# Spatiotemporal Deformable Transformer for Lung Cancer Risk Prediction Using Low-Dose CT Images

Ya-Han Chang, Yi-Hui Lin, Hsing-Chen Meng, Shawn Wu, Yen-Yu Lin, *Senior Member, IEEE*, and Chun-Rong Huang, *Senior Member, IEEE*

***Abstract*—Low-dose lung computed tomography (LDCT) has been set as the golden standard for lung cancer screening. Current state-of-the-art computer-aided diagnosis (CAD) methods aim to detect visible lung cancer from LDCT images based on well-labeled training data. To provide early alarms, we aim to predict lung cancer risk and locate the potential lung cancer region by utilizing LDCT images in two consecutive years, i.e. LDCT images of the patient captured in the first year and the second year without pixel-level labels. To achieve this goal, a novel weakly supervised spatiotemporal deformable transformer is proposed. Because LDCT images are captured at different time points without pre-alignment, our method proposes combining spatial attention maps with temporal deformable attention maps to learn spatiotemporal correlations which represent the tumor progression from LDCT images in consecutive years. To achieve weakly supervised learning of the model, we introduce three losses, including classification loss, patch token loss, and side loss. During inference, spatial attention maps with class information cooperate with spatiotemporal correlations of the test image to locate the potential lung cancer region. In the experiments, the proposed method outperforms the state-of-the-art methods and can localize the lung cancer region with only image-level labels.**

***Index Terms*—Lung cancer risk prediction, transformer, deformable attention.**

## I. Introduction

ACCORDING to the statistics from the World Health Organization (WHO) [1], lung cancer is the second most common cancer in the world with approximately $2.21$ million new cases each year. It is also the leading cause of cancer-related deaths, and around $1.8$ million deaths due to the disease annually. In Taiwan [2], lung cancer is also the leading cause of cancer-related deaths. There are approximately $16,000$ new cases and around $9,500$ patients die from lung cancer each year. At the time of initial diagnosis, the majority of lung cancer cases are in stage IV, accounting for $46\%$ of all stages [2]. If lung cancer can be diagnosed early, the mortality rate can be reduced with significant socioeconomic benefits.

The National Lung Screening Trial (NLST) [3] has demonstrated that using low dose lung computed tomography (LDCT) as a lung cancer screening tool can improve the survival rate of lung cancer patients. Based on [3], LDCT has become the gold standard tool for the detection of lung cancer. Benefited from the widespread prevalence of LDCT, $34\%$ of lung cancer can be diagnosed at stage I [2]. In the diagnostic criteria of Lung-RADS v2022 [4], lung nodules are categorized into seven levels according to their likelihood of malignancy: (1) negative, (2) benign, (3) probably benign, (4A) suspicious, (4B) very suspicious, (4X) category 3 or 4 with additional findings that increase suspicion, and (S) significant. For the levels (2) benign, (3) probably benign, and (4A) suspicious, it is recommended to repeat LDCT after a certain period of time to observe the likelihood of cancer due to insufficient evidence. Using LDCT images in two consecutive years, the accuracy of the diagnosis of lung cancer can be increased.

Recently, many deep learning methods have been applied for medical image analysis [5]–[11]. They are also applied to lung cancer classification [12]–[17]. Due to the limited information provided by the lung cancer classification methods, lung cancer detection methods [18]–[21] are proposed to locate the positions of the tumors. To obtain more accurate cancer regions, numerous lung cancer segmentation methods [22]–[26] are proposed. To achieve tumor detection or segmentation, these supervised methods require box- or pixel-level labels, which are a burden for radiologists.

The aforementioned lung cancer diagnosis methods classify, detect, or segment cancer lesions from CT images that contain these lesions. In preventive medicine, physicians aim to alarm patients before lung cancer occurs. Thus, given CT images of a patient, lung cancer prediction methods aim to predict if the patient will be diagnosed with cancer or not in the future. In the medical domain, lung cancer prediction is important to assess the cancer risk of patients. The timing of the next LDCT scan is typically based on factors such as the size (diameter, volume, or density) of the largest lung nodule in the previous LDCT or the presence of new nodules.

However, the diagnosis results can be variant based on the

Manuscript received Jul. 15, 2024, and revised Jan. 23, 2025 and Nov. 12, 2025. This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC 114-2221-E-A49-034-MY3, NSTC 114-2640-E-006-010, NSTC 114-2221-E-A49-038-MY3, NSTC 113-2634-F-006-002, and NSTC 112-2221-E-A49-090-MY3. (Corresponding authors: Chun-Rong Huang and Yen-Yu Lin).

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Y.-H. Chang is with the Institute of Multimedia Engineering, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan (e-mail: yhchang.cs10@nycu.edu.tw).

Y.-H. Lin is with the Department of Radiotherapy and Oncology, Pintung Veterans General Hospital, Pintung 900, Taiwan (e-mail: irene2023study@gmail.com).

H.-C. Meng, is with the Graduate Degree Program of Artificial Intelligence, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan. (e-mail: idfidfidf89520.ee11@nycu.edu.tw).

S. Wu is with the Department of Diagnostic Imaging, Medical Research Institute of Health, Middletown, NY 10940, USA. (e-mail: shengywu58@gmail.com).

Y.-Y. Lin and C.-R. Huang are with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan. (e-mail: lin@cs.nycu.edu.tw, crhuang@cs.nycu.edu.tw).

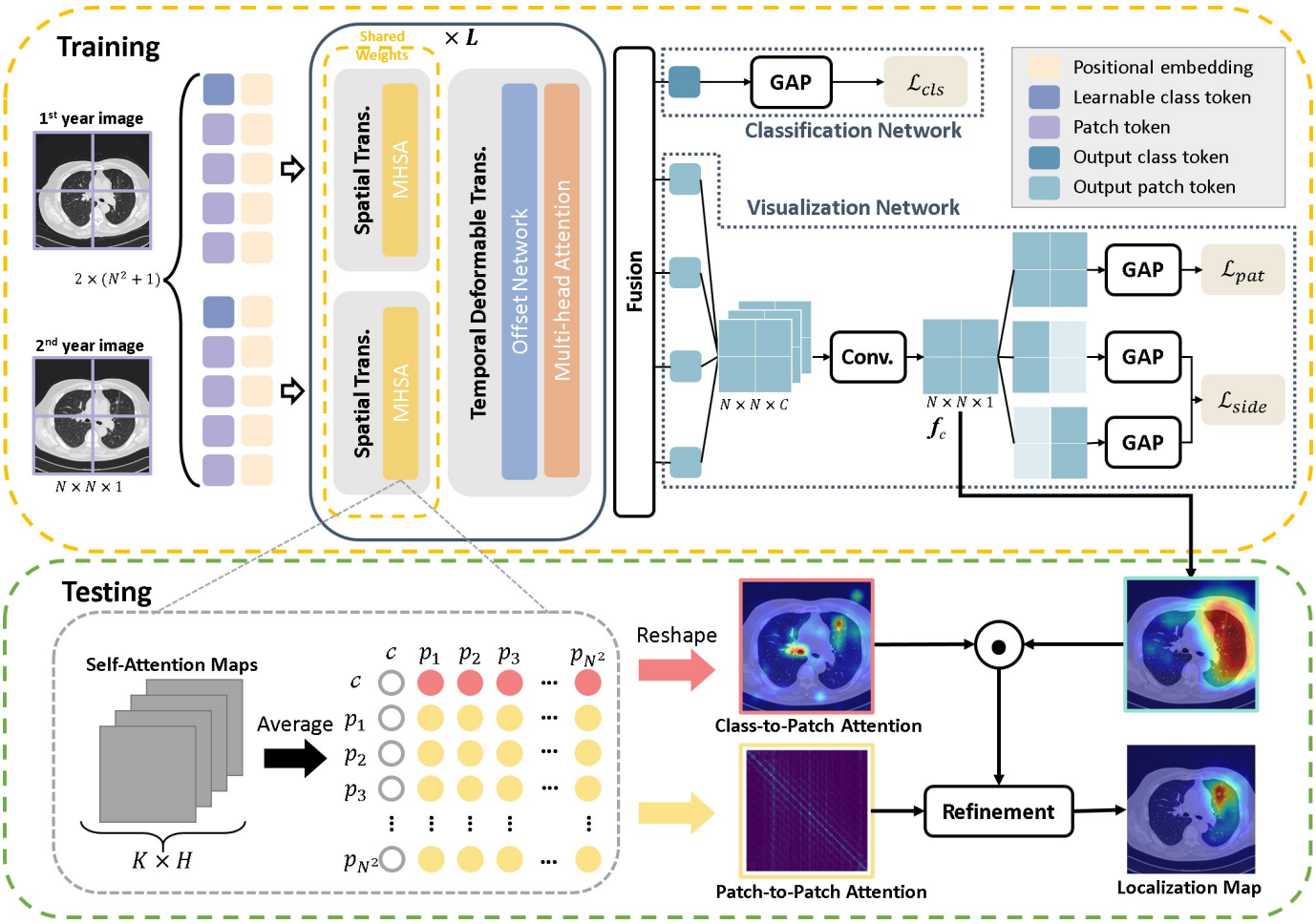



Fig. 1: The architecture of the proposed model. GAP represents the global average pooling layer. During testing, $c$, $p_1$, ..., and $p_{N^2}$ represent the class token and patch tokens of the global pairwise attention map $\boldsymbol{A}_g$.

experience of different radiologists. To address the issue, several deep learning-based lung cancer prediction methods [27]–[30] are presented. For example, Huang et al. [27] propose the DeepLR model considering clinical information and CT reading by the radiologist for lung cancer risk prediction. However, the use of radiomic features that are interpreted by experienced radiologists brings time-consuming burdens to radiologists. Lu et al. [28] propose a convolutional neural network, CXR-LC, to identify smokers at high risk of lung cancer based on a single chest radiograph image and basic information from electronic medical records (EMR). Using a tubular network and an image network to extract features from EMR information and chest radiograph image, respectively, CXR-LC predicts the incidence of lung cancer based on these features.

Ardila et al. [29] propose an end-to-end lung cancer risk prediction model that uses CT images alone. The proposed model is composed of four components, including a lung segmentation model, a cancer ROI detection model, a full-volume model, and a cancer risk prediction model. Although their model achieves good performance for cancer risk prediction and localization, the complicated architecture and object-level supervision become the shortcomings of this method. Mikhael et al. [30] propose Sybil to predict the future risk of lung cancer based on a single CT volume. Sybil not only infers biologically relevant information from the input CT images but also outputs the risk for the next 6 years. However, Sybil also requires object-level annotations that indicate the locations of suspicious nodules to train the designed guided attention module. In summary, [27] and [28] require additional clinical data or imaging features interpreted by the radiologists. In [29] and [30], object-level labels are employed to train the risk prediction models. Both manual interpretation of medical images and pixel-level labels impose a heavy burden on radiologists. Moreover, these methods are hard to provide explainable results, which lead to the black-box nature of the models.

To deal with these problems, we propose a novel spatiotemporal deformable transformer for lung cancer risk prediction and weakly supervised localization from LDCT images in two consecutive years. The proposed method predicts whether the patient will be diagnosed with cancer in the third year and the potential locations of tumors with only image-level labels. Fig. 1 shows the network architecture composed of a spatiotemporal deformable transformer encoder module, a classification network, and a visualization network. The trans-

former encoder module is proposed to learn representative features from LDCT images in two consecutive years. It contains cascaded spatiotemporal deformable transformer encoders, which consist of two shared-weights spatial transformer encoder and a temporal deformable transformer encoder. The spatial transformer encoder utilizes the standard self-attention mechanism to learn intra-image information. The temporal deformable transformer encoder is designed to capture inter-image information and cope with the alignment issue within two LDCT images due to the movement and positions of the patient captured at different time points and tumor progression.

The resultant features of the last spatiotemporal deformable transformer encoder are fed into the classification network and the visualization network. The front one is used for lung cancer risk prediction, and the latter one is used for the tumor localization in a weakly supervised manner. Three losses are used to guide the learning of the proposed model. The first one is the classification loss, which aims to align the output of the model with the image-level labels. The second loss is the patch token loss, which is used to obtain the locations of the tumors by considering the salient feature responses. If the reshaped patch tokens contain a tumor, salient feature responses should be learned. The third loss is the side loss, which helps the proposed model to learn more accurate tumor locations. The main contribution in this paper is as follows.

1) Our method is the first weakly supervised deep learning method for lung cancer risk prediction and localization from LDCT images in two consecutive years.
2) We propose the spatiotemporal deformable transformer to deal with the alignment problem and represent the tumor progression between the LDCT images of different years.
3) We integrate anatomical priors as weak labels into the proposed model and introduce the patch loss and side loss to obtain tumor localization maps. As a result, the proposed method outperforms state-of-the-art methods in both prediction and localization accuracy.

The remaining parts of the paper are organized as follows. In Sec. II, we present the materials and the proposed method. The results are shown in Sec. III. Finally, we give the conclusions in Sec. IV.

## II. Materials and Methods

### A. Patient Data

In the experimental results, the National Lung Screen Trial (NLST) [3] dataset is applied. The dataset was collected from 2002 to 2004 from 33 medical centers throughout the United States. These patients were randomly assigned to the CXR annual screening group or the LDCT group. Among the LDCT group, $142$ patients who were diagnosed with lung cancer in the third year are screened by considering if the patients performed the LDCT procedure in three consecutive years. In addition, $192$ normal patients who performed the LDCT procedure in three consecutive years were randomly selected from the NLST dataset as the negative data to avoid the problem of data imbalance. The training data and the testing data are partitioned with $70\%$ and $30\%$ in the experiments.

For each patient, we first align the 2-D LDCT images of the first year, the second year, and the third year in the same position on the $z$-axis based on the alignment of the carina and the thickness of the 3-D LDCT images of each year. Using the alignment process, the 2-D LDCT images aligned with the $z$-axis are extracted for the following model processing. In this way, the model can focus on discovering the temporal changes between the LDCT images of the first year and the second year. If the LDCT image in the third year is diagnosed with lung cancer, the label of the corresponding LDCT images captured in the first year and the second year is annotated with the diagnosis of lung cancer. The corresponding LDCT images serve as the input of the model, and the results of the lung cancer diagnosis in the third year serve as the weak label of the model. Our goal is to train a model using LDCT images of two consecutive years to predict lung cancer in the third year and to localize the tumor in a weakly supervised manner. If the model predicts the LDCT images of two consecutive years that contain lung cancer, the patient is predicted to have lung cancer in the third year at the case level.

### B. Overview

Let $I_1$ and $I_2$ be consecutive LDCT images captured in the first year and the second year of a patient, respectively. The image-level label $y$ of the patient is defined as $y \in \{0, 1\}$, where $0$ indicates that the patient is diagnosed without lung cancer and $1$ indicates that the patient is diagnosed with lung cancer, based on the image captured in the third year. The side label $z$ of the patient is defined as $z \in \{l, r\}$, where $l$ or $r$ indicate that the patient is diagnosed with lung cancer in the left lung or the right lung in the third year. We aim to develop a weakly supervised model to achieve lung cancer prediction and localize the potential cancer regions.

As shown in Fig. 1, the proposed model is composed of a spatiotemporal deformable transformer encoder module, a classification network, and a visualization network. Given $I_1$ and $I_2$, they are individually projected to multiple non-overlapping patch tokens. The patch tokens of each LDCT image are concatenated with a class token and combined with positional embeddings to generate the input tokens of the spatiotemporal deformable transformer encoder module. The module contains cascaded spatiotemporal deformable transformer encoders. Each encoder consists of two shared-weights spatial transformer encoders and a temporal deformable transformer encoder. The spatial transformer encoders aim to learn the spatial intra-image information of patch tokens from $I_1$ and $I_2$, respectively. To further discover the temporal inter-image correlations of spatial features of $I_1$ and $I_2$, the temporal deformable transformer encoder is proposed.

The output class token $\boldsymbol{T}_{cls}$ and the output patch token $\boldsymbol{T}_{pat}$ of the last spatiotemporal transformer encoder serve for different purposes. $\boldsymbol{T}_{cls}$ serves as the input to the classification network to predict if the patient has lung cancer in the third year and compute the classification loss $\mathcal{L}_{cls}$. $\boldsymbol{T}_{pat}$ serves as the input to the visualization network to generate an attention map to localize potential cancer regions based on weak labels $y$ and $z$. The learned patch tokens of $I_1$ and $I_2$ are concatenated

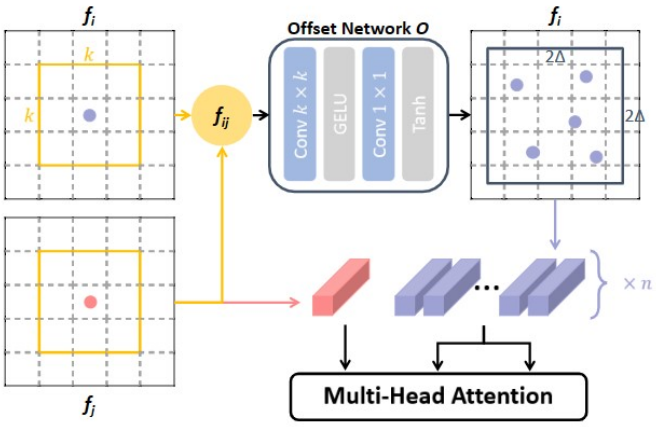



Fig. 2: Temporal Deformable Transformer Encoder.

and reshaped to 2-D feature $\boldsymbol{f}_c$ to compute the patch token loss $\mathcal{L}_{pat}$ and the side loss $\mathcal{L}_{side}$. $\mathcal{L}_{pat}$ is designed to make the learned features of the generated attention map consistent with the weak label, while $\mathcal{L}_{side}$ is designed to aid the model in locating the cancer lesion in the left lung or right lung. The total loss $\mathcal{L}$ is defined as

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{pat} + \mathcal{L}_{side}. \tag{1}$$

During inference, the learned patch tokens are cooperated with $\boldsymbol{f}_c$ to generate the tumor localization map.

### C. Spatiotemporal Deformable Transformer Encoder

As shown in Fig. 1, each LDCT image is divided into $N^2$ patches, and then each patch is projected onto a 1-D patch token using a linear layer with $D = 384$ neurons. A trainable class token is concatenated to the patch tokens and the positional embedding is appended to reserve the positional information of each patch [31]. Finally, the input tokens $\boldsymbol{T}_1 \in \mathbb{R}^{(1+N^2)\times D}$ and $\boldsymbol{T}_2 \in \mathbb{R}^{(1+N^2)\times D}$ of $I_1$ and $I_2$ are fed into the spatiotemporal deformable transformer encoder module.

The transformer module contains $L(= 12)$ spatiotemporal deformable transformer encoders. Each encoder consists of two shared-weights spatial transformer encoders $e_1^s$ and $e_2^s$, and a temporal deformable transformer encoder $e^t$. The spatial transformer encoder consists of a multi-head attention (MHA) module and a multilayer perceptron (MLP). A normalization layer is applied before the MHA module and the MLP, respectively. The spatial transformer modules $e_1^s$ and $e_2^s$ aim to capture intra-image information from $I_1$ and $I_2$, respectively.

A naive idea to discover the temporal inter-image correlations between features of $I_1$ and $I_2$ is to apply the multi-head self-attention scheme to patches of corresponding spatial positions across different time points which is commonly applied in video prediction methods [32], [33]. Because two LDCT images are captured at different time points, they are not aligned. Thus, the naive approach will lead to misalignment between the patches of $I_1$ and $I_2$. Correctly discovering the temporal correlations of the features of the patches of $I_1$ and $I_2$ under the unaligned condition remains a problem.

To cope with the alignment problem and learn the temporal inter-image correlations between $I_1$ and $I_2$, the temporal deformable transformer encoder is proposed. Because our objective is to capture the temporal correlations between the content of $I_1$ and $I_2$, we only consider patch tokens. Let $\hat{\boldsymbol{T}}_1 \in \mathbb{R}^{N^2\times D}$ and $\hat{\boldsymbol{T}}_2 \in \mathbb{R}^{N^2\times D}$ represent the patch tokens generated by $e_1^s$ and $e_2^s$, respectively. Then, we reshape $\hat{\boldsymbol{T}}_1$ and $\hat{\boldsymbol{T}}_2$ into 2-D feature maps $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, where the dimensions of $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are $\mathbb{R}^{N\times N\times D}$.

To make the model produce appropriate deformation points to align patch tokens between two LDCT images, we propose the temporal deformable transformer encoder as shown in Fig. 2. In the encoder, the differences between $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ serve as the input of the offset network $O$ to generate the deformable offset, which provides the spatial positions of the corresponding patches at different time points. Based on the deformable offset, the encoder computes the multi-head attention based on different queries to obtain key temporal features to represent the changes between two images, i.e., to learn the tumor progression between two images.

Let the features of $I_1$ be the basis and the features of $I_2$ be the query. We compute the fused feature map $\boldsymbol{f}_{12}$ with respect to $I_1$ by considering the differences between $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ as follows:

$$\boldsymbol{f}_{12} = sigmoid(\boldsymbol{f}_1 - \boldsymbol{f}_2) \times \boldsymbol{f}_1, \tag{2}$$

where $sigmoid(\cdot)$ is the sigmoid function. Then $\boldsymbol{f}_{12}$ is entered into the offset network. The offset network is composed of a $5\times 5$ convolutional layer, a GELU layer, a $1\times 1$ convolutional layer, and a Tanh layer. The offset $\boldsymbol{\Delta}_{12}$ is generated by feeding $\boldsymbol{f}_{12}$ into a offset network $O$ as follows:

$$\boldsymbol{\Delta}_{12} = O(\boldsymbol{f}_{12}). \tag{3}$$

Given the centers $\boldsymbol{c}_2$ of the patches in $I_2$, the candidate centers $\boldsymbol{c}_1$ of the corresponding patches in $I_1$ are computed as $\boldsymbol{c}_1 = \boldsymbol{c}_2 + \boldsymbol{\Delta}_{12}$. In our method, the number $S$ of candidate centers is 11. We then sample the features at $\boldsymbol{c}_1$ through bilinear interpolation. These generated deformable features form a set $\mathcal{X}_1 = \{\boldsymbol{x}_1^1, ..., \boldsymbol{x}_1^s, ..., \boldsymbol{x}_1^S\}$, where the dimension of each feature in $\mathcal{X}_1$ is $\mathbb{R}^{N\times N\times D}$. Let the feature of the corresponding patch of $I_2$ with respect to $\mathcal{X}_1$ be $\boldsymbol{x}_2^q$, which serves as the query patch for the multi-head attention computation. The deformable multi-head attention (DA) based on the query $\boldsymbol{x}_2^q$ in $I_2$ and the feature $\boldsymbol{x}_1^s$ in $\mathcal{X}_1$ is computed as follows:

$$DA(\boldsymbol{x}_2^q, \boldsymbol{x}_1^s) = softmax(\frac{L^q(\boldsymbol{x}_2^q)L^k(\boldsymbol{x}_1^s)^\intercal}{\sqrt{D}})L^v(\boldsymbol{x}_1^s), \tag{4}$$

where $L^q(\cdot)$, $L^k(\cdot)$ and $L^v(\cdot)$ are linear projection functions to generate query, key, and value to compute deformable multi-head attention. Then, the temporal deformable attention feature $\tilde{\boldsymbol{T}}_1$ of $I_1$ is obtained by concatenating all deformable multi-head attention features.

Similarly, we can consider the features of $I_2$ as the basis and the features of $I_1$ as the query. Then, we compute the fused feature map $\boldsymbol{f}_{21}$ with respect to $I_2$ by considering the differences between $\boldsymbol{f}_2$ and $\boldsymbol{f}_1$ as follows:

$$\boldsymbol{f}_{21} = sigmoid(\boldsymbol{f}_2 - \boldsymbol{f}_1) \times \boldsymbol{f}_2. \tag{5}$$

TABLE I: Comparisons with Competing Methods on the NLST dataset.

| Methods | Accuracy | Precision | Recall | F1-score | mIoU |
|---|---|---|---|---|---|
| MCTformer [34] | 0.8193 | 0.8125 | 0.7429 | 0.7761 | 0.5167 |
| TimeSformer [32] | 0.7711 | 0.8077 | 0.6000 | 0.6885 | 0.5002 |
| MViT [33] | 0.7229 | 0.6429 | 0.7714 | 0.7013 | 0.4997 |
| SIFA [35] | 0.7470 | 0.6750 | 0.7714 | 0.7200 | 0.4996 |
| Uniformer [36] | 0.7590 | 0.6667 | 0.8571 | 0.7500 | 0.4993 |
| Proposed | **0.9398** | **0.9412** | **0.9143** | **0.9276** | **0.5345** |

$\boldsymbol{f}_{21}$ is then served as the input of the offset network to obtain the deformable feature set $\mathcal{X}_2 = \{\boldsymbol{x}_2^1, ..., \boldsymbol{x}_2^s, ..., \boldsymbol{x}_2^S\}$ using $\boldsymbol{\Delta}_{21}$. Then, the deformable multi-head attention based on the query $\boldsymbol{x}_1^q$, which is the corresponding patch of $I_1$ with respect to $\mathcal{X}_2$, and the feature $\boldsymbol{x}_2^s$ in $\mathcal{X}_2$ is computed as follows:

$$DA(\boldsymbol{x}_1^q, \boldsymbol{x}_2^s) = softmax(\frac{L^q(\boldsymbol{x}_1^q)L^k(\boldsymbol{x}_2^s)^\intercal}{\sqrt{d}})L^v(\boldsymbol{x}_2^s). \quad (6)$$

Then, by concatenating all deformable multi-head attention features, the temporal deformable attention feature $\tilde{\boldsymbol{T}}_2$ of $I_2$ is obtained. With the proposed temporal deformable transformer encoder, the model will not be limited to only calculating the attention of tokens at the same location of different time points. It can find suitable tokens in a wider range to calculate deformable multi-head attention and represent the tumor progression. Moreover, by considering deformable attention from two different bases, the offset network can better learn the temporal inter-image correlations and solve the alignment problem.

### *D. Classification Network and Visualization Network*

To achieve cancer prediction and weakly supervised localization, we apply $\tilde{\boldsymbol{T}}_1$ and $\tilde{\boldsymbol{T}}_2$. The output class token $\boldsymbol{T}_{cls}$ of the last spatiotemporal transformer encoder serves as the input of the classification network. $\boldsymbol{T}_{cls}$ is passed to a global average pooling layer to extract the key features related to tumors for the classification loss $\mathcal{L}_{cls}$ computed based on the image-level labels as follows:

$$\mathcal{L}_{cls} = -(y \times \log(G(\boldsymbol{T}_{cls})) + (1-y) \times \log(1-G(\boldsymbol{T}_{cls}))), \quad (7)$$

where $y$ is the image-level label that indicates whether the patient is diagnosed with the lung cancer in the third year, and $G(\cdot)$ is the global average pooling function for the output class token.

Because $\tilde{\boldsymbol{T}}_1$ and $\tilde{\boldsymbol{T}}_2$ contain deformable temporal differences, we compute the difference map $\tilde{\boldsymbol{T}}_2 - \tilde{\boldsymbol{T}}_1$ and combine the difference map with $\tilde{\boldsymbol{T}}_2$ to generate the output patch token $\boldsymbol{T}_{pat}$ which contains $N^2$ tokens. This integration is informed by the prior knowledge suggesting that lesions with significant changes could develop into malignant tumors by considering the difference between $I_1$ and $I_2$, i.e., the tumor progression between $I_1$ and $I_2$.

The visualization network is designed to enable the model to generate precise tumor localization maps based on weak labels. Specifically, the output patch token $\boldsymbol{T}_{pat}$ is reshaped into $N \times N \times C$ and fed into a $3\times3$ convolutional layer with 1 channel to obtain the 2-D feature map $\boldsymbol{f}_c$. When a patient has lung cancer in the third year, $\boldsymbol{f}_c$ should contain the feature responses of the tumors correlated with the image label. Otherwise, $\boldsymbol{f}_c$ should not have feature responses of tumors. To address this issue, we propose the patch token loss $\mathcal{L}_{pat}$ as follows:

$$\mathcal{L}_{pat} = -(y \times \log(G(\boldsymbol{f}_c)) + (1-y) \times \log(1-G(\boldsymbol{f}_c))). \quad (8)$$

In this way, we can enforce the network to represent the feature responses of tumors using $\boldsymbol{f}_c$.

While the patch token loss helps learn the existence of tumors, the side loss $\mathcal{L}_{side}$ helps learn the potential locations of tumors based on weak labels $z$. Because $\boldsymbol{f}_c$ is a 2-D feature map to represent the correlations of the tokens in $I_1$ and $I_2$, we split $\boldsymbol{f}_c$ into $\boldsymbol{f}_c^l$ and $\boldsymbol{f}_c^r$ which are the left side and right side of $\boldsymbol{f}_c$ to represent the left and right lung regions of $\boldsymbol{f}_c$ as shown in Fig. 1. If the tumor is located in the left lung, i.e., $z \in l$, $\lambda$ is set to 1. Otherwise, when $z \in r$, $\lambda$ is set to 0. The side loss $\mathcal{L}_{side}$ is then defined as follows:

$$\mathcal{L}_{side} = -y(\lambda \times \log(G(\boldsymbol{f}_c^l)) + (1-\lambda) \times \log(G(\boldsymbol{f}_c^r))). \quad (9)$$

Please note that the side loss of a normal patient is 0 to indicate that there are no feature responses of tumors for both sides of the lung. By utilizing the labels that indicate the potential location of the tumor, the side loss enables the proposed model to generate more accurate visualization maps.

### *E. Inference of Localization Map*

During inference, the class-to-patch attention map, patch-to-patch attention map [34] and $\boldsymbol{f}_c$ are utilized to generate the localization maps of predicted tumors. To identify suspicious regions of the LDCT image captured in the second year, $K \times H$ self-attention maps are extracted from $H(=6)$ heads of the last $K(=3)$ spatial transformer encoders which are used to learn intra-image information from $I_2$. We average the self-attention maps to obtain the global pairwise attention map $\boldsymbol{A}_g$ [34], which consists of the correlations of class and patch tokens. The class-to-patch attention map $\boldsymbol{A}_c = \boldsymbol{A}_g[1:2, 2:1+N^2]$ is represented by the red dots shown in Fig. 1, which indicates the correlations of the class token $c$ and patch tokens $p_1$, ..., $p_{N^2}$. Here, $\boldsymbol{A}_c$ signifies the responses of the patch tokens with respect to the class token. Then, $\boldsymbol{A}_c$ is reshaped to $\mathbb{R}^{N\times N\times 1}$ to provide the weights of the patches for $\boldsymbol{f}_c$ to indicate potential tumor locations.

In addition to the class-to-patch attention map, the patch-to-patch attention map $\boldsymbol{A}_p$ is used to refine the tumor localization map. $\boldsymbol{A}_p$ is defined as $\boldsymbol{A}_p = \boldsymbol{A}_g[2:1+N^2, 2:1+N^2]$, which are indicated as the yellow dots shown in Fig. 1. $\boldsymbol{A}_p$ indicates the correlations of different patch tokens by considering the

affinity among the patch tokens and the dimension of $\boldsymbol{A}_p$ is $\mathbb{R}^{N^2 \times N^2}$. With $\boldsymbol{A}_c$, $\boldsymbol{A}_p$, and $\boldsymbol{f}_c$, the localization map $\boldsymbol{A}_f$ of the tumor is obtained as follows:

$$\boldsymbol{A}_f = \boldsymbol{A}_p \cdot Re(\boldsymbol{f}_c \odot \boldsymbol{A}_c), \tag{10}$$

where $\cdot$ represents matrix multiplication, $\odot$ represents element-wise multiplication, and the reshape function $Re(\cdot)$ changes a $N \times N \times 1$ feature map to a $N^2 \times 1$ feature map. Because the dimension of $\boldsymbol{A}_f$ is $N^2 \times 1$, we reshape $\boldsymbol{A}_f$ to $N \times N \times 1$ to obtain the final tumor localization map during inference.

### F. Implementation Details

The proposed method is implemented in PyTorch 1.10.0 with a GeForce GTX 1080 Ti GPU. We initialize the model by using the DeiT-S [37] pre-trained on ImageNet [38] as the backbone network. We randomly resize the training images and then crop them in $224 \times 224$. We use AdamW [39] as an optimizer to train the proposed model. The batch size and weight decay are set to 24 and 0.05, respectively. The warm-up strategy is employed to train the proposed model. The initial learning rate is $1e-6$, and increases linearly based on the training strategy in [34], [37] during the first 5 epochs. Then, the decay of the cosine learning rate is used until 200 epochs.

## III. Experimental Results

We employ accuracy, precision, recall, and F1-score to assess the cancer prediction performance of the proposed method and competing methods. Additionally, to evaluate weakly supervised localization performance, we utilize the mean intersection over union (mIoU) metric. Given the localization map $\boldsymbol{A}_f$, we apply a threshold 0.93 to generate the tumor segmentation map. The tumor segmentation map is then compared with the ground truth region labeled by the doctors to compute the mIoU as follows:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}, \tag{11}$$

where $k = 1$, $p_{ii}$ represents the number of true positive pixels, and $p_{ij}$ and $p_{ji}$ represent the numbers of false positive pixels and false negative pixels, respectively.

### A. Quantitative Results

To the best of our knowledge, the proposed method is the first weakly supervised method for lung cancer prediction and localization from LDCT images captured in two consecutive years. Therefore, the proposed method was compared with similar state-of-the-art transformer-based methods in the computer vision domain. First, a weakly supervised approach, MCTformer [34] served as a baseline. We only input the LDCT image of the second year into the model. Second, several transformer-based video recognition methods including TimeSformer [32] MViT [33], SIFA [35], and Uniformer [36] were compared for lung cancer prediction from LDCT images in two consecutive years. Furthermore, Grad-CAM [40] was applied to produce heat maps for these methods to locate the predicted cancer regions.

Table I shows the lung cancer prediction and localization results of the proposed method and the state-of-the-art methods. Because MCTformer [34] only considers the LDCT image captured in the second year without temporal information, its recall values are not satisfactory. Such results imply the importance of imposing temporal information for cancer prediction. Because TimeSformer [32] and MViT [33] take into account information from patches of the same location at different time points, they achieve worse results for lung cancer prediction when the LDCT images are not aligned.

SIFA [35] considers stand-alone inter-fame attention, which re-scales the offset by using the difference between two images. Nevertheless, such a scheme is still hard to capture the subject variations in different years. By considering hierarchical architectures of CNN and vision transformer, Uniformer [36] achieves a better recall value compared with the above-mentioned methods. Because of subject variations and the nonalignment of images, we propose the spatiotemporal deformable transformer encoder to capture temporal correlations in consecutive LDCT images. The proposed temporal deformable attention mechanism considers using the offset network to learn the corresponding patches for cross attention computation of two images. Thus, the proposed method significantly outperforms the competing methods.

For the weakly supervised lung cancer localization, the results of the video recognition methods struggle to accurately identify the locations of tumors and result in low mIoU. In contrast, the weakly supervised method, MCTformer, produces more effective attention maps by learning multiple class tokens compared with video recognition methods. Because of the absence of temporal information, the resulting heat maps of MCTformer are still not precise. In contrast, the proposed method applies the visualization network with the patch token loss and side loss to locate the tumors, and thus achieves better mIoU in a weakly supervised manner compared with competing methods.

### B. Qualitative Results

Fig. 3 shows the visualization results of the proposed method and competing methods for patients diagnosed with lung cancer in the third year. Images with red borders indicate that the model gives incorrect predictions. Fig. 3(a) and (b) show the LDCT images captured in the first year and the second year, where these images are not aligned. Fig. 3(c), (d), (e), and (f) show the tumor localization maps of MCTformer, SIFA, Uniformer, and the proposed method, respectively. The ground truth locations of the tumors of the third year are shown in Fig. 3(g).

The first two rows shown in Fig. 3 show patients who are correctly diagnosed with tumors in the third year. Compared with SIFA and Uniformer, MCTformer achieves better visualization results because it considers patch similarity to discover salient regions that correspond to the class labels. However, the results of MCTformer are more diverse and contain false alarms compared to the ground truth shown in Fig. 3(g). Although SIFA and Uniformer achieve better recall results shown in Table I, their localization maps do not correlate

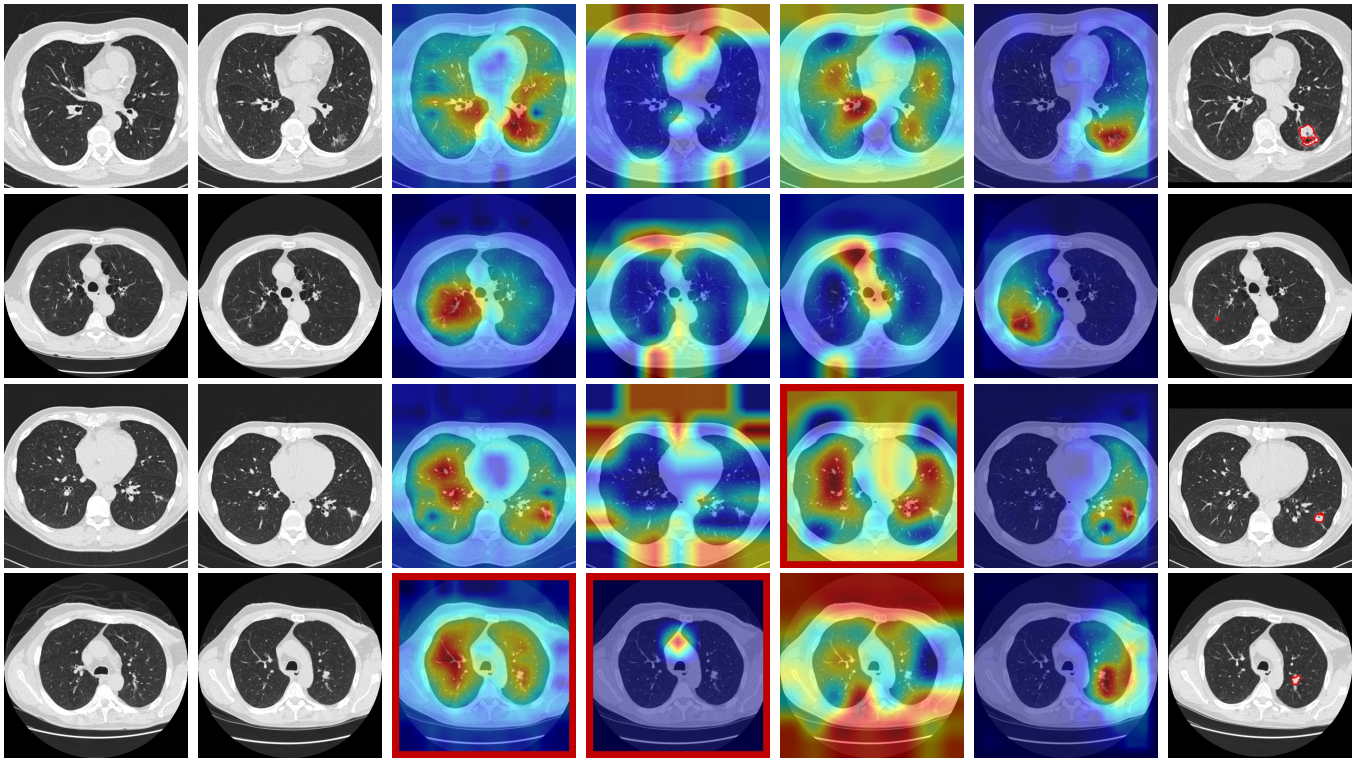



Fig. 3: Lung tumor localization results for positive cases. (a) The LDCT image captured in the first year, (b) The LDCT image captured in the second year, (c) Results of MCTformer, (d) Results of SIFA, (e) Results of Uniformer, (f) Results of the proposed method, and (g) the ground truth on the LDCT images captured in the third year. The red borders indicate that the model gives wrong predictions.

TABLE II: Ablation Study of Different Attention Mechanisms on the NLST dataset.

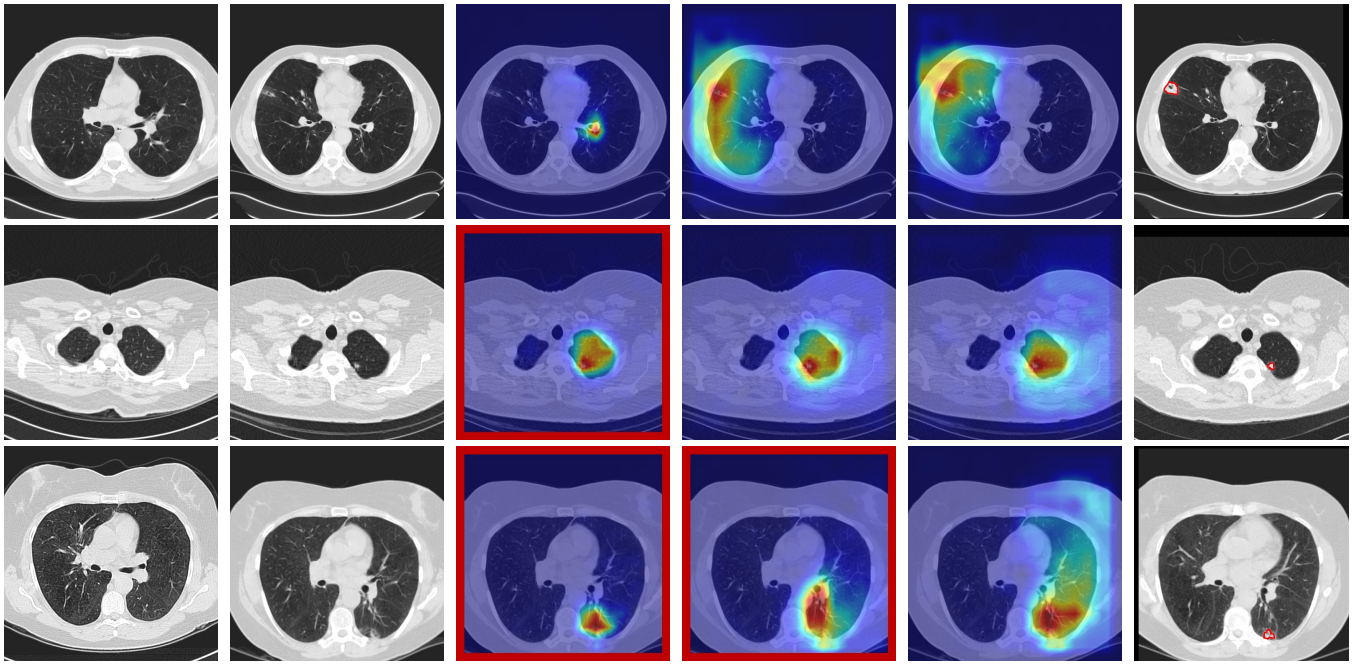
| Input | Attention | Accuracy | Precision | Recall | F1-score | mIoU |
|---|---|---|---|---|---|---|
| $1^{st}$ & $2^{nd}$ | Spatial Attention [41] | 0.8438 | 0.8667 | 0.7429 | 0.8000 | 0.5196 |
| | Spatial & Temporal Attention [32] | 0.8675 | 0.8333 | 0.8571 | 0.8450 | 0.5292 |
| | Spatial & Temporal Deformable Attention | **0.9398** | **0.9412** | **0.9143** | **0.9276** | **0.5345** |

with the ground truth shown in Fig. 3(g) because these video transformer-based methods do not consider patch similarity. Such results lead to unexplainable results for radiologists. In contrast, the results of the proposed method shown in Fig. 3(f) can accurately focus on tumor regions compared to competing methods due to the proposed spatiotemporal deformable encoder, patch loss, and side loss.

The last two rows of Fig. 3 show incorrect predictions of the competing methods. Due to the absence of temporal information, MCTformer fails to classify the images of the last patient. Furthermore, without considering patch similarity in a weakly supervised manner, SIFA and Uniformer fail to correctly classify the third and fourth patients as shown in Fig. 3(d) and (e), respectively. As shown in Fig. 3(f), the results show the effectiveness of the proposed temporal deformable attention encoder and the side loss to drive the model to learn the potential regions of the tumors.

### C. Ablation Study

*1) Performance of different attention mechanisms:* In the ablation study, we compared the proposed method with different attention schemes, including spatial attention [41], spatial attention followed by temporal attention [32], and spatial attention followed by temporal deformable attention, respectively. The results of spatial attention and spatial attention followed by temporal attention are shown in the first row and second row of Table II, respectively. We infer that the precision of the first row in Table II is higher than that of the second row because the spatial attention mechanism utilizes global information, making it more robust against noise. In contrast, spatial attention followed by temporal attention only gets information from the same position, allowing it to concentrate more on temporal changes and thus yields a higher recall value. Compared with the previous two attention mechanisms, the combination of spatial attention and temporal deformable attention automatically selects surrounding suitable tokens, which copes with the alignment problem between two LDCT images. Therefore, the third row of Table II shows that spatial



(a) (b) (c) (d) (e) (f)

Fig. 4: Comparison different attention mechanisms. (a) LDCT image captured in the first year, (b) LDCT image captured in the second year, (c) Results with only spatial attention, (d) Results with spatial attention followed by temporal attention, (e) Results with spatial attention followed by temporal deformable attention, and (f) the ground truth on LDCT images captured in the third year. The localization maps with red borders indicate that the model gives incorrect predictions.

TABLE III: Ablation Study of Different Loss Functions on the NLST dataset.

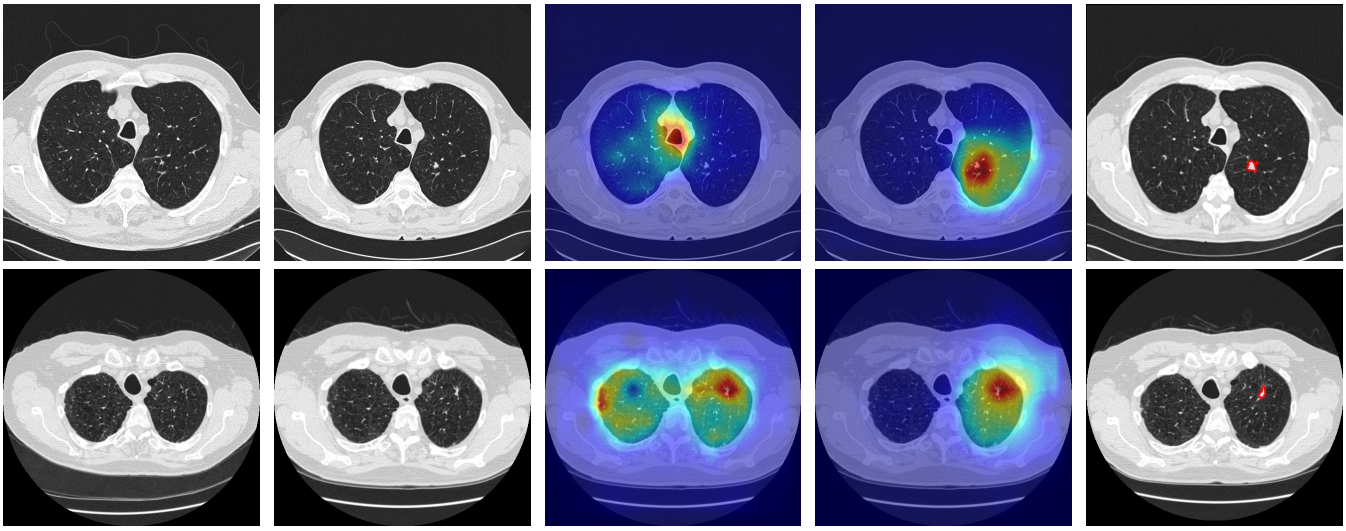
| $\mathcal{L}_{cls}$ | $\mathcal{L}_{pat}$ | $\mathcal{L}_{side}$ | **Accuracy** | **Precision** | **Recall** | **F1-score** | **mIoU** |
|---|---|---|---|---|---|---|---|
| ✓ | | | 0.8072 | 0.7879 | 0.7429 | 0.7647 | 0.4996 |
| ✓ | ✓ | | 0.8795 | 0.8788 | 0.8286 | 0.8530 | 0.5271 |
| ✓ | | ✓ | 0.8916 | 0.8611 | 0.8857 | 0.8732 | 0.5303 |
| ✓ | ✓ | ✓ | **0.9398** | **0.9412** | **0.9143** | **0.9276** | **0.5345** |

attention followed by temporal deformable attention achieves the best results.

Fig. 4 shows the visualization results of different attention mechanisms. The localization maps with red borders indicate that the model gives wrong predictions. The LDCT images captured in the first year and the second year are shown in Fig. 4(a) and (b), respectively. The localization maps of the proposed method with different attention mechanisms including spatial attention [41], spatial attention followed by temporal attention [32], and spatial attention followed by temporal deformable attention are shown in Fig. 4(c), (d), and (e), respectively. The ground truth of the LDCT image captured in the third year is shown in Fig. 4(f). As shown in Fig. 4(a) and (b), the LDCT images captured in the first year and the second year are not well-aligned. Without temporal deformable attention, the models shown in Fig. 4(c) and (d) do not predict cancer risk even when attention maps might be correct as shown in the third row of Fig. 4(d). Thus, the proposed spatial attention followed by temporal deformable attention, which helps solve the alignment problem, can then achieve the best visualization results.

*2) Performance of different loss functions:* Table III shows the results of the evaluation of the effectiveness of loss functions. The results of using only the classification loss $\mathcal{L}_{cls}$ are shown in the first row of Table III. Due to the lack of the patch token loss $\mathcal{L}_{pat}$ and side loss $\mathcal{L}_{side}$, not only the accuracy drops, but also the mIoU value is low, which implies that with only $\mathcal{L}_{cls}$ fails to drive the model to localize the tumors. When combining $\mathcal{L}_{cls}$ and $\mathcal{L}_{pat}$, both the accuracy and mIoU increase compared to only using $\mathcal{L}_{cls}$. When combining $\mathcal{L}_{cls}$ and $\mathcal{L}_{side}$, the accuracy and mIoU are even better than those of combining $\mathcal{L}_{cls}$ and $\mathcal{L}_{pat}$. This is because $L_{side}$ further hints at the locations of the tumors to drive the learning of the model to grasp more accurate tumor locations. As shown in the last row of Table III, combing $\mathcal{L}_{cls}$, $\mathcal{L}_{pat}$, and $\mathcal{L}_{side}$ simultaneously achieves the best results because $\mathcal{L}_{pat}$ and $\mathcal{L}_{side}$ drive the model to learn the locations of tumors.

The visualizations of the results of different losses can be observed in Fig. 5. The LDCT images captured in the first year and the second year are shown in Fig. 5(a) and (b), respectively. The localization maps of combining $\mathcal{L}_{cls}$ and $\mathcal{L}_{pat}$, and combining $\mathcal{L}_{cls}$ and $\mathcal{L}_{side}$ are shown in Fig. 5(c) and (d), respectively. The ground truth of the image captured in the third year is shown in Fig. 5(e). As shown in Fig. 5(c), although combining $\mathcal{L}_{cls}$ and $\mathcal{L}_{pat}$ can provide tumor



(a) (b) (c) (d) (e)

Fig. 5: Comparison between $\mathcal{L}_{pat}$ and $\mathcal{L}_{side}$. (a) LDCT image captured in the first year, (b) LDCT image captured in the second year, (c) Results with $\mathcal{L}_{cls}$ and $\mathcal{L}_{pat}$, (d) Results with $\mathcal{L}_{cls}$ and $\mathcal{L}_{side}$, and (e) the ground truth on LDCT images captured in the third year.

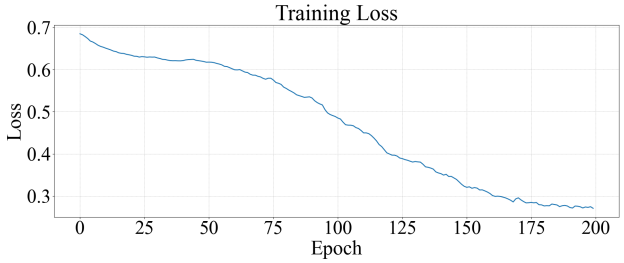



Fig. 6: The training loss.

responses based on weak labels, the locations are not accurate. In contrast, when combining $\mathcal{L}_{cls}$ and $\mathcal{L}_{side}$, the model can better focus on the lung regions guided by the weak labels of the regions, i.e., the left or right lungs, and thus achieves better results. Such results indicate the importance of $\mathcal{L}_{side}$ for tumor localization.

### *D. Discussions*

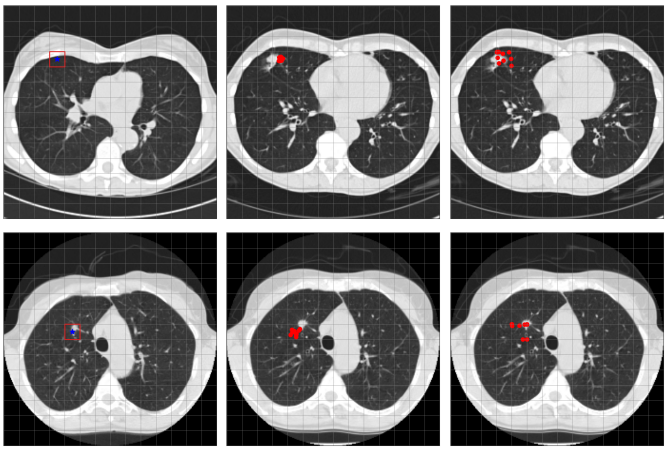
Due to the movement and positions of the patient at different time points, the images of the first year and the second year are not aligned as shown in Fig. 3(a) and (b). Moreover, the variance in tumor locations due to tumor progression also makes the images of the first year and the second year become different in lung regions. These consecutive LDCT images are thus difficult to be registered using image registration methods [42], [43]. In addition, image registration methods aim to align images instead of discovering temporal changes of tumors between two LDCT images. In contrast, our method considers temporal deformable attention among local patches to discover different parts between two LDCT images. The model can then better represent the temporal changes of growing tumor regions between two LDCT images and facilitate cancer prediction.

The training loss is shown in Fig. 6. As shown in Fig. 6, the training loss gradually decreases and converges around 200 epochs. Such results show that the training strategy helps the model learn representative features from the dataset.

The offset network considers the differences between the feature maps $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, and helps discover the temporal changes between two consecutive LDCT images. When the temporal changes of tumor progression are represented by the offset network, the deformable multi-head attention will then have higher feature responses. In this case, the feature responses will be more correlated with the image label when applying the global average pooling function to the features to compute the patch token loss and the side loss. In other words, reducing the patch token loss and the side loss will guide the offset network to discover temporal changes of tumor progression to learn representative deformable multi-head attention features.

Fig. 7 shows the results of the offset network. Fig. 7(a) shows a blue star, which indicates the patch center in the LDCT image captured in the first year. The red points shown in Fig. 7(b) are the outputs of the initial offset network in the LDCT image captured in the second year based on the given patch center. With the guide of the losses, the red points shown in Fig. 7(c) are mapped to different positions to present the temporal changes between the LDCT images of two years. More specifically, some points are moved toward the tumor regions, which indicates that the proposed losses help drive the offset network to discover the temporal changes between two LDCT images captured at different years.

In the case of lung cancer screening, nodules smaller than 0.5 centimeters in LDCT images cannot be definitively classified as benign or malignant, and indiscriminate surgical removal may lead to overtreatment. Follow-up LDCT images conducted after a certain period can assess changes in nodule sizes, with rapidly expanding nodules indicating a higher



(a) (b) (c)

Fig. 7: The results of the offset network. (a) The blue star indicates the patch center in the LDCT image captured in the first year, (b) The red points are initial outputs of the offset network in the LDCT image captured in the second year based on the given patch center, and (c) The red points are outputs of the offset network in the LDCT image captured in the second year after training based on the given patch center. Some offset points are moved toward the tumor regions, which indicate the proposed losses help drive the offset network to discover the temporal changes between two consecutive LDCT images captured at different years.

likelihood of malignancy. LDCT images captured at various time points can be utilized to assess disease trends, cancer screening, and post-treatment follow-up. By analyzing LDCT images from the previous two years, our method can estimate the probability that a nodule becomes malignant in the third year. If the probability is high, early surgical intervention can be performed to eliminate lung cancer. If the probability is low, continued observation or extended intervals between follow-ups can be adopted, thereby reducing the waste of medical resources and demonstrating significant clinical value. There are few models capable of assessing temporal medical imaging. The model developed in this study is innovative and may also be applicable to other diseases in the future.

## IV. Conclusions

In this paper, we propose a novel spatiotemporal deformable transformer for lung cancer risk prediction and weakly supervised localization. The proposed model predicts if a patient will develop cancer in the third year based on two consecutive years of LDCT images taken during routine examinations. The proposed spatiotemporal deformable transformer encoder module captures intra-image and inter-image correlations to represent the tumor progression in LDCT images. To cope with the alignment issues of the LDCT image captured in different years, the temporal deformable attention mechanism is proposed in the encoder module. The patch token loss and side loss are specially designed to drive the model to effectively localize the tumors. With the losses, the proposed method is end-to-end trainable and achieves tumor localization in a weakly supervised manner with only image-level labels. As shown in the experiments, it also outperforms state-of-the-art video classification methods and weakly supervised methods. Currently, we align the 2-D LDCT images of the first year, the second year, and the third year in the same position on the $z$-axis based on the alignment of the carina and the thickness of the 3-D LDCT images of each year to reduce the model training burden of aligning slices on the $z$-axis. In the future, we will consider deformable attention between slices on the $z$-axis to achieve alignment on the $z$-axis within the model architecture to further extend the proposed method to 3-D LDCT image datasets.

## Acknowledgement

All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this paper. The authors would like to thank the National Center for High-Performance Computing (NCHC) for providing computational and storage resources.

## References

[1] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray, "Global cancer observatory: Cancer today. lyon, france: International agency for research on cancer.," 2020.

[2] H. promotion administration, "2020 yearly report of cancer registry."

[3] N. L. S. T. R. Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, 2011.

[4] A. C. of Radiology, "Lung-rads 2022." (Accessed: 2022-11-09).

[5] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Trans. on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.

[6] Z. Chen, Z. Tian, J. Zhu, C. Li, and S. Du, "C-cam: Causal cam for weakly supervised semantic segmentation on medical image," in *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 11666–11675, 2022.

[7] Y.-H. Chang, M.-Y. Lin, M.-T. Hsieh, M.-C. Ou, C.-R. Huang, and B.-S. Sheu, "Multiple field-of-view based attention driven network for weakly supervised common bile duct stone detection," *IEEE J. of Translational Engineering in Health and Medicine*, vol. 11, pp. 394–404, 2023.

[8] C.-W. Chao, D. W. Hwang, H.-W. Tsai, S.-H. Lin, W.-L. Chen, C.-R. Huang, and P.-C. Chung, "Multi-magnification attention convolutional neural networks [ai-explained]," *IEEE Computational Intelligence Magazine*, vol. 18, no. 3, pp. 54–55, 2023.

[9] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen, and D. Visvikis, "Current and emerging trends in medical image segmentation with deep learning," *IEEE Trans. on Radiation and Plasma Medical Sciences*, vol. 7, no. 6, pp. 545–569, 2023.

[10] S.-K. Huang, Y.-T. Yu, C.-R. Huang, and H.-C. Cheng, "Cross-scale fusion transformer for histopathological image classification," *IEEE J. of Biomedical and Health Informatics*, vol. 28, no. 1, pp. 297–308, 2024.

[11] Y.-M. Kuo, J.-C. Sheng, C.-H. Lo, Y.-J. Wu, and C.-R. Huang, "Cross-scale guidance integration transformer for instance segmentation in pathology images," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 6, pp. 413–419, 2025.

[12] S. Shen, S. X. Han, D. R. Aberle, A. A. Bui, and W. Hsu, "An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification," *Expert systems with applications*, vol. 128, pp. 84–95, 2019.

[13] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing covid-19, pneumonia, and lung cancer chest diseases," *Computers in biology and medicine*, vol. 132, p. 104348, 2021.

[14] H. Li, Q. Song, D. Gui, M. Wang, X. Min, and A. Li, "Reconstruction-assisted feature encoding network for histologic subtype classification of non-small cell lung cancer," *IEEE J. of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4563–4574, 2022.

[15] H. Wang, H. Zhu, and L. Ding, "Accurate classification of lung nodules on ct images using the transunet," *Frontiers in Public Health*, vol. 10, p. 1060798, 2022.

[16] Z. Liao, Y. Xie, S. Hu, and Y. Xia, "Learning from ambiguous labels for lung nodule malignancy prediction," *IEEE Trans. on Medical Imaging*, vol. 41, no. 7, pp. 1874–1884, 2022.

[17] L. Yi, L. Zhang, X. Xu, and J. Guo, "Multi-label softmax networks for pulmonary nodule classification using unbalanced and dependent categories," *IEEE Trans. on Medical Imaging*, vol. 42, no. 1, pp. 317–328, 2022.

[18] A. Masood, B. Sheng, P. Li, X. Hou, X. Wei, J. Qin, and D. Feng, "Computer-assisted decision support system in pulmonary cancer detection and stage classification on ct images," *Journal of biomedical informatics*, vol. 79, pp. 117–128, 2018.

[19] N. Nasrullah, J. Sang, M. S. Alam, M. Mateen, B. Cai, and H. Hu, "Automated lung nodule detection and classification using deep learning combined with multiple strategies," *Sensors*, vol. 19, no. 17, p. 3722, 2019.

[20] O. Ozdemir, R. L. Russell, and A. A. Berlin, "A 3d probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose ct scans," *IEEE Trans. on Medical Imaging*, vol. 39, no. 5, pp. 1419–1429, 2019.

[21] L. Chen, K. Liu, H. Shen, H. Ye, H. Liu, L. Yu, J. Li, K. Zhao, and W. Zhu, "Multimodality attention-guided 3-d detection of nonsmall cell lung cancer in 18f-fdg pet/ct images," *IEEE Trans. on Radiation and Plasma Medical Sciences*, vol. 6, no. 4, pp. 421–432, 2022.

[22] A. Kumar, M. Fulham, D. Feng, and J. Kim, "Co-learning feature fusion maps from pet-ct images of lung cancer," *IEEE Trans. on Medical Imaging*, vol. 39, no. 1, pp. 204–217, 2019.

[23] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation," *IEEE J. of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3507–3516, 2021.

[24] H. Seo, L. Yu, H. Ren, X. Li, L. Shen, and L. Xing, "Deep neural network with consistency regularization of multi-output channels for improved tumor detection and delineation," *IEEE Trans. on Medical Imaging*, vol. 40, no. 12, pp. 3369–3378, 2021.

[25] B. Zhang, S. Qi, Y. Wu, X. Pan, Y. Yao, W. Qian, and Y. Guan, "Multi-scale segmentation squeeze-and-excitation unet with conditional random field for segmenting lung tumor from ct images," *Computer Methods and Programs in Biomedicine*, vol. 222, p. 106946, 2022.

[26] A. Bhattacharjee, R. Murugan, T. Goel, and S. Mirjalili, "Pulmonary nodule segmentation framework based on fine-tuned and pretrained deep neural network using ct images," *IEEE Trans. on Radiation and Plasma Medical Sciences*, vol. 7, no. 4, pp. 394–409, 2023.

[27] P. Huang, C. T. Lin, Y. Li, M. C. Tammemagi, M. V. Brock, S. Atkar-Khattra, Y. Xu, P. Hu, J. R. Mayo, H. Schmidt, *et al.*, "Prediction of lung cancer risk at follow-up screening with low-dose ct: a training and validation study of a deep learning method," *The Lancet Digital Health*, vol. 1, no. 7, pp. e353–e362, 2019.

[28] M. T. Lu, V. K. Raghu, T. Mayrhofer, H. J. Aerts, and U. Hoffmann, "Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model," *Annals of Internal Medicine*, vol. 173, no. 9, pp. 704–713, 2020.

[29] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[30] P. G. Mikhael, J. Wohlwend, A. Yala, L. Karstens, J. Xiang, A. K. Takigami, P. P. Bourgouin, P. Chan, S. Mrah, W. Amayri, *et al.*, "Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography," *Journal of Clinical Oncology*, vol. 41, no. 12, pp. 2191–2200, 2023.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int'l Conf. Computer Vision*, pp. 10012–10022, 2021.

[32] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proc. Int'l Conf. Machine Learning*, vol. 2, p. 4, 2021.

[33] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proc. Int'l Conf. Computer Vision*, pp. 6824–6835, 2021.

[34] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 4310–4319, 2022.

[35] F. Long, Z. Qiu, Y. Pan, T. Yao, J. Luo, and T. Mei, "Stand-alone inter-frame attention in video models," in *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 3192–3201, 2022.

[36] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2023.

[37] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int'l Conf. Computer Vision*, pp. 10347–10357, 2021.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. Int'l Conf. Computer Vision*, pp. 618–626, 2017.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, X. Z. D. Weissenborn, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int'l Conf. Learning Representations*, 2021.

[42] J. Lu, R. Jin, and E. Song, "Pyramid convolutional recurrent network for serial medical image registration with adaptive motion regularizations," *IEEE Trans. on Radiation and Plasma Medical Sciences*, vol. 8, no. 7, pp. 800–813, 2024.

[43] R. Miura, M. Nakamura, and M. Nakao, "Image-to-volume deformable registration by learning displacement vector fields," *IEEE Trans. on Radiation and Plasma Medical Sciences*, vol. 9, no. 1, pp. 69–82, 2025.