Conclusion & future work

In this work, a fully unsupervised WSI classification algorithm is proposed using a Transformer Pseudo Label Generator (TPLG) and Transformer Label Cleaner (TLC). In TPLG, the instances are projected to a latent space and then inverse-projected to the original space using a projector and inverse projector. Based on the transformation error, the instances are assigned pseudo labels of being normal vs. cancerous. These pseudo labels are then cleaned using a label cleaning mechanism employed by TLC. Both components mutually learn from each other to obtain better labels iteratively. Based on the cleaned labels estimated by TLC, a discriminative learning mechanism is employed in the TPLG module so that the transformation error increases for the positive instances and decreases for the negative instances. Experiments are performed in fully unsupervised as well as weakly supervised settings for cancer vs. normal WSI classification on four different datasets. For downstream analysis, cancer subtype classification is performed using weak supervision for TLC fine-tuning. The proposed algorithm has demonstrated better performance compared to SOTA methods. As a future direction, investigating clinical tasks such as survival prediction using the proposed algorithm may be performed.

Table 7
Performance comparison of the proposed D-UMTL algorithm with self-supervised learning methods on two different datasets. The AUC is reported using the test split.

Methods	TCGA-LC	TCGA-RCC
H2T (Vu et al., 2023)	0.802	0.993
HIPT (Chen et al., 2022a)	0.952	0.980
SRCL (Wang et al., 2022)	0.973	0.991
Prop. D-UMTL	0.976	0.991

CRediT authorship contribution statement

Sajid Javed: Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Arif Mahmood:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Talha Qaiser:** Validation, Project administration, Investigation, Conceptualization. **Naoufel Werghi:** Resources. **Nasir Rajpoot:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This publication acknowledges the support provided by the Khalifa University of Science and Technology, United Arab Emirates under Faculty Start-Up grants FSU-2022-003 Award No. 8474000401.

References

Abbet, C., Zlobec, I., Bozorgtabar, B., Thiran, J.P., 2020. Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. In: MICAAI.
Alizadeh, A.A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., et al., 2015. Toward understanding and exploiting tumor heterogeneity. Nature Med. 21 (8), 846–853.
Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Med. 25 (6), 954–961.
Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv:1607.06450.
Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.
Bilal, M., Raza, S.E.A., Azam, A., Graham, S., Ilyas, M., Cree, I.A., Snead, D., Minhas, F., Rajpoot, N.M., 2021. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. LDH 3 (12), e763–e772.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Med. 25 (8), 1301–1309.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: ECCV.

Chapel Hill Kimryn Rathmell, W., et al., 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499 (7456), 43–49.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022a. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: IEEE CVPR.

Chen, C.L., Chen, C.C., Yu, W.H., Chen, S.H., Chang, Y.C., Hsu, T.I., Hsiao, M., Yeh, C.Y., Chen, C.Y., 2021a. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. NC 12 (1), 1–13.

Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F., 2020. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE TMI.

Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F., 2021. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: ICCV.

Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al., 2022b. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. CC 40 (8), 865–878.

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W., 2021b. Pre-trained image processing transformer. In: IEEE CVPR.

Chikontwe, P., Sung, H.J., Jeong, J., Kim, M., Go, H., Nam, S.J., Park, S.H., 2022. Weakly supervised segmentation on neural compressed histopathology with self-equivariant regularization. Med. Image Anal. 80, 102482.

Chuang, W.Y., Chen, C.C., Yu, W.H., Yeh, C.J., Chang, S.H., Ueng, S.H., Wang, T.H., Hsueh, C., Kuo, C.F., Yeh, C.Y., 2021. Identification of nodal micrometastasis in colorectal cancer using deep learning on annotation-free whole-slide images. MP 34 (10), 1901–1911.

Ciga, O., Xu, T., Martel, A.L., 2022. Self supervised contrastive learning for digital histopathology. MLA 7, 100198.

Cui, M., Zhang, D.Y., 2021. Artificial intelligence and computational pathology. LI 101 (4), 412–422.

Davis, C., Ricketts, C.J., Wang, M., Yang, L., Cherniack, A., Shen, H., Buhay, C., Kang, H., Kim, S., Fahey, C., et al., 2014. The somatic genomic landscape of chromophobe renal cell carcinoma. CC 26 (3), 319–330.

Di, D., Zou, C., Feng, Y., Zhou, H., Ji, R., Dai, Q., Gao, Y., 2022. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. IEEE TPAMI.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nature Med. 25 (1), 24–29.

Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I., Akinyelu, A.A., 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. EAAI 110, 104743.

Fitzgerald, R.C., Antoniou, A.C., Fruk, L., Rosenfeld, N., 2022. The future of early cancer detection. Nature Med. 28 (4), 666–677.

Fuchs, T.J., Buhmann, J.M., 2011. Computational pathology: challenges and promises for tissue analysis. CMIG 35 (7–8), 515–530.

Ge, L., Wei, X., Hao, Y., Luo, J., Xu, Y., 2022. Unsupervised histological image registration using structural feature guided convolutional neural network. IEEE TMI 41 (9), 2414–2431.

Guan, Y., Zhang, J., Tian, K., Yang, S., Dong, P., Xiang, J., Yang, W., Huang, J., Zhang, Y., Han, X., 2022. Node-aligned graph convolutional network for whole-slide image representation and classification. In: IEEE CVPR.

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: CVPR.

He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., Zhang, K., 2019. The practical implementation of artificial intelligence technologies in medicine. Nature Med. 25 (1), 30–36.

Hosseini, M.S., Brawley Hayes, J.A., Zhang, Y., Chan, L., Plataniotis, K.N., Damaskinos, S., 2019. Focus quality assessment of high-throughput whole slide imaging in digital pathology. IEEE TMI 39 (1), 62–74.

Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2016. Patch-based convolutional neural network for whole slide tissue image classification. In: IEEE CVPR.

Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., Wu, H., 2021. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In: MICCAI.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: ICML.

Ilya, S., et al., 2012. Comprehensive molecular portraits of human breast tumours. Nature 490 (7418), 61–70.

Jaume, G., Pati, P., Bozorgtabar, B., Foncubierta, A., Anniciello, A.M., Feroce, F., Rau, T., Thiran, J.P., Gabrani, M., Goksel, O., 2021. Quantifying explainers of graph neural networks in computational pathology. In: IEEE CVPR.

Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE TPAMI 43 (11), 4037–4058.

Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., Yamazaki, K., Takeo, S., Iizuka, O., Tsuneki, M., 2020. Weakly-supervised learning for lung carcinoma classification using deep learning. Sci. Rep. 10 (1), 1–11.

Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A., Bankhead, P., et al., 2020. Pan-cancer image-based detection of clinically actionable genetic alterations. NC 1 (8), 789–799.

Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al., 2019. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Med. 25 (7), 1054–1056.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2021. Transformers in vision: A survey. CSUR.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907.

Le-Khac, P.H., Healy, G., Smeaton, A.F., 2020. Contrastive representation learning: A framework and review. IEEE Access 8, 193907–193934.

Lerousseau, M., Vakalopoulou, M., Classe, M., Adam, J., Battistella, E., Carré, A., Estienne, T., Henry, T., Deutsch, E., Paragios, N., 2020. Weakly supervised multiple instance learning histopathological tumor segmentation. In: MICAAI.

Li, B., Li, Y., Eliceiri, K.W., 2021a. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: IEEE CVPR.

Li, J., Lin, T., Xu, Y., 2021b. Sslp: Spatial guided self-supervised learning on pathological images. In: MICCAI.

Li, W., Nguyen, V.D., Liao, H., Wilder, M., Cheng, K., Luo, J., 2019. Patch transformer for multi-tagging whole slide histopathology images. In: MICCAI.

Lipkova, J., Chen, T.Y., Lu, M.Y., Chen, R.J., Shady, M., Williams, M., Wang, J., Noor, Z., Mitchell, R.N., Turan, M., et al., 2022. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. Nature Med. 28 (3), 575–582.

Liu, Y., Jin, M., Pan, S., Zhou, C., Zheng, Y., Xia, F., Yu, P., 2022. Graph self-supervised learning: A survey. IEEE KDE.

Lu, M.Y., Chen, T.Y., Williamson, D.F., Zhao, M., Shady, M., Lipkova, J., Mahmood, F., 2021a. AI-based pathology predicts origins for cancers of unknown primary. Nature 594 (7861), 106–110.

Lu, W., Toss, M., Dawood, M., Rakha, E., Rajpoot, N., Minhas, F., 2022. SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer. MIA 102486.

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021b. Data-efficient and weakly supervised computational pathology on whole-slide images. NBE 5 (6), 555–570.

Marusyk, A., Polyak, K., 2010. Tumor heterogeneity: causes and consequences. BBA 1805 (1), 105–117.

Network, C.G.A.R., 2016. Comprehensive molecular characterization of papillary renal-cell carcinoma. NEJM 374 (2), 135–145.

Network, C.G.A.R., et al., 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511 (7511), 543.

Rawat, R.R., Ortega, I., Roy, P., Sha, F., Shibata, D., Ruderman, D., Agus, D.B., 2020. Deep learned tissue "fingerprints" classify breast cancers by ER/PR/Her2 status from H&E images. Scie. Rep. 10 (1), 1–13.

Rindi, G., Klimstra, D.S., Abedi Ardekani, B., Asa, S.L., Bosman, F.T., Brambilla, E., Busam, K.J., de Krijger, R.R., Dietel, M., El Naggar, A.K., et al., 2018. A common classification framework for neuroendocrine neoplasms: an international agency for research on cancer (IARC) and world health organization (WHO) expert consensus proposal. MP 31 (12), 1770–1786.

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H., 2022. Transformers in medical imaging: A survey. arXiv:2201.09873.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. NIPS 34.

Sharma, Y., Shrivastava, A., Ehsan, L., Moskaluk, C.A., Syed, S., Brown, D., 2021. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In: MIDL.

Srinidhi, C.L., Ciga, O., Martel, A.L., 2021. Deep neural network models for computational histopathology: A survey. MIA 67, 101813.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CJC 71 (3), 209–249.

Tellez, D., Litjens, G., van der Laak, J., Ciompi, F., 2019. Neural image compression for gigapixel histopathology image analysis. IEEE transactions on pattern analysis and machine intelligence 43 (2), 567–578.

Thandiackal, K., Chen, B., Pati, P., Jaume, G., Williamson, D.F., Gabrani, M., Goksel, O., 2022. Differentiable zooming for multiple instance learning on whole-slide images. In: ECCV2022.

Tizhoosh, H.R., Pantanowitz, L., 2018. Artificial intelligence and digital pathology: challenges and opportunities. JOPI 9 (1), 38.
Tomczak, K., Czerwińska, P., Wiznerowicz, M., 2015. Review the cancer genome atlas (TCGA): an immeasurable source of knowledge. CO 2015 (1), 68–77.
Turashvili, G., Brogi, E., 2017. Tumor heterogeneity in breast cancer. FIM 4, 227.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. NIPS 30.
Vu, Q.D., Rajpoot, K., Raza, S.E.A., Rajpoot, N., 2023. Handcrafted histological transformer (H2T): Unsupervised representation of whole slide images. MEDIA 102743.
Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W., 2018. Revisiting multiple instance neural networks. PR 74, 15–24.
Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. MEDIA 81, 102559.
Wu, H., Wang, Z., Song, Y., Yang, L., Qin, J., 2022. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In: IEEE CVPR.
Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W., 2019. Camel: A weakly supervised learning framework for histopathology image segmentation. In: ICCV.
Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y., 2022a. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: IEEE CVPR.
Zhang, J., Zhang, X., Ma, K., Gupta, R., Saltz, J., Vakalopoulou, M., Samaras, D., 2022b. Gigapixel whole-slide images classification using locally supervised learning. In: MICCAI.
Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B., 2022. A graph-transformer for whole slide image classification. IEEE TMI 41 (11), 3003–3015.
Zhu, W., Lou, Q., Vang, Y.S., Xie, X., 2017. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: MICCAI.