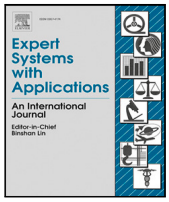
Expert Systems With Applications 258 (2024) 125067

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Segmentation prompts classification: A nnUNet-based 3D transfer learning framework with ROI tokenization and cross-task attention for esophageal cancer T-stage diagnosis

Chen Li [a,b,1], Runyuan Wang [a,1], Ping He [d], Wei Chen [b,*], Wei Wu [c,*], Yi Wu [a,*]

[a] Department of Digital Medicine, College of Biomedical Engineering and Medical Imaging, Army Medical University (Third Military Medical University), Chongqing, China
[b] School of Computer, National University of Defense Technology, Changsha, China
[c] Department of Thoracic Surgery, Southwest Hospital, Army Medical University (Third Military Medical University), Chongqing, China
[d] Department of Cardiac Surgery, Southwest Hospital, Army Medical University (Third Military Medical University), Chongqing, China

ARTICLE INFO

Keywords:
Transfer learning
Cross-task attention
ROI tokenization
nnU-Net
3D segmentation
3D classification
Esophageal cancer T-stage diagnosis
CT images

ABSTRACT

The computer-aided diagnosis system for esophageal cancer (EC) holds vital significance in EC diagnosis and treatment making, with a primary focus on accurate segmentation of EC-related organs and classification of EC's T-stage. Above two tasks are closely related and crucial in assisting surgeon segment and diagnose cancer early. Note that this paradigm is still at its infancy and limited by closely related open issues: (1) how to link the complementary relationship between these two tasks and improve the originally poor performance? and (2) how to determine whether the tumor has invaded the surrounding muscle layers from CT images? Aiming at these issues, this study develops nn-TransEC, a 3D transfer learning framework that builds upon nnU-Net and synergizes segmentation and classification. nn-TransEC focuses on prompting fine-grained classification of EC's T-stage with the aid of prior segmentation, which is implemented in two parts: (1) A nnUNet-configured multi-task learning network (nn-MTNet) is designed for complementary segmentation of EC-related organs and classification of EC's T-stages with cross-task attention gates and transfer learning. (2) A knowledge-embedded ROI tokenization method (KRT) is defined to mimic the diagnostic workflow of doctors for classifying EC's T-stage. KRT is implemented by cropping the most concerned regions from entire CT volume based on prior segmentation. Experiments have been conducted on a private dataset collected from 169 patients with confirmed EC through pathological diagnosis. Our proposed nn-TransEC is compared against the state-of-the-art counterparts (e.g., nnU-Net and nnFormer), and results demonstrate that: nn-TransEC excels in all compared methods in multi-organ segmentation and classification of EC's T-stages, with 3D Dice of EC and average AUC of T-stages reaching 0.844 and 0.941, respectively. In contrast, the state-of-the-art method nnFormer achieves 0.814 and 0.927, respectively. Meanwhile, nn-TransEC also outperforms state-of-the-art multi-task learning models in joint segmentation and classification, with Hausdorff Distance of EC and average precision of T-stages reaching 8.497 and 0.845, respectively. In contrast, the state-of-the-art method TransMT-Net achieves 12.206 and 0.730, respectively.

1. Introduction

Esophageal cancer (EC), as a lethal malignancy, is the sixth leading cause of cancer-related deaths worldwide (Enzinger & Mayer, 2003), especially in the Asia and Africa (Malhotra et al., 2017; Rustgi & El-Serag, 2014). Nowadays, EC has become one of the deadliest but least studied cancers in the world. Compared to other cancers, EC is more aggressive and has a worse prognosis rate, i.e., a only 21% overall 5-year survival (Atlanta, 2023; Siegel, Miller, Wagle, & Jemal, 2023). Therefore, accurate diagnosis of EC is of great clinical significance and research value.

Early detection and treatment of EC still remain the most effective means of reducing mortality rates and improving survival prognosis (Hosseini, Asadi, Emami, & Ebnali, 2023). The TNM system is most widely-used to stage EC in clinical practice, which effectively evaluates

* Corresponding authors.
E-mail addresses: lichen14@nudt.edu.cn (C. Li), liyaolareina@stu.xjtu.edu.cn (R. Wang), hp@tmmu.edu.cn (P. He), chenwei@nudt.edu.cn (W. Chen), ww@tmmu.edu.cn (W. Wu), wuy1979@tmmu.edu.cn (Y. Wu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.eswa.2024.125067
Received 7 February 2024; Received in revised form 1 July 2024; Accepted 7 August 2024
Available online 22 August 2024

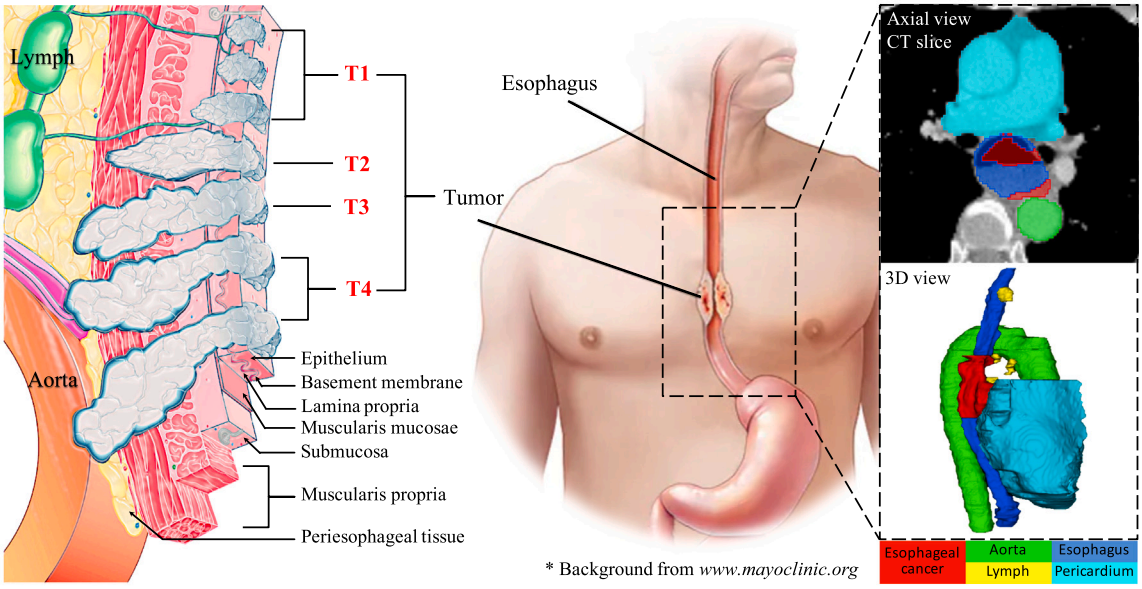


Fig. 1. Illustration of esophageal cancer (EC) in the human body and the basis for determining the tumor's T-stages. As can be observed, tumors represented in CT images are irregular in shape, vary significantly in size, have a similar appearance to adjacent tissues with ambiguous boundaries, and are grossly imbalanced in category. Therefore, recognizing EC accurately and then performing T-staging diagnosis can be challenging due to the issues mentioned above.
Source: The anatomy of human tissue on the left is repainted from Pennathur, Gibson, Jobe, and Luketich (2013), Rice, Ishwaran, Ferguson, Blackstone, and Goldstraw (2017).

how advanced the tumor is and if it has spread, thus helping doctors make treatment decisions. In the TNM system, T-stage is utilized as the main metric for determining the depth of the primary tumor in the esophagus, as well as the extent of its spread in the oesophagus and nearby organs. In the traditional T-staging diagnosis of EC, manual labeling the tumor and its adjacent structures from medical images is an indispensable step to provide precise references for subsequent T-stage diagnosis and treatment decisions for EC. However, this time-consuming and inefficient manual labeling increases the burden of doctors and does not contribute to the reasonable utilization of medical resources. Besides, it remains challenging for even experienced doctors to accurately segment EC and determine its T-stages because tumors and adjacent structures often have a similar appearance, resulting in a severe imbalance in categories, as shown in Fig. 1.

Numerous attempts have been made to develop automatic EC diagnosis systems, especially based on *neural networks*. Cutting-edge methods largely focus on tumor segmentation (Fechter et al., 2017; Guo et al., 2020; Lin et al., 2023; Yousefi et al., 2021, 2018) or/and classification (Ferreira, Domingues, Sousa, Sampaio, & Santos, 2020; Sui et al., 2021; Van Riel, Van Der Sommen, Zinger, Schoon, & de With, 2018; Zhao, Li, Liu, Yu, & Zhao, 2022). In the above studies, endoscopy and CT images are the two most commonly used medical images to identify esophageal diseases. However, the studies mentioned above only performed simple organ-tumor segmentation or benign-malignant discrimination. Few of them explored the strong complementary relationship between these two critical tasks to the extent that this issue has long been underestimated, thus limiting their clinical application and performance (marked as issue #1).

To address issue #1, there is been a growing body of research centered around multi-task learning (Takeuchi et al., 2021; Tang, Yu, et al., 2022; Tang et al., 2023; Wu, Ge et al., 2021; Yu, Tang, Cheang, Yu, & Choi, 2022), which has increasingly emphasized the joint segmentation and classification. Most of them are designed for endoscopic images and few for CT images. However, it is impossible for doctors to discriminate the EC's T-stage directly from endoscopic images, and the illumination conditions of endoscopic images may introduce additional noise and artefacts. These factors limit the clinical applications of endoscopic images in the esophageal cancer T-stage diagnosis. In contrast, CT images are more standardized and easier to use, providing a non-invasive and holistic view of the tumor and surrounding tissues, reducing patient discomfort and minimizing artefacts. Therefore, CT images have become a preferred choice to carry out computer-aided EC diagnosis, assisting doctors to determine whether the tumor has invaded the surrounding muscle layer and diagnose EC's T-stage. However, there were only a few studies on multi-task learning based on CT images of EC, and all of them neglected the important issue of fine-grained classification of EC's T-stage and overlooked the incorporation of prior knowledge, resulting in a deficiency in the guidance of cross-task learning training (marked as issue #2).

Consequently, computer-aided system for EC's T-stage diagnosis is still at its infancy. When handling challenging but critical applications in clinical practice, such as segmenting EC from CT images and identifying its T-stages, this paradigm is refrained by the above two issues. Aiming at these issues, this paper argues that an ideal diagnosis system for EC would be expected to have at least two capabilities:

- A medically sound diagnostic process that mimics the surgeon's T-staging diagnosis of EC is essential for computer-aided diagnostic (CAD) systems. The idea of incorporating clinical diagnostic expertise into CAD systems is now a well-established consensus, recognized for its potential to increase the accuracy and reliability (Begon, Lockhart, Metreau, & Dhumeaux, 1979; Xie et al., 2021; Yanase & Triantaphyllou, 2019; Yang et al., 2021). In this study, we have embraced this consensus and conducted extensive clinical surveys to draw inspiration for network design. Specifically, in the preliminary phase, this study has consulted with several senior surgeons to understand their diagnostic approaches to EC, which invariably involves three sequential steps: (1) identifying the region of interest (ROI) from the CT image, such as EC, esophagus, and its surrounding organs; (2) determining whether the tumor has invaded the surrounding muscle layers based on the selected ROI; and (3) classifying the T-stage of EC.[2] After that, this study, for the first time, conceptualizes the aforementioned diagnostic process and bridges the initial two

[2] More details about the T-staging diagnosis of EC in clinical practice are released in Appendix A.

steps through a Knowledge-embedded ROI Tokenization method (KRT). More details are introduced in Section 3.4.
- An advanced multi-task learning framework that enables the model to adapt parameters based on multi-organ segmentation and fine-grained T-stage classification. nnU-Net[3] is a self-configuring deep learning framework designed for medical images, which can automatically and systematically configure the key hyperparameters of neural networks according to different tasks (Isensee, Jaeger, Kohl, Petersen, & Maier-Hein, 2021). Since its proposal, nnU-Net and its variants have achieved remarkable success in various medical image segmentation tasks (Isensee et al., 2024; McConnell, Ndipenoch, Cao, Miron, & Li, 2023). Consequently, this paper selects the widely recognized nnU-Net as the baseline model and then redesigns a multi-task learning framework for EC diagnosis.

Therefore, it is desirable to design a multi-task learning framework based on an advanced neural network, where joint training of multi-organ segmentation and fine-grained T-stage classification from CT images are carried out. Then, this study focuses on prompting fine-grained classification of EC's T-stage with the aid of prior segmentation results, improving the performance.

Specifically, this study proposes a 3D transfer learning framework (**nn-TransEC**) that builds upon a nnUNet-configured multi-task learning network (nn-MTNet) and works with a knowledge-embedded ROI tokenization method (KRT). nn-MTNet consists of a 3D encoder–decoder for segmentation of EC-related organs and a shared 3D encoder for fine-grained classification of tumor's T-stage. Then, a novel cross-task attention gate is designed to bridge the segmentation encoder and classification encoder for highlighting features in ROI. After that, nn-MTNet is configured by the modified nnU-Net self-configuration strategy, and then works in three phases:

- Coarse segmentation for ROI localization and pre-training (Section 3.3): segmentation network takes 3D CT volumes as input and outputs 3D segmentation masks, which contains pixel-wise predictions about five EC-related thoracic organs. Above segmentation masks are then sent to the second phase for the subsequent tokenization. The pretrained encoder of segmentation network is transferred to the third phase for better initialization.
- ROI tokenization for cropping the CT volume based on KRT method (Section 3.4): The KRT method is motivated by mimicking the diagnostic workflow of doctors for classifying EC's T-stage. The KRT first separates the locations of our concerned EC as ROI from the previous segmentation predictions, which are then used to crop the entire CT volume to obtain patch relevant for T-staging diagnosis.
- Fine classification with pre-trained weights and cropped CT patch (Section 3.5): classification network is initialized by the pre-trained encoder in the first phase and then takes the cropped CT patch as input. Besides, cross-task attention gates are used to highlight the features extracted from segmentation encoder. Finally, network outputs the predicted probabilities of the four stages from T1 to T4 and selects the category with the highest probability as the final T-staging result.

Extensive experiments have been performed on a private dataset collected from 169 patients with confirmed esophageal cancer. The proposed nn-TransEC is compared against the state-of-the-art counterparts (e.g., nnU-Net (Isensee et al., 2021) and nnFormer (Zhou et al., 2023)) and multi-task learning methods (e.g., TransMT-Net (Tang et al., 2023)). nn-TransEC's performances on segmentation of EC-related organs and fine-grained classification of EC's T-stage have been evaluated. Ablation studies are also conducted to analyze the contribution of each component in nn-TransEC. Supplementary tests have been made to examine the effectiveness of the proposed ROI tokenization with other state-of-the-art methods. nn-TransEC's model interpretation has also been evaluated with GradCAM (Selvaraju et al., 2017, 2020). The main **contributions** of this study are as follows:

- This paper proposes a 3D transfer learning framework to mimic how doctors diagnoses EC's T-stages from CT images. This framework successfully links the complementary relationship between segmentation of EC-related organs and classification of tumor's T-stage, achieving high accuracy.
- The volumes of EC-related organs measured by nn-TransEC achieve a good agreement with doctors' manually annotated results, i.e., the mean difference of volume between the ground truth and prediction is 1.911 cm^3 for the EC, 1.124 cm^3 for aorta, 2.880 cm^3 for esophagus, and 1.130 cm^3 for lymph. Thus the proposed nn-TransEC can serve as a relief from time-consuming pixel-by-pixel manual annotation in the esophageal cancer diagnosis.
- The proposed KRT and transfer learning methods contribute to aligning regions of interest with doctors in T-stage diagnosis, thus providing a clinically interpretable enhancement for fine T-staging classification of EC. After introducing the KRT method into the multi-task learning framework, the average AUC for classification of EC's T-stages improves from 0.894 to 0.925.

2. Related work

Extensive research has been directed towards enhancing medical image analysis through deep learning, with a particular focus on organ-tumor segmentation and benign-malignant classification with diverse medical imaging modalities. This section will introduce the relevant cutting-edge methods associated with this domain.

2.1. Medical image segmentation of tumors

Accurate tumor segmentation from medical images is of significant importance, as it assists doctor in making more precise diagnoses and formulating more accurate surgical and radiotherapy plans by delineating tumor boundaries. There exists several deep learning-based studies focused on segmenting tumors and their associated surrounding organs.

Table 1 gives an overview of some state-of-the-art classification methods. Detailed introductions are released as follows.

Zhou, Siddiquee, Tajbakhsh, and Liang (2019) design a nest neural network (UNet++) to enhance the medical image segmentation. The UNet++ model addresses the limitations of existing U-Net (Ronneberger, Fischer, & Brox, 2015) and FCN (Long, Shelhamer, & Darrell, 2015) variants by proposing an efficient ensemble of U-Nets with varying depths. It redesigns skip connections and employ deep supervision to allow for more flexible feature fusion across different semantic scales and resolutions, leading to improved segmentation quality for objects of varying sizes. The redesigned skip connections allow for better integration of features from different layers, leading to more robust and accurant segmentation in six medical image segmentation datasets. However, the cascaded skip connections in nested U-Net might increase the complexity of the network, which could affect computational efficiency.

Hatamizadeh, Tang, et al. (2022) have made a contribution to medical image segmentation with the introduction of the UNETR model, which represents an integration of U-Net and Vision Transformer (ViT) (Dosovitskiy et al., 2020). UNETR directly receives 3D patches as input and connects the CNN-based decoder and ViT encoder via skip connection. This innovative approach has been effectively applied to the 3D segmentation of brain tumors in MRI images, achieving new state-of-the-art performance on the BTCV[4] leaderboard and

[3] nnU-Net stands for the "no new U-Net".

[4] https://www.synapse.org/Synapse:syn3193805/wiki/217785

outperforming most methodologies on tasks in the MSD challenge. Building upon UNETR, the same team proposes the Swin UNETR model (Hatamizadeh et al., 2021; Tang, Yang, et al., 2022) by replacing the Vision Transformer with the Swin Transformer (Liu et al., 2021) in the UNETR. The proposed Swin UNETR has further advanced the state-of-the-art in brain tumor segmentation, achieving top rank in the MSD challenge.[5] However, these Transformer-based methods have a larger number of parameters, requiring substantial computing resources and high-quality annotated data for training. Besides, limited data constrains their ability to segment tiny organs or tumors across various medical imaging modalities.

Isensee et al. (2021) proposes a self-configuring deep learning framework (nnU-Net) for medical image segmentation, which extracts data fingerprints based on the characteristics of the dataset and then provides key hyper-parameters of neural networks. As an out-of-the-box tool, nnU-Net has achieved top performance in the *Medical Segmentation Decathlon*[6] challenge and sets a new state of the art in 33 of 53 target structures evaluated, demonstrating its effectiveness in capturing intricate anatomical details and its potential utility in diverse clinical applications. However, few studies have investigated the application of nnU-Net in multi-task learning tasks, and there is a lack of optimization for nnU-Net in the domain of medical image classification. Sequentially, Zhou et al. (2023) propose nnFormer that introduces local and global volume-based self-attention mechanism to segment 3D tumors from medical images. nnFormer complements nnU-Net in ensemble modeling and replaces the traditional concatenation operations in skip connections with skip attention, delivering greater computational efficiency with lower HD95. There have been subsequent studies applying nnU-Net and nnFormer to various other tumor segmentation tasks, such as Hatamizadeh, Xu, et al. (2022) implemented nnU-Net and nnFormer on liver tumor segmentation task from CT images, Sang et al. (2024) implemented nnU-Net and nnFormer on 3D rectal tumor segmentation from MRI images. However, these methods necessitate a significant volume of annotated data to train the Transformer model, which is prone to overfitting when the amount of data is constrained.

Drawing inspiration from the state-of-the-art segmentation models, particularly the nnU-Net, which has demonstrated exceptional performance and robust generalization across a variety of organs and tumors within diverse medical imaging modalities, this study decides to utilize the nnU-Net as the backbone. To address the defects of nnU-Net in the concerned diagnosis of esophageal cancer, this paper introduces targeted modifications to both its network and self-configuration strategy, tailored to enhance its performance in multi-task learning. Meanwhile, the paper of nnFormer (Zhou et al., 2023) also suggests that using pre-training for neural networks may be a solution to achieve better training outcomes.

2.2. Medical image classification of tumors

Accurate tumor classification from medical images, facilitated by deep learning, provides essential qualitative information about tumors, aiding doctors in the recognition of various tumor types. This advancement significantly aids in the early diagnosis and the development of personalized treatment plans, thereby reducing the workload for medical professionals and enhancing the efficiency of the diagnostic process. Current research typically includes coarse-grained discrimination between normal, benign and malignant tumor/lesion, and fine-grained staging of tumor malignancy.

Table 2 gives an overview of some state-of-the-art classification methods. Detailed introductions are released as follows.

Table 1
An overview of some state-of-the-art methods for medical image segmentation of tumors.

Zhou et al. (2019)	Task info	Segmentation of cell, tumor, liver and lung nodule in microscopy, CT and MRI images.
	Method	Explore to process and integrate information at different scales or resolutions in a Nested UNet (UNet++) for preserving spatial information.
	Achievement	The redesigned skip connections allow for better integration of features from different layers, leading to more robust segmentation.
	Limitation	Cascaded skip connections might increase the complexity of the network, which could affect computational efficiency.
Hatamizadeh, Tang, et al. (2022)	Task info	3D segmentation of abdominal organs, brain tumors in CT and MRI images.
	Method	Propose the UNETR model that harnesses the strengths of U-Net and ViT to capture long-range dependencies and global multi-scale features for medical image segmentation.
	Achievement	UNETR demonstrates new state-of-the-art performance on the BTCV leaderboard and outperforms most methodologies on tasks in the MSD dataset.
	Limitation	Fail to segment tiny organ or tumor within diverse medical imaging modalities.
Hatamizadeh et al. (2021)	Task info	3D segmentation of brain tumors in MRI images.
	Method	Propose the Swin UNETR model that integrates a hierarchical Swin Transformer as encoder to extract feature at different resolutions for medical image segmentation.
	Achievement	The hierarchical encoder and effective modeling of long-range dependencies contribute to its top rank in the BraTS 2021 challenge.
	Limitation	Swin Transformer has a larger number of parameters, requiring substantial computing resources and high-quality annotated data for training.
Isensee et al. (2021)	Task info	Various organ/tumor segmentation in multi-modality medical images.
	Method	Propose a self-configuring deep learning framework (**nnU-Net**) that provides key hyper-parameters of neural networks for medical image segmentation.
	Achievement	As an out-of-the-box tool, nnU-Net surpasses most specialized segmentation methods and sets a new state of the art in 33 of 53 target structures evaluated.
	Limitation	⋆Fail to transfer its self-configuring capability to multi-task learning. ⋆Lack targeted optimization for medical image classification.
Zhou et al. (2023)	Task info	3D segmentation of tumors, cardiac and abdominal organs in MRI and CT images.
	Method	Integrate CNNs and Transformers in a hybrid architecture (**nnFormer**) and introduce local and global self-attention mechanism for enhancing segmentation.
	Achievement	nnFormer complements nnU-Net in ensemble modeling, delivering greater computational efficiency with lower HD95.
	Limitation	Require a substantial amount of annotated data for training Transformer, which is susceptible to overfitting with limited data.

Chattopadhyay et al. (2022) introduce a dual-shuffle attention guided deep learning model for breast cancer classification, which is constructed based on the ShuffleNet (Zhang, Zhou, Lin, & Sun, 2018) architecture and incorporates a dual-shuffle **residual** block and a channel **attention** block to augment the encoder's learning capabilities. This study has demonstrated superior performance over several state-of-the-art methods in classifying breast cancer in histopathological images. However, the model's generalizability is influenced by the

[5] https://decathlon-10.grand-challenge.org/evaluation/challenge/leaderboard/

[6] http://medicaldecathlon.com/results/

Table 2
An overview of some state-of-the-art methods for medical image classification of tumors.

Mishra et al. (2022)	Task info	Classification of breast tumor in histopathological images.
	Method	Incorporate a dual-shuffle residual block and a channel attention block to augment the CNN's classifying capability.
	Achievement	Integration of residual connections and attention mechanisms contributes to achieving top performance in classification of breast cancer across various magnification levels.
	Limitation	⋆ The proposed model exhibits limited generalizability for other tumor classification tasks. ⋆ The model's training is longer compared to other state-of-the-art models due to its complexity.
Sun, Chen, Fu, and Liu (2023)	Task info	Classification of skin lesion in dermatoscopic images
	Method	Propose a data-driven based deep supervision method that selects network layers matching the shape of ROI, thereby improving the extraction of classification features.
	Achievement	The effectiveness of deep supervision is demonstrated through experiments on various datasets, achieving comparable or better performance than the challenge winners.
	Limitation	The method's accuracy is limited by object masks generated via activation mapping, which is sensitive to the threshold values and not as accurate as segmentation methods do to extract ROI.
Zhu, Jin, et al. (2024)	Task info	Classification of tumor in adrenocortical carcinoma in CT images
	Method	Create a rapid and accurate TNM staging system that uses a self-supervised contrastive learning method (SimCLR) to pre-train a ResNet50 for TNM staging without manual labeling.
	Achievement	Transfer learning works in the tumor's TNM staging system with generalizability and reduces reliance on surgical invasions and subjective biases in traditional staging methods.
	Limitation	⋆The used neural network is too simple to support the complex TNM staging diagnosis for tumors. ⋆The scale of the experiments limits further method validation processes.
	Task info	Classification of lung cancer in CT images
	Method	Integrate both pathological and radiological features under a feature fusion framework (SGHF-Net) to generate hybrid features for accurate classification.
	Achievement	It demonstrates the performance improvement by self-generating hybrid features from single-modality inputs, which is a novel approach in medical image classification.
	Limitation	Simple network architecture struggles to provide high-quality representations for the classification of tiny tumors.

small scale of the dataset when applied to a variety of tumor classification tasks. Besides, the model's training period is longer compared to other state-of-the-art networks due to its complexity.

Mishra et al. (2022) propose a data-driven**deep supervision** strategy to enhance the feature extraction capabilities of the encoder with the help of activation mapping. The integration of deep supervision into the CNN framework has led to significant performance improvements in classification of various skin lesions and tumors from dermatoscopic images task. The effectiveness of deep supervision is demonstrated through experiments on various datasets, achieving comparable or better performance than the ISIC 2016, 2017, and 2018 (Codella et al., 2018) challenge winners.[7] However, the reliance on activation mapping for object size approximation could be sensitive to the choice of threshold values, which is not as accurate as segmentation methods do to extract ROI and thus may affect the accuracy of the object mask.

Sun et al. (2023) employ a self-supervised contrastive learning method (SimCLR (Chen, Kornblith, Norouzi, & Hinton, 2020)) for pre-training the ResNet50 network. Subsequently, the initialized classification model is utilized for TNM staging of adrenocortical carcinoma tumor. This work has demonstrated the effectiveness of the transfer learning mechanism in the tumor classification task in medical images. However, the neural network RestNet50 used in this paper is too simplistic, making it difficult to support the complex task of TNM staging diagnosis for tumors, especially with lower accuracy in T-staging diagnosis. Besides, the insufficient scale of experiments in the paper is also limited.

Zhu, Jin, et al. (2024) introduce an effective feature fusion framework, namely SGHF-Net. This framework integrates synthetic pathological features with radiological features for accurately classifying lung cancer subtypes on CT images. A Pathological Feature Synthetic Module (PFSM) is proposed to derive ground truth label from corresponding pathological images, and a Radiological Feature Extraction Module (RFEM) is proposed to extract features from CT images. This paper demonstrates that such hybrid features generation from CT images input leads to improved classification accuracy. However, the network architecture is too simple to effectively support the classification of tiny tumors by the SGHF-Net in a variety of applications.

The successful application of these literature has not only enhanced the performance of neural networks but also provided valuable insights that have informed the approach of this paper. At the same time, this study has also found that there is a noticeable imbalance in the research focus on tumor classification tasks, where binary classification studies on tumor benignancy or positivity far outweighing fine-grained multi-class tumor classification, such as classification of malignancy levels or TNM-stages of tumors. In clinical practice, the demand for fine-grained multi-class tumor classification far exceeds that for binary classification of tumor malignancy. This disparity stems from the increased complexity associated with multi-class classification, which often leads to suboptimal performance when approached with direct end-to-end classification models. Building upon the existing literature on medical image classification of tumors, where researchers have adopted strategies to enhance the encoder to augment the neural network's ability to extract tumor's feature, this paper plans to leverage **residual connections** and **attention mechanism** to further refine the feature extraction. This study aims to tackle sophisticated multi-class classification challenges involving tiny tumors, while also incorporating the **transfer learning** mechanism to improve performance.

2.3. Multi-task learning for joint segmentation and classification

As discussed in the aforementioned introduction to tumor segmentation and classification, conducting these tasks separately has notable limitations in cancer diagnosis, including: (1) **Information isolation**, where independent training limits the classifier's access to spatial and contextual information and hinders the segmentation model from utilizing discriminative class features, potentially degrading accuracy in complex tasks, and (2) **Reduced efficiency**, independent training can compromise neural network efficiency and necessitates retraining of parameters for cross-task learning, thus increasing time costs. Consequently, researchers have endeavored to integrate these two closely interrelated tasks into a unified framework that employs **multi-task learning** holds the promise of mutual enhancement. By leveraging the synergies between segmentation and classification, this integrated framework is expected to improve overall performance by providing a more comprehensive understanding of tumor characteristics.

Multi-task learning (MTL) belongs to transfer learning (Pan & Yang, 2009) while MTL tries to learn the target and source task simultaneously. Rich Caruana (Caruana, 1997) gives the widely recognized definition of MTL, which is defined as a learning paradigm to transfer

[7] https://challenge.isic-archive.com

domain knowledge between related tasks so that improves generalization. In deep learning, there are two most commonly implementations (Zhang & Yang, 2018) of MTL:

- **Hard parameter sharing** represents that different tasks share the same parameters in first few hidden layers of the network, but have their own output structures in the later layers. This implementation is the most common strategy used for MTL, which helps the model to better understand and extract the underlying data representations. This yields a multi-task generalist model rather than a model that specializes in certain tasks. Representative works in this area include (Caruana, 1997; Harouni, Karargyris, Negahdar, Beymer, & Syeda-Mahmood, 2018; Mormont, Geurts, & Marée, 2020). However, there are still some limitations. For instance, hard parameter sharing neglects to explore the interaction between tasks, which means if the correlation between tasks is weak or the optimization process is misleading, existing parameters may not be able to fit MTL efficiently and thus can lead to performance degradation.
- **Soft parameter sharing** is different from hard parameter sharing in that each task has an independent network structure (Misra, Shrivastava, Gupta, & Hebert, 2016; Strezoski, van Noord, & Worring, 2019), and MTL performs co-optimization of parameters by setting constraints. For example, (Han, Zhang, Song, & Xie, 2014; Lozano & Swirszcz, 2012; Wang, Bi, Yu, & Sun, 2014) utilized ℓ_1 regularization for optimization. In addition, soft parameter sharing encourages cross-task interactions by transferring task-related knowledge to help improve the performance, such as (Chen, Wang, Shi, Liu, & Yu, 2018; Wu, Gao, et al., 2021; Xie, Zhang, Xia, & Shen, 2020). However, such constraint-based implementation is inefficient and requires more computational overhead and storage space to train and store multi-models due to the use of multi-models.

Therefore, both hard parameter sharing and soft parameter sharing MTL approaches possess their own inherent limitations. Given this, an approach that combines the strengths of both methods is highly desirable. Specifically, it would be beneficial to retain the network structure design characteristic of hard parameter sharing, while simultaneously enabling the cross-task interaction inherent to soft parameter sharing. This viewpoint provides the fundamental idea for our design of a MTL framework that integrates hard parameter sharing and soft parameter sharing.

In the field of medical image segmentation and classification, which is of interest in this paper, MTL has also received an increasing amount of research and application.

Table 3 gives an overview of some state-of-the-art MTL methods. For example,

Zhou et al. (2021) propose a soft parameter sharing MTL framework for joint segmentation and classification of tumors in breast ultrasound images. The proposed framework consists of an auto-encoder for segmentation and a single encoder for classification. Besides, an iterative training strategy is proposed to refine prediction from previous iterations. They have demonstrated that sharing a same feature extractor can leverage the correlation between breast tumor classification and segmentation to enhance the performance of both tasks. However, the predictions of classification and segmentation networks are decoupled, implying the loss of some associated semantic information, which consequently leads to performance gains that are lower than expected. Zhu et al. (2021) propose a soft parameter sharing MTL model (DSI-Net) for joint classification and segmentation of vascular lesions and inflammatory using endoscope images. In order to enable the deep synergistic interaction between multi-task, DSI-Net consists of three independent branches with a lesion location mining module and a category-guided feature generation module. This work stands out for utilizing segmentation-derived lesion location information to enhance classification accuracy, by identifying overlooked lesion areas and correcting misclassified background regions. The limitation of this approach is its focus on differentiating between foreground and the background, which hampers its application in complex multi-task learning scenarios characterized by numerous or tiny segmentation targets.

Table 3
An overview of some state-of-the-art methods for medical image segmentation and classification of tumors.

Zhou et al. (2021)	Task info	3D segmentation and classification of breast tumors in ultrasound images.
	Method	Integrate an encoder–decoder for segmentation and a shared multi-scale encoder for classification, employing an iterative training strategy to refine feature maps.
	Achievement	It demonstrates that sharing a same feature extractor can leverage the correlation between breast tumor classification and segmentation to enhance the performance of both tasks.
	Limitation	The results of the tightly interlinked segmentation and classification tasks are disconnected, resulting in inefficient use of semantic tumor information and reduced accuracy.
Zhu, Chen, and Yuan (2021)	Task info	3D segmentation and classification of inflammatory and vascular lesions in endoscopic images.
	Method	Provide lesion location information for classification by highlighting the lesion areas and suppressing the background based on coarse segmentation.
	Achievement	It demonstrates that interactive training between segmentation and classification tasks can guide classification by mining neglected lesion areas and erasing misclassified background areas.
	Limitation	The model only distinguishes between the foreground and background during segmentation, making it difficult to apply in scenarios with multiple segmentation targets or small segmentation objects.
Wu, Ge et al. (2021)	Task info	3D segmentation and classification of esophageal lesions in endoscopic images.
	Method	Propose a MTL framework (ELNet) that carries out esophageal lesion classification and segmentation via preprocessing, localizing lesions, classifying and segmenting diverse kinds of lesions.
	Achievement	Explore the potential of using multi-task learning for joint segmentation and classification to diagnose esophageal lesions, as this is crucial for facilitating early detection and treatment.
	Limitation	The networks used for highly-coupled segmentation and classification tasks are isolated and lack connectivity at the feature level.
Tang et al. (2023)	Task info	3D segmentation and classification of polyp and cancer in endoscopic images.
	Method	Integrates CNN and Transformer blocks to capture common features and task-specific features for classification and segmentation tasks.
	Achievement	This work shows potential as a valuable tool for endoscopists, particularly when integrated with active learning to address the challenge of limited labeled images in medical imaging.
	Limitation	The computational complexity of the proposed network is relatively high compared to other models, with more floating-point operations (FLOPs) and parameters.

For our interest in esophageal cancer and its associated diagnosis, MTL studies in endoscopic images are more widely studied. For example, Wu, Ge et al. (2021) build a deep convolutional neural network (ELNet) for esophageal lesion classification and segmentation in endoscopic images. ELNet integrates dual-view contextual lesion information to extract global features and local features for MTL via four functional modules in order: Preprocessing, Location, Classification and Segmentation. This study successfully validates the use of multi-task learning for the concurrent segmentation and classification

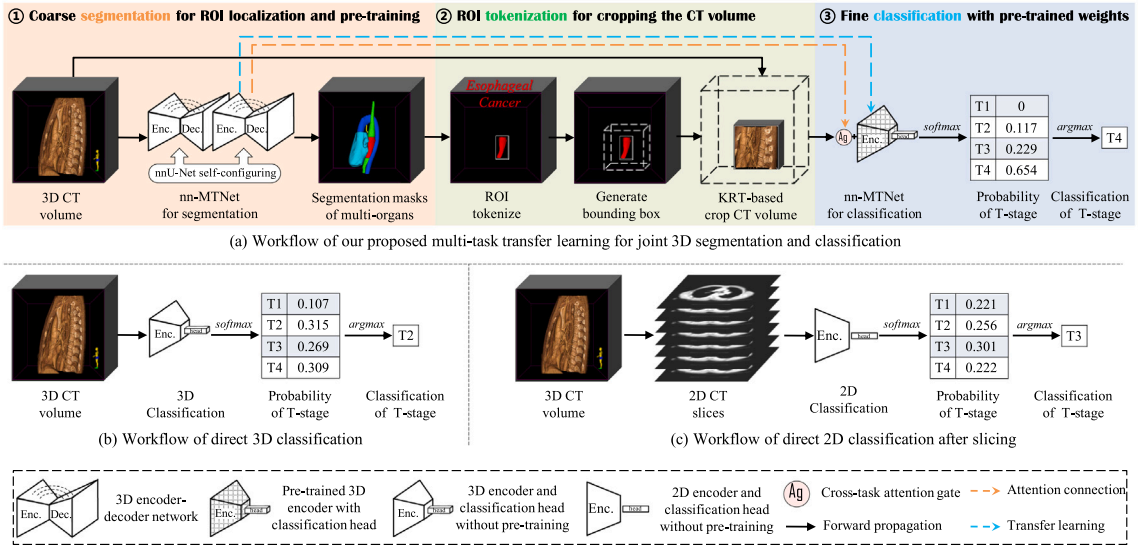


Fig. 2. [Zoom in for more details.] Fig. (a) gives an overview of the workflow of proposed framework (nn-TransEC) for esophageal cancer segmentation and classification. Fig. (b) and Fig. (c) depict two other workflows of direct 3D classification and 2D classification after slicing, respectively.

of EC, which aids in the early diagnosis and treatment of the disease. Nonetheless, it falls short in providing fine-grained TNM-staging of tumor malignancy in endoscopic images, thus lacking sufficient clinical significance. Additionally, there is an absence of feature-level enhancement between the segmentation and classification networks. Tang et al. (2023) introduce the transformer blocks into CNN to obtain a new MTL model (TransMT-Net). TransMT-Net combines global features and local features to improve prediction accuracy in identifying the lesion types and regions. This work shows potential as a valuable tool for endoscopists but with high computational complexity, particularly when integrated with active learning to address the challenge of limited labeled images in medical imaging. Yu et al. (2022) do the similar study, but both of them ignore the fine classification of TNM-stages of esophageal cancer and segmentation of thoracic organs around the esophagus. In contrast, there are fewer MTL studies designed for CT images on esophageal cancer diagnosis. Takeuchi et al. (2021) propose a VGG16-based model for classifying the EC's T-stage from 2D CT slices. Unfortunately, such 2D diagnostic method ignores the utilization of spatial information in 3D CT volume, which inevitably violates the actual diagnostic process for esophageal cancer T-staging.

In summary, inspired by the state-of-the-art work introduced in this section, our study aims to explore the application of multi-task learning to the segmentation of EC-related organs and the T-stage classification of tumors from three distinct perspectives: (1) to build upon the widely recognized nnU-Net framework and optimize the backbone for joint classification and segmentation, (2) to enhance the connections between classification and segmentation networks and predictions by integrating hard parameter sharing and soft parameter sharing multi-task learning, and (3) to achieve high-quality cross-task interaction via attention mechanism and transfer learning.

3. Method

This section describes the proposed nnUNet-based transfer learning framework (**nn-TransEC**) that enables the multi-task learning between segmentation of EC-related organs and classification of EC's T-stages. nn-TransEC focuses on how representations learned from prior segmentation can be utilized to prompt the subsequent fine-grained classification.

3.1. Overview

The proposed nn-TransEC framework builds upon a nnUNet-configured multi-task learning network (nn-MTNet) and works with a knowledge-embedded ROI tokenization method (KRT). nn-MTNet is designed for complementary segmentation and classification, consisting of a 3D encoder–decoder architecture for segmentation and a shared 3D encoder for classification (see Section 3.2 for more details). A novel cross-task attention gate is bridged between the segmentation encoder and classification encoder for highlighting features in ROI. The KRT method utilizes segmentation masks as an intermediate result to crop the CT volume into a patch that is relevant for T-stage diagnosis. Based on that, nn-TransEC takes 3D CT volumes as inputs and produces two outputs in sequence, including a 3D segmentation mask of EC-related organs and a classification probability of EC's T-stages. In summary, nn-TransEC works in three phases for EC diagnosis. Fig. 2(a) depicts the workflow of nn-TransEC, where each phase is represented by a different colored region:

(1) **Coarse segmentation for ROI localization and pre-training**. In this phase, the proposed multi-task learning network (nn-MTNet) receives 3D CT image as input and adopts a cascaded 3D encoder–decoder for coarse segmentation, which is initialized by modified nnU-Net self-configuration strategy (see Section 3.3 for more details). The output is a 3D segmentation mask, containing pixel-wise predictions about five EC-related thoracic organs. Above segmentation mask is then sent to the second phase for tokenization. The well-trained segmentation encoder is transferred to the third phase for better initialization.

(2) **ROI tokenization for cropping the CT volume**. Considering that EC and its adjacent tissues are closely related to the diagnosis of EC's T-stages, a knowledge-embedded ROI tokenization method (KRT) is defined in this phase to obtain the diagnosis-concerned region. The KRT method generates a bounding box for each patient, which is specific in location and size. The bounding box is centered on the geometric center of the EC's segmentation result, with a size equal to the pre-calculated maximum box size. The entire CT volume is then cropped with the bounding box to obtain a patch that is closely related to the diagnosis of EC. The cropped CT patch is sent to next phase for T-stage classification (Section 3.4).

(3) **Fine classification with pre-trained weights and cropped CT patch**. In this phase, the cropped CT patch is input into the classification encoder of nn-MTNet, which is initialized by the well-trained encoder in the first phase. Besides, cross-task attention gates are connected between the segmentation encoder and classification encoder for highlighting features in ROI. A linear classification head with leakyReLU activation is attached as the classification head to predict classification probabilities for EC's T-stages. The category with the highest probability is output as the final T-staging result (Section 3.5).

Above design is motivated by mimicking how a thoracic surgeon would focus on the most significant regions from entire 3D CT volumes for diagnosing EC's T-stages. In addition, Fig. 2 compares nn-TransEC with another two classification workflows of esophageal cancer T-staging, where Fig. 2(b) depicts the directly classification of 3D CT volume and Fig. 2(c) depicts the classification of 2D CT slices. In contrast, our method automatically learns the ROI derived from the priori coarse segmentation without manual intervention.

3.2. nnU-Net-based multi-task transfer learning network design

nnU-Net is a self-configuring deep learning framework proposed by German Cancer Research Center (Isensee et al., 2021) for medical image segmentation. This framework is able to provide key hyper-parameters of neural networks according to the proposed the concepts of data fingerprint, pipeline fingerprint, and heuristic rules.[8]

- **Data fingerprint** represents key attributes of the dataset, such as volume shape, voxel spacing distribution, image modality, number of classes, number of cases, intensity distribution.
- **Pipeline fingerprint** represents key attributes of the training setup, such as the basic architecture, loss function, data augmentation, patch size, batch size, and other hyperparameters. Pipeline fingerprint consists of blueprint, inferred and empirical parameters.
- **Heuristic rules** derive from theoretical knowledge and tuning experiences of the nnU-Net's researchers, who condenses them and then guides the construction of inferred pipeline fingerprint from data fingerprints.

In this way, nnU-Net is able to extract data fingerprints based on the characteristics of the dataset. Subsequently, it infers the corresponding pipeline fingerprints using heuristic rules. Lastly, it follows a four-step learning scheme: (1) pre-processing, (2) training, (3) inference and (4) post-processing. The entire process is automated without human intervention.

For the challenging esophageal cancer diagnosis, spatially accurate identification of the cancer and its adjacent tissues is required. nnU-Net is expected to be a reliable solution due to its outstanding performance in the *Medical Segmentation Decathlon* challenge. However, the backbone of nnU-Net inherently suffers from defects such as insufficient receptive fields and multi-scale information, which makes it difficult to provide robust features for the challenging esophageal cancer recognition, resulting in degraded performance. Besides, the vanilla nnU-Net lacks a unified architecture to jointly train the two related segmentation and classification. At the same time, these defects can hardly be compensated by pre-processing or optimizing the hyper parameters.

Therefore, this study designs a novel multi-task learning network (nn-MTNet) to replace the backbone of the nnU-Net. Fig. 3 provides an overview of the proposed nn-MTNet network. As can be observed, nn-MTNet is an extension of the nnU-Net's backbone,[9] tailored to enhance **cross-task learning** capabilities for esophageal cancer diagnosis. Specifically, three modifications have been implemented to adapt to the requirements of **cross-task learning**, thereby mitigating the limitations inherent in the vanilla nnU-Net.

(1) To enhance the deep information extracted by the encoder, deep supervision-based residual connections are introduced into the backbone. Section 3.2.1 reports more details.
(2) To reinforce the cross-task interaction within our framework, cross-task attention gates have been strategically bridged between the segmentation encoder and the classification encoder. Section 3.2.2 reports more details.
(3) To enhance classification performance through task-related initialization, cross-task transfer learning is employed, facilitating the transfer of knowledge from the segmentation encoder to the classification encoder. Section 3.2.3 reports details.

In addition to modifying the network architecture, this study optimizes the parameter configuration strategy of nnU-Net when applying to esophageal cancer diagnosis, as detailed in Section 3.3.

3.2.1. Deep supervision-based residual backbone

In order to boost deep information extracted by segmentation encoder of nn-MTNet, this study introduces deep supervision from ANU-Net (Li et al., 2020) and residual connections from ResNet (He, Zhang, Ren, & Sun, 2016) into the backbone.

Deep supervision is designed to add auxiliary supervision to intermediate layers of the segmentation encoder. It enables the nn-MTNet to operate in two phases: (1) training phase wherein the outputs from all segmentation branches are jointly calculated the loss, and (2) inference phase wherein the final segmentation output is selected from only the top segmentation branch. Besides, in order to ensure that shallow representations can be passed to deeper layers and loss can be back-propagated to shallow layers, residual connections are also incorporated into the segmentation encoder and the classification encoder.

Given that deep supervision and residual connections are not the primary innovative elements of this paper, this section has been streamlined to present a concise introduction to these concepts. Additional details have been included in Appendix C and Fig.D for readers seeking a more comprehensive understanding.

3.2.2. Cross-task attention mechanism

Given the fact that the segmentation encoder and decoder play a role to classify each pixel in the input CT volume, we believe that the representations embedded in the segmentation encoder contribute to the related classification. Inspired by that, this section proposes a novel **cross-task attention gate** and introduces it into the joint segmentation and classification.

Fig. 4 shows the structure of the proposed cross-task attention gate. Each gate at depth d receives two inputs: (1) the feature maps F^d_{Cls} from the classification encoder, and (2) the feature maps F^d_{Enc} extracted by the segmentation encoder at corresponding depth. Both of them have the same resolution. After that, F^d_{Dec} are redirected to the classification encoder as gating signals through the attention gate to enhance the learning of F^d_{Cls}. Attention coefficient $\alpha \in [0, 1]$ is calculated via the attention gate. Finally, the gate multiples the encoder feature F^d_{Cls} with α pixel by pixel and outputs the results $\hat{F}^d_{Cls}$.

Fig. 3 depicts the role of the proposed attention gate in the multi-task learning network (nn-MTNet), which tightly connects the segmentation encoder and the classification encoder. It is worthy-noting that

[8] This paper inherits nnU-Net's definition of the above concepts, so we pre-define them here for subsequent use. More details are published in the Appendix and the *Nature Methods* publication (Isensee et al., 2021).

[9] Fig. A in the Appendix compares the backbone of nn-MTNet and other U-Net's variants.

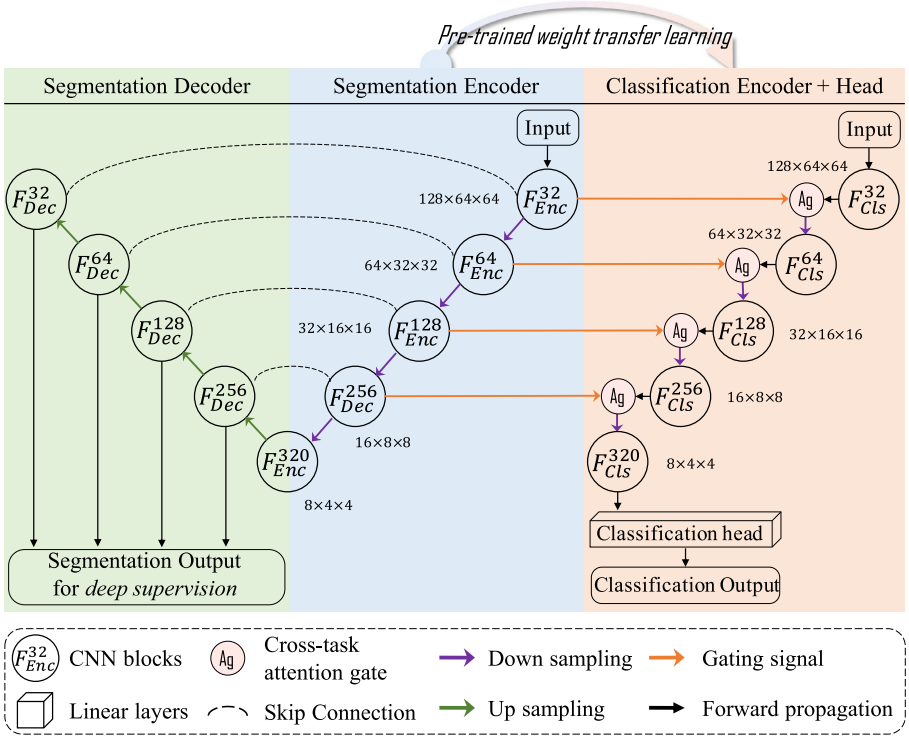


Fig. 3. An overview of the proposed multi-task learning network (nn-MTNet). Detailed structure is shown in Fig. D in Appendix A.

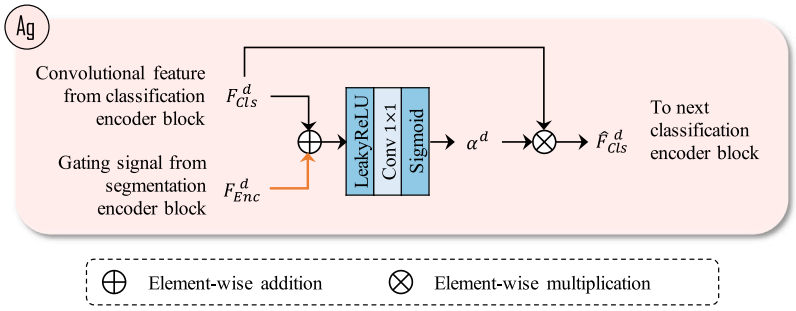


Fig. 4. An overview of the proposed cross-task attention gate with inputs derived from features extracted from classification and segmentation networks, enhancing T-stage classification with the help of multi-organ segmentation.

our proposed cross-task attention gate is different from the previous attention gates, such as Attention U-Net (Oktay et al., 2018; Schlemper et al., 2019), which use the current depth decoder's feature map to refine the skip connection of the upper depth encoder. Both inputs derive from the single segmentation task. In contrast, our proposed gate achieves cross-task connection between segmentation and classification, where the current depth segmentation encoder's feature maps serve as the gating signal for the current depth classification encoder.

3.2.3. Cross-task transfer learning

As observed in Fig. 3 and Fig.D, the proposed segmentation encoder and classification encoder share the same network structure, except for the last layer.[10]

Consequently, it is technically feasible to carry out cross-task transfer learning from segmentation encoder to classification encoder. Specifically, segmentation of EC-related organs serves as a pre-training task while classification of EC's T-stages is regraded as a downstream task. The well-trained segmentation encoder provides the classification encoder with better initialization weights than random initialization, enhancing the classification of EC's T-stages.

3.3. nnUNet-configured coarse segmentation for ROI localization and pre-training

Based on the multi-task learning network (nn-MTNet), this section performs coarse segmentation of five EC-related thoracic organs, i.e., esophageal cancer, aorta, esophagus, lymph and pericardium. Fig. 5 illustrates the coarse segmentation.

First, in order to configure the network parameters of nn-MTNet, this section modifies the vanilla parameter configuration strategy of nnU-Net (Isensee et al., 2021). As shown in the top green block in Fig. 5, the data fingerprint is extracted according to the dataset. The pipeline fingerprint is calculated from data fingerprint and pre-defined heuristic rules. Apart from that, there are three modifications made for optimizing the nnUNet-based self-configuration strategy:

- Heuristic rules: nnU-Net proposes to condense theoretical knowledge and tuning experiences of the researchers into a set of

[10] The classification encoder additionally adds a linear classification head in the last layer for obtaining the predicted probabilities of four T-stages.

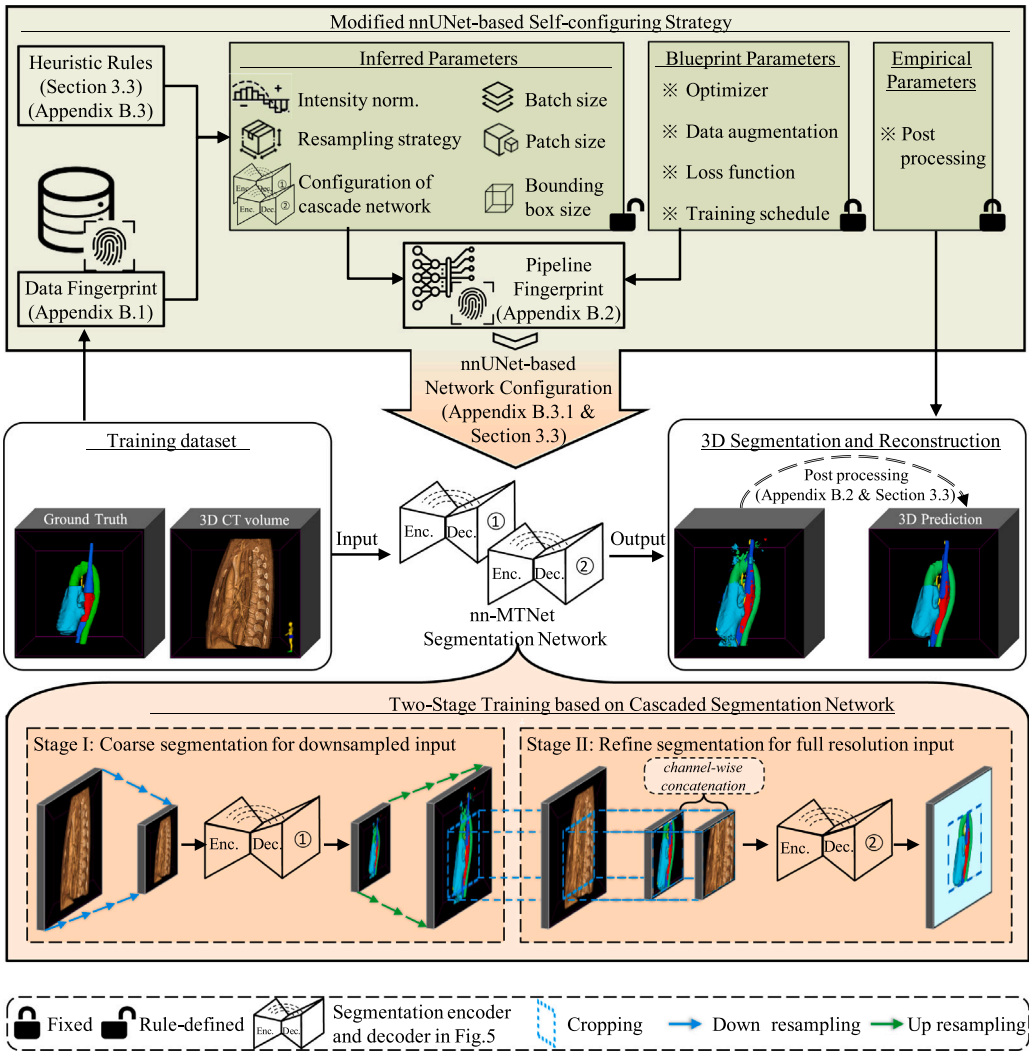


Fig. 5. Illustration of the coarse segmentation of thoracic multi-organs. The segmentation network is configured by the modified nnUNet-based self-configuration strategy.

heuristic rules. These rules guide the construction of various inter-dependent parameter mappings to infer partial pipeline parameters required for training from the data fingerprints. Fig. 6 reports these mappings relationship. Our study updates heuristic rules with the **maximum bounding box size** and **loss parameters**, which are detailed as follows:

(1) the data fingerprint is updated with new-added shape attribute of the ROI's annotation. Based on that, we statistically count the EC's labeled shapes of all cases in the training dataset and calculate their greatest common divisors $< H, W, D >$ in the three dimensions as the **maximum bounding box size**, which plays a key role in Section 3.4 for cropping the 3D CT volume.

(2) the amount of patients with different T-stages in the dataset is counted for designing the hyper-parameter α_2 of the classification loss function, which is formulated in Eq. (4). More details are reported in Appendix B.3.

- Segmentation loss function: considering the imbalance between the targets and background, this study takes advantages of pixel-wise multi-class focal loss and integrates the segmentation predictions at four depth $\{\hat{Y}^k, k \in \{1,2,3,4\}\}$ into joint loss calculation, which is formulated as follows:

$$\mathcal{L}oss_{seg}(Y,\hat{Y}) = \frac{\sum_{c=1}^{C}\sum_{k=1}^{4}\left(-\alpha_1(1-\hat{Y}_c^k)^{\gamma_1}Y_c log(\hat{Y}_c^k)\right)}{-4}, \quad (1)$$

where $\alpha_1 \in [0,1]$ is a vector of class weights, $\gamma_1 \geq 0$ is a scaling coefficient. This study empirically set $\alpha_1 = 0.25$ and $\gamma_1 = 2$ according to Lin, Goyal, Girshick, He, and Dollár (2017). Y_c denotes the ground truth of class c, $\hat{Y}_c^k$ indicates the probability predictions of class c from depth k, and C indicates the number of classes. In this loss function, deep supervision assist to regularize the calculation of the loss at multiple semantic levels and back propagation of the error from various depths. Fig. B depicts above process.

- Post-processing: this study utilizes the medical a priori knowledge in the post-processing, i.e., most thoracic organs usually have only one subject. Guided by it, the post-processing utilizes the connected component analysis before testing. The connected components are obtained by calculating the regions that have the same label and are adjacent to each other in all the foreground regions. Filtering the connected components helps to remove the spurious false positive predictions. More details are reported in Appendix B.2.
- Activation functions: this study replaces the ReLU activation functions in the vanilla U-Net with Leaky ReLU. Leaky ReLU adds a small slope (0.01 in this study) for negative inputs to protect that neuron from dying and allowing training to continue. Eq. (2) gives the formulation, where $\lambda = 0.99$.

$$Leaky\ \text{ReLU}(x) = x - \lambda \cdot min(0,x) \quad (2)$$

- Normalization: Original batch normalization struggle to evaluate the distribution of the mean and standard deviation of the training data. Previous studies (Casella et al., 2021; Huang & Belongie,

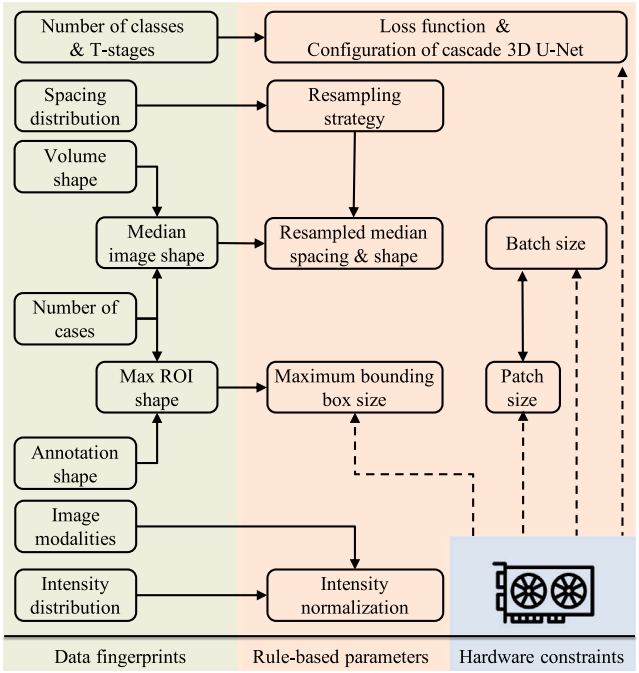


Fig. 6. Heuristic rules mapping functions.

2017; Shen et al., 2020) have demonstrated that instance normalization (Ulyanov, Vedaldi, & Lempitsky, 2016) outperforms batch normalization in tasks such as image segmentation with small batchsize. Therefore, this study replaces the batch normalization in the vanilla U-Net with instance normalization.

Second, after configuration, the segmentation encoder–decoder of nn-MTNet are cascaded sequentially, which receives 3D CT volume X as input and outputs pixel-wise 3D prediction $\hat{Y}$ about five related organs. Fig. 5 illustrates the two-stage training of the cascaded segmentation network. As can be observed, the first stage's segmentation encoder–decoder receives the downsampled CT volume and outputs the segmentation map at low resolution. Then, the segmentation map is upsampled to full resolution and sent to the next stage. In the second stage, we channel-wisely concatenate the segmentation maps to the original CT volume as input, and then sent it to the second stage's segmentation encoder–decoder for training at full resolution. The output of the decoder is refined as the segmentation mask. Fig. 2 illustrates above process in pink. Note that the cross-task attention gate (Section 3.2.2) and transfer learning (Section 3.2.3) are only applied to the second stage's segmentation encoder–decoder.

Finally, the pixel-wise prediction $\hat{Y}$ is the segmentation output after post-processing, which provides the location of EC-related organs. $\hat{Y}$ is able to assist volume quantization via 3D reconstruction (Section 4.4) and sent to the second phase (see Section 3.4 for more details). The well-trained encoder of the cascade network is transferred to the third phase for better initialization (see Section 3.5 for more details).

3.4. Knowledge-embedded ROI tokenization for cropping the 3D CT volume

Motivated by diagnostic workflow of doctors for classifying EC's T-stage, the second phase defines a knowledge-embedded ROI tokenization method (KRT). Specifically, the KRT method works in three steps to separate the most concerned CT patch with the help of prior multi-organ segmentation results $\hat{Y}$.

(1) In order to take advantage of the ROI localization provided by coarse segmentation, KRT outlines the bounding box of each EC-related organ, including esophageal cancer, aorta, esophagus, lymph, and pericardium.
(2) In order to capture the patch that related to diagnosis of T-stage, KRT constructs a cropping bounding box that centered on the geometric center of the EC's bounding box. The size of KRT's cropping bounding box is set to the pre-calculated maximum box size $< H, W, D >$, which is defined as inferred parameters via heuristic rules in Section 3.3.
(3) According to the KRT's cropping bounding box, the entire 3D CT volume X is cropped to get a patch $\hat{X}$ that is closely related to the diagnosis of EC. Eq. (3) formulates the KRT-based cropping operation:

$$\hat{X} = 3D\_Crop(X, Center(\hat{Y}), \langle H, W, D \rangle), \tag{3}$$

where $3D\_Crop(\cdot)$ is a center crop function that crops the 3D volume X to the cropping bounding box size$< H, W, D >$, $Center(\cdot)$ denotes the geometric center of EC in the segmentation prediction $\hat{Y}$.

Above design is motivated by how doctors would transfer their knowledge of ROI learned from prior tasks to a subsequent task, the KRT's cropping bounding box is able to capture the EC and its surrounding tissues since EC is the main concern and its surrounding tissues are also involved in diagnosis of EC's T-stage. Besides, the cropping bounding box is specialized for each patient and different in terms of location and size.

Finally, the cropped CT patch $\hat{X}$ is sent to the third phase for the subsequent fine-grained classification of EC's T-stages.

3.5. Fine classification with pre-trained weights and cropped CT patch

After completing the model pre-training in Section 3.3, this section classifies the EC's T-stages from cropped CT patch. Fig. 2 illustrates this phase in blue.

As shown in Fig. D in the Appendix, the classification network of nn-MTNet consists of a classification encoder and a classification head. The classification encoder utilizes cross-task attention gates at each depth for enhancing the semantic information from the segmentation encoder, which is detailed in Section 3.2.2. The classification head is composed of a fully-connected layer, a LeakyReLU layer and a fully-connected layer, where the output channel of the last fully-connected layer is set as Appendix B.3. The yellow box in Fig. 3 shows the specific architecture of the classification network.

After finishing coarse segmentation, the weights of well-trained segmentation encoder embed diverse textural feature, which then continues to play a role in transfer learning. Specifically, these weights serve as initialization for the classification encoder, which is benefiting to the fine-grained classification of EC's T-stages. Fig. 3 illustrates the transfer learning between segmentation and classification.

Next, classification network receives the cropped CT patch $\hat{X} \in R^{H \times W \times D}$ from Section 3.4 as an input and outputs the predicted probability ($pred$) about four T-stages.

Finally, the manual-labeled T-stage labels are encoded into one-hot vectors, which are sent to the classification loss function ($\mathcal{L}oss_{cls}$) for model training and optimization.

$$\mathcal{L}oss_{cls} = -\sum_{t=1}^{4} \alpha_2{}^t (1 - pred^t)^{\gamma_2} \log(pred^t), \tag{4}$$

where $pred^t$ is the predicted probability of T-stage t ($t \in \{1, 2, 3, 4\}$). This study empirically sets the scaling coefficient $\gamma_2 = 2$ according to Lin et al. (2017). $\alpha_2^t \in [0, 1)$ is a weight parameter for T-stage t. We argue that smaller sample amount makes it harder to learn this category. Therefore, α_2 is negatively correlated with the percentages of patients with different T-stages as a way to enhance learning from hard samples,

which are counted via heuristic rules in Section 3.3. Specifically, α_2 is set to [0.20, 0.16, 0.10, 0.54] in this study.

During inference, the classification network receives the cropped CT patch $\hat{X}$ and outputs the highest probability $T = argmax(pred)$ as the prediction of EC's T-stages.

4. Experiments

Extensive experiments were conducted on the diagnosis of esophageal cancer: (1) to evaluate the performance of nn-TransEC in thoracic multi-organ coarse segmentation and compare it with the state-of-the-art counterparts (Section 4.2.1), (2) to evaluate the performance of nn-TransEC in fine classification of EC's T-stages and compare it with the state-of-the-art counterparts (Section 4.2.2), (3) to compare nn-TransEC with the state-of-the-art multi-task learning frameworks (Section 4.2.3), (4) to validate the effectiveness of nn-TransEC via ablation studies based on classification and segmentation (Section 4.3), (5) to visualize the segmentation predictions in four views and quantify the discrepancy between the predicted volumes (Section 4.4), (6) to estimate the performance improvement with the proposed ROI tokenization for cropping CT volume as input (Section 4.5), and (7) to explore the model interpretation for fine classification of EC's T-stages (Section 4.6).

4.1. Datasets and experiment settings

This subsection described the experimental protocols, including datasets acquisition and collection, annotation and pre-processing, evaluation metrics, and experiment settings.

4.1.1. Datasets acquisition and collection

In this study, private datasets were collected from patients who underwent radical esophageal cancer surgery at the Department of Thoracic Surgery of the First Affiliated Hospital of the Army Medical University and Shanxi Provincial Cancer Hospital during the period of January 2018 to April 2022.

All patients were confirmed to have esophageal cancer by postoperative histopathology and signed an informed consent form and received breathing exercises in advance. Image sampling was performed via a Siemens 64-bit multi-detector spiral CT scanner, with scanning coverage of the chest or from the neck to the upper abdomen, and the layer thickness of the CT image was less than 2 mm.

We designed three inclusion criteria: (i) esophageal squamous cell carcinoma and esophageal adenocarcinoma were confirmed by postoperative histopathology, (ii) undergone a preoperative thin-layer enhanced CT scan (layer thickness <2 mm) with an identifiable tumor lesion > mm in diameter. (iii) complete and accurate clinical-pathologic information was available. Besides, exclusion criteria were established: (i) poor CT volume quality, e.g., artifacts that significantly interfered with segmentation or diagnosis, (ii) incomplete CT volume, e.g., lack of neck or gastroesophageal conjunction image, and (iii) pathologic types of esophageal cancer other than adenocarcinoma or squamous carcinoma.

Based on the above inclusion and exclusion criteria, this study collected a total of 169 patients with esophageal cancer, including their DICOM data and tumor T-stage labels. For the convenience of reading, the above dataset was named **EC169**[11] in this paper. Table 4 reported the detailed data-fingerprint of EC169.

[11] The study on EC169 dataset was approved by the Ethics Committee of the First Affiliated Hospital of Army Military Medical University [(B)KY2021165] and Shanxi Provincial Cancer Hospital (2021050).

Table 4
Data fingerprint generated by nnU-Net for the EC169 dataset.

median space distribution	[1.0, 0.73, 0.73]
median volume shape	237 × 256 × 256
number of cases	169
number of modalities	["CT"]
intensity distribution	Range [−1024, 3071]
number of segmentation classes	[0:"Background", 1:"Esophageal cancer", 2:"Aorta", 3:"Esophagus", 4:"Lymph", 5:"Pericardium"]
number of classification T-stages	["T1", "T2", "T3", "T4"]

4.1.2. Annotation and pre-processing

The **annotation** process was strictly double-blind, completed by one radiologist and reviewed by another senior radiologist, who manually labeled EC-related organs on the thin-layer enhanced CT images in DICOM format using *Amira* software (version 6.0.0). For EC, the contours were drawn around the tumor volume of interest, thus assisting the annotator in accurately identifying the boundaries of the corresponding structures. Afterwards, using the SEGMENTATION module of *Amira*, radiologists manually outlined the tumor area along the edge of the tumor slice by slice. Four more tissues surrounding the tumor including esophagus, pericardium, aorta, and lymph nodes were annotated for segmentation training in the same way. Besides, tumor T-stage labels from postoperative histopathology were used for fine classification of EC's T-stages.

After the aforementioned preprocessing, this section also presented a statistical analysis of the distribution of various categories within the **EC169** dataset. This included the proportion of annotations for five EC-related thoracic organs and the distribution of four T-stages of EC. Above statistical results were depicted in Fig. 7. As illustrated, the esophageal cancer diagnosis task based on the EC169 dataset exhibited significant class imbalance. Specifically, the largest organ constituted 77.63% annotations, while the smallest accounted for a mere 0.27%. In the distribution of T-stage classifications across cases, the T-stage with the largest number of cases accounted for 54.44% of the cases, while the smallest accounted for only 10.06%.

After that, we performed **pre-processing** for all datasets. Since the CT image data came from two different hospitals, some uncontrollable factors such as room temperature, humidity and patient's body position during the acquisition of the images resulted in different image intensities, which were likely to affect the recognition of the tumor edge of EC and further affect the accuracy of segmentation. Therefore, we performed intensity normalization and image quantization on the segmented region through gray scale range selection to reduce the error caused by the change of image intensity. Considering the existence of different sampling spacings of slice in the EC169 dataset, in order to make the network learn the spatial semantic information better, the images and mask annotations of all the patients were resampled to the same spacing, i.e., this paper adopted the median of the voxel spacing of the EC169 dataset. Then third order spline interpolation was used for the CT volume, while nearest-neighbor interpolation was used for the segmentation mask. After resampling, the median image shape was 237 × 256 × 256.

This paper adopted the extensive **data augmentations** provided by nnU-Net (Isensee et al., 2021) such as rotation, scaling, Gaussian noise, Gaussian blurring, brightness adjustment, contrast adjustment, Gamma adjustment, multi-angle mirroring, and so on.

4.1.3. Evaluation metrics

This study used *3D Dice*, *IoU*, HD_{95} (the 95th percentile of Hausdorff Distance to eliminate the influence of a small subset of outliers) and *ASSD* (Average symmetric surface distance) as segmentation evaluation metrics. *AUC* (area under the receiver operating characteristic curve), *precision*, *sensitivity*, *specificity*, *NPV* (negative predictive value) and *F1-score* were classification evaluation metrics.

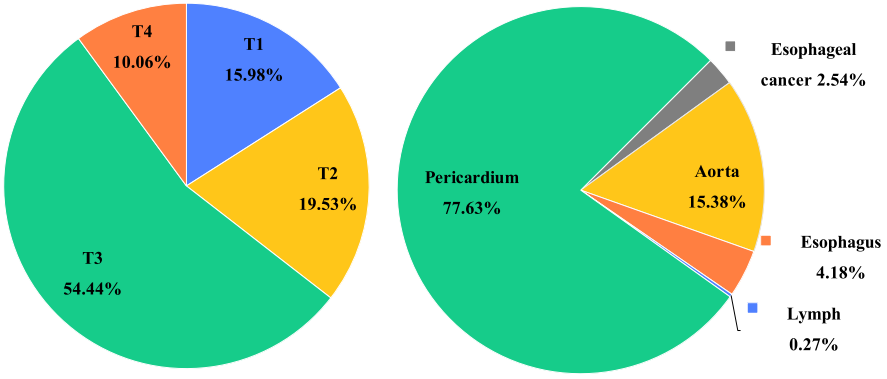


Fig. 7. Statistical assessment of class imbalance in the EC169 dataset for esophageal cancer diagnosis.

Table 5
Evaluation Metrics. Y and $\hat{Y}$ denote the sets of surface vertices of ground truth and segmentation prediction, respectively. $|Y|$ and $|\hat{Y}|$ denote the total numbers of vertices in those two sets. TP is true positive, TN is true negative, FP is false positive, FN is false negative. $\|a-b\|$ is the Euclidean distance between a and b.

3D Dice	$2 \times \mathrm{TP}/(2 \times \mathrm{TP+FP+FN})$
IoU	TP/(TP+FP+FN)
ASSD	$\frac{\sum_{a\in Y} \min_{b\in\hat{Y}} \|a-b\| + \sum_{b\in\hat{Y}} \min_{a\in Y} \|b-a\|}{\|Y\| + \|\hat{Y}\|}$
HD_{95}	$\max\left\{ \max_{a\in Y} \min_{b\in\hat{Y}} \|a-b\|, \max_{b\in Y} \min_{a\in\hat{Y}} \|a-b\| \right\}$

(a) Segmentation evaluation metrics.

AUC	Area Under the ROC Curve
Precision	TP/(TP+FP)
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
NPV	TN/(TN+FN)
F1-score	1/Precision + 1/Sensitivity

(b) Classification evaluation metrics.

Table 5 formulated these metrics.

In addition, this paper evaluated the correlation between segmentation results and ground truth with the *Pearson correlation coefficient*, which further quantified the volumes of multi-organ segmentation. Pearson's r is a standardized coefficient that varies between −1 (perfect negative correlation) and +1 (perfect positive correlation). When the value is 0, it means that there is no any linear correlation between variants. Researchers generally agreed that when the absolute value of $r \geq 0.8$ and p-value <0.05, it testified a strong and satisfactory correlation with statistically significance (Profillidis & Botzoris, 2019). In this study, r was formulated as follows.

$$r = \frac{\sum_{i=1}^{N} \left[V_{ol}(Y_i) - \overline{V_{ol}(Y)} \right] \times \left[V_{ol}(\hat{Y}_i) - \overline{V_{ol}(\hat{Y})} \right]}{\sqrt{\sum_{i=1}^{N} \left(V_{ol}(Y_i) - \overline{V_{ol}(Y)} \right)^2} \sqrt{\sum_{i=1}^{N} \left(V_{ol}(\hat{Y}_i) - \overline{V_{ol}(\hat{Y})} \right)^2}}, \tag{5}$$

where $V_{ol}(\cdot)$ was the function to calculate the volume, $\overline{V_{ol}(\hat{Y})}$ denoted the average volume of prediction and $\overline{V_{ol}(Y)}$ denoted the average volume of ground truth.

4.1.4. Experiment protocol

For the coarse segmentation of EC-related organs, this study randomly selected 100 cases from the EC169 dataset after pre-processing, denoted as **EC100**. EC100 was then divided into training dataset and test dataset with the split ratio 4:1.[12] The number of training epochs was 100. Limited by the hardware constraints, this paper set small batch size (2) for 3D volume segmentation.

For the fine classification of EC's T-stage, this paper used the entire **EC169** as the training and testing dataset, employing a five-fold cross-validation to evaluate the generalization ability and prediction performance of nn-TransEC and other comparative methods. The batch size and number of epochs were 12, 1000, respectively.

For the proposed ROI tokenization for cropping the 3D CT volume, this study set the number of segmentation tokens as same as the number of classes (6) where only the bounding box of esophageal cancer involved in the proposed KRT methods. The cropping bounding box was 128 × 64 × 64 pre-defined by heuristic rules.

In order to validate the superior of the proposed nn-TransEC, Sections 4.2.1 and 4.2.2 conducted two comparison experiments on coarse segmentation of EC-related organs and fine classification of EC's T-stages. State-of-the-art 3D models were involved and implemented in the same experiment protocol for fair comparison with nn-TransEC, including nnFormer (Zhou et al., 2023), Swin UNETR (Hatamizadeh et al., 2021), UNETR (Hatamizadeh, Tang, et al., 2022), nnU-Net (Isensee et al., 2021), TransUNet (Chen et al., 2021), UNet++ (Zhou et al., 2019), SegResNet (Myronenko, 2019), Dense-UNet (Li et al., 2018), and V-Net (Milletari, Navab, & Ahmadi, 2016a).

In order to assess multi-task learning capabilities of the proposed framework, Section 4.2.3 re-implemented various state-of-the-art multi-task learning counterparts, including ELNet (Wu, Ge et al., 2021), DSI-Net (Zhu et al., 2021), Zhou et al. (2021) and TransMT-Net (Tang et al., 2023). Above framework were designed for medical image analysis and all tailored for cancer segmentation and classification. They were re-implemented under the same experimental setup and then included in the comparison with nnTrans-EC.

Furthermore, Section 4.3 presented comprehensive ablation studies to evaluate the individual contributions of the components within the proposed nn-TransEC framework. The focus of these studies was on the segmentation of EC-related organs and classification of EC's T-stages. The segmentation network comprised four distinct modules, while the classification network was structured into five integral components. In order to validate the generalization capability of the proposed KRT method for cropping 3D CT volumes as classification input, Section 4.5 conducted experiments to evaluate the performance of state-of-the-art 3D models before and after ROI tokenization. Through a systematic process of modular and component removal, these sections endeavored to quantify the specific impact of each component on the overall performance of the nn-TransEC framework.

The run-time infrastructure for this study was mainly formed by PyTorch 1.7.1 with CUDA 12.2 over NVIDIA TITAN X Pascal GPU. The

[12] Given that the objective of coarse segmentation was to delineate the ROI related to esophageal cancer diagnosis and to facilitate model pre-training for subsequent fine classification tasks, the validation dataset was not divided, so the segmentation model was not fine-tuned.

Table 6
The topology of the cascaded segmentation network of nn-MTNet. Below parameters are generated based on the designed heuristic rules. The 3D U-Net low-resolution refers to the first stage of the cascade network. The 3D U-Net full-resolution refers to the second stage.

Rule-based parameters	Low-resolution U-Net	Full-resolution U-Net
Resampled median image shape	118 × 128 × 128	237 × 256 × 256
Patch size	128 × 64 × 64	128 × 64 × 64
Batch size	2	2
Network depth	5	5
Number of down-sampling per axis	[4, 4, 4]	[4, 4, 4]
Down sampling strides	[[2, 2, 2], [2, 2, 2], [2, 2, 2], [2, 2, 2]]	
Convolution kernel sizes	[[3, 3, 3], [3, 3, 3], [3, 3, 3], [3, 3, 3], [3, 3, 3]]	

Adam optimizer (Kingma & Ba, 2015) and cosine learning rate decay schedulers with initial learning rate $1e^{-3}$ were applied.

Table 6 released network typology of our proposed cascaded segmentation network in Section 3.3, which was automatically generated by the modified nnUNet-based self-configuration strategy.

4.2. Comparison with the state-of-the-art models

This study conducted two comparison experiments on coarse segmentation of EC-related organs and fine classification of EC's T-stages. State-of-the-art 3D models were involved to validate the superior of our proposed model, including nnFormer (Zhou et al., 2023), Swin UNETR (Hatamizadeh et al., 2021), UNETR (Hatamizadeh, Tang, et al., 2022), nnU-Net (Isensee et al., 2021), TransUNet (Chen et al., 2021), UNet++ (Zhou et al., 2019), SegResNet (Myronenko, 2019), Dense-UNet (Li et al., 2018), and V-Net (Milletari et al., 2016a). 3D U-Net (Çiçek, Abdulkadir, Lienkamp, Brox, & Ronneberger, 2016) was implemented as the baseline model. Besides, some multi-task learning methods were also compared.

4.2.1. Results of coarse segmentation

Table 7 reported the results of coarse segmentation of EC-related organs on the EC100 dataset. Considering that the purpose of coarse segmentation was to localize ROI about esophageal cancer diagnosis and perform model pre-training for subsequent fine classification, the segmentation task was not fine-tuned. Even so, results indicated that the proposed nn-TransEC still reached the state-of-the-art among the existing methods when segmenting the concerned esophageal cancer. nn-TransEC achieved a *3D Dice* improvement of 2.55 percentage points over the second-best method (nnU-net), improved IoU from 0.649 to 0.654, reduced $ASSD$ from 3.452 to 1.402.

Besides, almost all SOTA methods performed poorly in the segmentation of EC-related small organs, such as esophageal cancer and lymph, which reflected the weakness of existing methods. However, it was noteworthy that our proposed method achieved the top performance in the segmentation of these organs in terms of 3D Dice. The performance gaps between nn-TransEC and the second-place method were statistically significant (p-value <0.05), demonstrating the superiority of our proposed framework.

Above results demonstrated the advance of nn-TransEC for promising performance in segmentation of EC-related organs, especially for our concerned EC. The quantization analysis about segmentation was reported in Section 4.4. Additionally, Section 5.2 delved into an analysis from a methodological perspective, discussing the underlying reasons for the performance discrepancies between the proposed nn-TransEC and other state-of-the-art methods.

4.2.2. Results of fine classification

The section implemented encoders of aforementioned state-of-the-art 3D models and attached the same classification head.

Table 8 reported the fine classification results on the EC169 dataset. Results indicated that the proposed nn-TransEC still reached the state-of-the-art among the existing methods in the classification of EC's T-stages task, which indicated that:

- In all listed metrics, nn-TransEC achieved the best average performance from T1 to T4 classification, which reflected the superiority of nn-TransEC in classifying EC's T-stages. Specifically, nn-TransEC improved 1.51 percentage points over the second-best method (nnFormer) on AUC, improved 9.74 percentage points over the second-best method (Swin UNETR) on *precision*, improved 3.37 percentage points over the second-best method (nnFormer) on *sensitivity*, improved 1.94 percentage points over the second-best method (nnFormer) on *specificity*, improved 2.62 percentage points over the second-best method (nnFormer) on *NPV*, and improved 10.47 percentage points over the second-best method (Swin UNETR) on *F1*.
- It was noting worthy that most SOTA models suffered performance degradation in the most challenging classification on **T3** stage. In contrast, our proposed nn-TransEC still achieved the first place performance and the gap with the second-place method was significant. Specifically, nn-TransEC improved 6.65 percentage points over the second-best method (nnFormer) on *AUC*, improved 11.95 percentage points over the second-best method (DenseUNet) on *sensitivity*, improved 9.67 percentage points over the second-best method (DenseUNet) on *NPV*, improved 7.48 percentage points over the second-best method (DenseUNet) on *F1*. Above improvements indicated the superiority and robustness of our proposed method in classifying challenging esophageal cancer T-stages.

  Table A in the Appendix released the detailed results.

Fig. 8 compared the confusion matrix of our proposed nn-TransEC and other SOTA methods on the classification of EC's T-stages. Fig. 9 depicted the ROC curves of the classification results of the above nine methods from T1 to T4. As can be observed, our proposed nn-TransEC achieved best AUC in most classification tasks, especially in the most challenging T3 classification task.

Additionally, Section 5.2 delved into an analysis from a methodological perspective, discussing the underlying reasons for the performance discrepancies between the proposed nn-TransEC and other state-of-the-art methods.

4.2.3. Comparison with the multi-task learning methods

In order to evaluate the superiority of nn-TransEC on multi-task learning, this section compared it with other state-of-the-art multi-task learning models, including ELNet (Wu, Ge et al., 2021), DSI-Net (Zhu et al., 2021), Zhou et al. (2021) and TransMT-Net (Tang et al., 2023). Above models were designed for medical images and trained on the same dataset for fair comparison.

Table 9 summarized the experimental results in joint classification and segmentation. The results demonstrated that our proposed nn-TransEC achieved the best performance in all segmentation and classification metrics compared to listed models. Specifically, in the esophageal cancer segmentation task, our proposed model outperformed the second-best counterpart (Zhou et al. (2021) and TransMT-Net (Tang et al., 2023)) by 8.07, 30.39, 10.31, and 60.32 percentage points in 3D Dice, HD95, IoU, and ASD metrics, respectively. Meanwhile, in the classification task, our proposed model outperformed the second-best counterpart (TransMT-Net (Tang et al., 2023)) by 3.75, 15.75, 7.49, 3.05, 3.42 and 13.37 percentage points in AUC, Precision, Sensitivity, Specificity, NPV and F1 metrics, respectively. Above improvements indicated that nn-TransEC could be considered as a promising multi-task learning model in both segmentation and classification.

Table 7
Comparison with existing state-of-the-art 3D segmentation models on the EC100 dataset for multi-organ segmentation, where esophageal cancer (EC) is our main concern. Results style: **best**, second-best.

3D segmentation models	3D Dice↑					HD95↓				
	EC	Aorta	Esophagus	Lymph	Pericardium	EC	Aorta	Esophagus	Lymph	Pericardium
3D U-Net (Çiçek et al., 2016)	0.520	0.884	0.573	0.362	0.937	65.988	15.814	18.25	46.783	15.586
V-Net (Milletari et al., 2016a)	0.598	0.934	0.585	0.404	0.947	42.553	2.239	20.836	31.235	3.865
DenseUNet (Li et al., 2018)	0.739	0.955	0.745	0.475	0.967	18.314	1.118	13.034	22.583	2.720
SegResNet (Myronenko, 2019)	0.747	0.955	0.758	0.491	0.967	17.888	1.133	14.130	23.629	2.820
UNet++ (Zhou et al., 2019)	0.741	0.955	0.758	0.498	0.967	17.374	**1.111**	10.415	22.630	2.832
TransUNet (Chen et al., 2021)	0.749	**0.956**	0.758	0.503	0.968	17.025	1.130	10.376	24.861	2.752
UNETR (Hatamizadeh, Tang, et al., 2022)	0.785	0.951	0.794	0.483	**0.970**	13.488	1.416	8.196	14.377	2.759
Swin UNETR (Hatamizadeh et al., 2021)	0.813	0.951	**0.821**	0.550	0.967	18.126	1.175	9.787	19.743	**2.655**
nnU-Net (Isensee et al., 2021)	0.822	0.951	0.817	0.568	0.969	11.991	1.451	10.245	13.710	4.282
nnFormer (Zhou et al., 2023)	0.814	0.950	0.805	0.523	0.969	**11.112**	1.288	**5.569**	**11.753**	2.742
nn-MTNet (Ours)	**0.843**	0.955	0.811	**0.605**	0.969	11.773	1.340	10.304	14.880	2.900

3D segmentation models	IoU ↑					ASSD↓				
	EC	Aorta	Esophagus	Lymph	Pericardium	EC	Aorta	Esophagus	Lymph	Pericardium
3D U-Net (Çiçek et al., 2016)	0.338	0.790	0.437	0.234	0.881	23.545	4.618	5.066	15.379	4.696
V-Net (Milletari et al., 2016a)	0.393	0.868	0.456	0.270	0.899	16.484	0.924	5.992	7.748	1.116
DenseUNet (Li et al., 2018)	0.573	0.910	0.657	0.331	0.936	7.535	0.375	2.573	5.566	0.635
SegResNet (Myronenko, 2019)	0.568	0.909	0.664	0.344	0.935	7.110	0.375	2.371	5.707	0.64
UNet++ (Zhou et al., 2019)	0.566	0.909	0.669	0.353	0.935	6.799	0.386	2.068	6.235	0.632
TransUNet (Chen et al., 2021)	0.559	0.911	0.666	0.356	0.937	6.925	0.380	1.793	6.153	0.628
UNETR (Hatamizadeh, Tang, et al., 2022)	0.623	0.904	0.678	0.338	**0.942**	4.020	0.486	1.409	4.111	**0.558**
Swin UNETR (Hatamizadeh et al., 2021)	0.545	0.908	0.689	0.331	0.940	2.049	**0.370**	2.421	3.856	0.623
nnU-Net (Isensee et al., 2021)	**0.659**	0.902	0.666	**0.395**	0.936	1.952	0.430	2.967	**2.295**	0.663
nnFormer (Zhou et al., 2023)	0.649	0.903	**0.692**	0.358	0.940	3.452	0.471	**1.056**	3.544	0.584
nn-MTNet (Ours)	0.654	**0.912**	0.674	0.364	0.938	**1.402**	0.374	2.099	5.072	0.627

Table 8
Performance of the state-of-the-art 3D models on the EC169 dataset for classification of EC's T-stages. All reported results are the average values derived from *5-fold cross-validation* using the same experimental protocol. The 95% confidence interval (CI) for AUC is reported in parentheses. Results style: **best**, second-best.

3D models	▷ AUC (95% CI) over T1–T4 (n = 169)				
	T1	T2	T3	T4	Average
nnU-Net (Isensee et al., 2021)	0.954 (0.898 - 1.000)	**0.933 (0.872 - 0.993)**	0.798 (0.731 - 0.864)	0.932 (0.849 - 1.000)	0.904 (0.873 - 0.936)
SegResNet (Myronenko, 2019)	0.885 (0.801 - 0.970)	0.879 (0.801 - 0.958)	0.672 (0.592 - 0.753)	0.877 (0.768 - 0.985)	0.829 (0.788 - 0.869)
DenseUNet (Li et al., 2018)	0.814 (0.772 - 0.855)	0.875 (0.788 - 0.962)	0.862 (0.780 - 0.945)	0.667 (0.587 - 0.748)	0.814 (0.772 - 0.855)
UNet++ (Zhou et al., 2019)	0.889 (0.806 - 0.972)	0.897 (0.824 - 0.970)	0.754 (0.682 - 0.826)	0.868 (0.757 - 0.980)	0.852 (0.814 - 0.890)
TransUNet (Chen et al., 2021)	0.837 (0.740 - 0.934)	0.912 (0.843 - 0.980)	0.733 (0.659 - 0.808)	0.865 (0.753 - 0.978)	0.837 (0.797 - 0.876)
UNETR (Hatamizadeh, Tang, et al., 2022)	0.883 (0.798 - 0.968)	0.905 (0.835 - 0.976)	0.764 (0.694 - 0.835)	0.938 (0.858 - 1.000)	0.873 (0.837 - 0.908)
Swin UNETR (Hatamizadeh et al., 2021)	0.939 (0.875 - 1.000)	0.913 (0.845 - 0.981)	0.805 (0.740 - 0.870)	0.925 (0.837 - 1.000)	0.895 (0.863 - 0.928)
nnFormer (Zhou et al., 2023)	**0.989 (0.960 - 1.000)**	0.913 (0.845 - 0.981)	0.842 (0.784 - 0.901)	0.962 (0.898 - 1.000)	0.927 (0.899 - 0.954)
nn-TransEC (Ours)	0.958 (0.905 - 1.000)	0.925 (0.862 - 0.989)	**0.898 (0.851 - 0.945)**	**0.982 (0.936- 1.000)**	**0.941 (0.916 - 0.966)**
	▷ Average Precision ↑	▷ Average Recall ↑	▷ Average Specificity ↑	▷ Average NPV ↑	▷ Average F1 ↑
nnU-Net (Isensee et al., 2021)	0.702	0.832	0.911	0.899	0.732
SegResNet (Myronenko, 2019)	0.681	0.700	0.880	0.876	0.688
DenseUNet (Li et al., 2018)	0.692	0.698	0.883	0.879	0.692
UNet++ (Zhou et al., 2019)	0.679	0.769	0.906	0.894	0.709
TransUNet (Chen et al., 2021)	0.694	0.768	0.905	0.894	0.711
UNETR (Hatamizadeh, Tang, et al., 2022)	0.732	0.786	0.902	0.894	0.750
Swin UNETR (Hatamizadeh et al., 2021)	0.770	0.800	0.922	0.912	0.773
nnFormer (Zhou et al., 2023)	0.746	0.861	0.929	0.914	0.783
nn-TransEC (Ours)	**0.845**	**0.890**	**0.947**	**0.938**	**0.865**

[1] Note all listed models receive the same CT patch cropped by our proposed KRT method as input for fair comparison.
[2] Performance of these 3D models that receive the entire CT volume without cropping is reported in Table 13.

Table 9
Comparison with existing multi-task learning frameworks on joint segmentation and classification. Performances were reported for esophageal cancer segmentation and average values of esophageal cancer classification from T1 to T4. Results style: **best**, second-best.

Multi-task learning models	Segmentation of EC				Classification of EC's T-stages					
	3D Dice↑	*HD95↓*	*IoU↑*	*ASD↓*	*AUC↑*	*Precision↑*	*Sensitivity↑*	*Specificity↑*	*NPV↑*	*F1↑*
ELNet (Wu, Ge et al., 2021)	0.615	29.571	0.395	14.187	0.821	0.688	0.680	0.874	0.875	0.684
DSI-Net (Zhu et al., 2021)	0.753	17.623	0.566	7.073	0.812	0.656	0.712	0.879	0.874	0.678
Zhou et al. (2021)	0.781	13.450	0.621	3.905	0.841	0.721	0.696	0.885	0.884	0.701
TransMT-Net (Tang et al., 2023)	0.773	12.206	0.614	3.647	0.907	0.730	0.828	0.919	0.907	0.763
nn-TransEC (Ours)	**0.844**	**8.497**	**0.685**	**1.447**	**0.941**	**0.845**	**0.890**	**0.947**	**0.938**	**0.865**

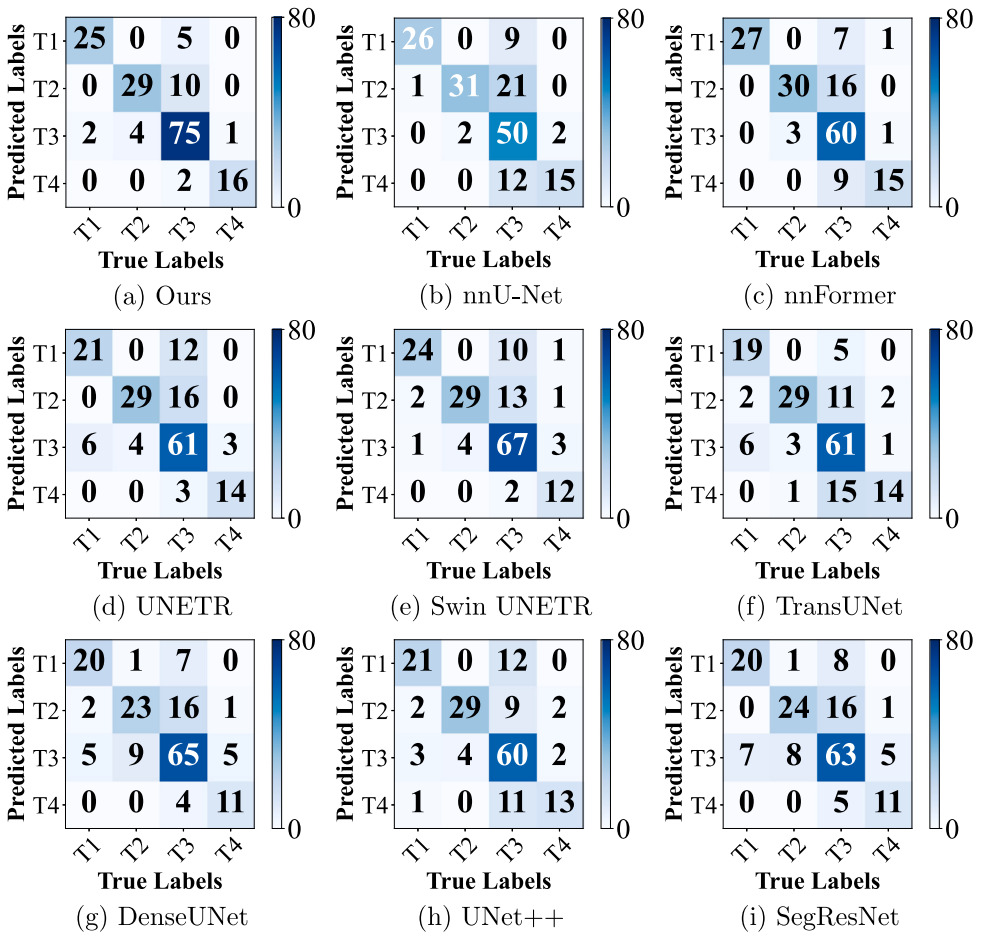


Fig. 8. Confusion matrix of nine methods on the classification of EC's T-stages. Note all listed models receive the same cropped CT patch after ROI tokenization for fair comparison.

4.2.4. Model and computational complexity

Table 10 presented a comparative analysis of the computational complexity among various 3D models, factoring in model parameters, FLOPs, and average inference times across 10 independent trials. The assessment was conducted using an input patch size of 128 × 64 × 64. The proposed nn-TransEC model was characterized by a moderate parameter count of 205.19M and a computational complexity of 22.43 GFLOPs. It exhibited competitive model complexity and significantly outperforms other Transformer-based and CNN-based models. For instance, Swin UNETR, with 196.77M parameters and 61.99 GFLOPs, nnFormer with 158.99M parameters and 128.80 GFLOPs, and V-Net with 198.31M parameters and 45.61 GFLOPs, were all surpassed by nn-TransEC in terms of parameters and FLOPs. Furthermore, the inference speed of nn-TransEC was the top-tier, following closely behind nnU-Net and V-Net, and substantially outpacing the remaining models in average inference time. It was noteworthy that the average inference times of the top-tier models were not significantly different, as all were capable of predicting a single input test 3D patch in under 100 ms.

Table 10
Comparison of the numbers of parameters, FLOPs and averaged inference time among various models that segment 3D medical images directly.

3D models	Params (M)	FLOPs (G)[a]	Inference time (ms)[a]
V-Net (Milletari et al., 2016a)	198.31	45.61	83.72
SegResNet (Myronenko, 2019)	325.47	42.28	150.07
UNet++ (Zhou et al., 2019)	690.92	26.64	239.73
UNETR (Hatamizadeh, Tang, et al., 2022)	334.45	115.18	144.20
Swin UNETR (Hatamizadeh et al., 2021)	196.77	61.99	169.21
nnU-Net (Isensee et al., 2021)	133.60	31.20	45.46
nnFormer (Zhou et al., 2023)	158.99	128.80	172.75
nn-MTNet (Ours)	205.19	22.43	96.77

[a] The numbers of FLOPs and inference time are computed with the input patchsize 128 × 64 × 64.

4.3. Ablation study

Comprehensive ablation studies had been conducted to verify the effectiveness of different components of nn-TransEC.

Table 11 reported the performance between the individual components and the combined scheme on the multi-organ segmentation. Experimental results indicated that:

- When deep supervision and residual connection were sequentially added to the segmentation baseline (a 3D U-Net), all segmentation metrics of esophageal cancer and the average values were significantly improved. And then the entire segmentation performance was improved even more after integrating the three components.
- The vanilla nnUNet-based strategy achieved moderately high performance on the segmentation of esophageal cancer and other related organs. The average performance of 3D Dice and IoU is further improved after modifying the backbone using deep supervision and residual connections.
- The overall segmentation performance was then further improved after integrating these three components.

Table 12 reported the performance between the individual components and the combined scheme on the classification of EC's T-stages. Experimental results indicated that:

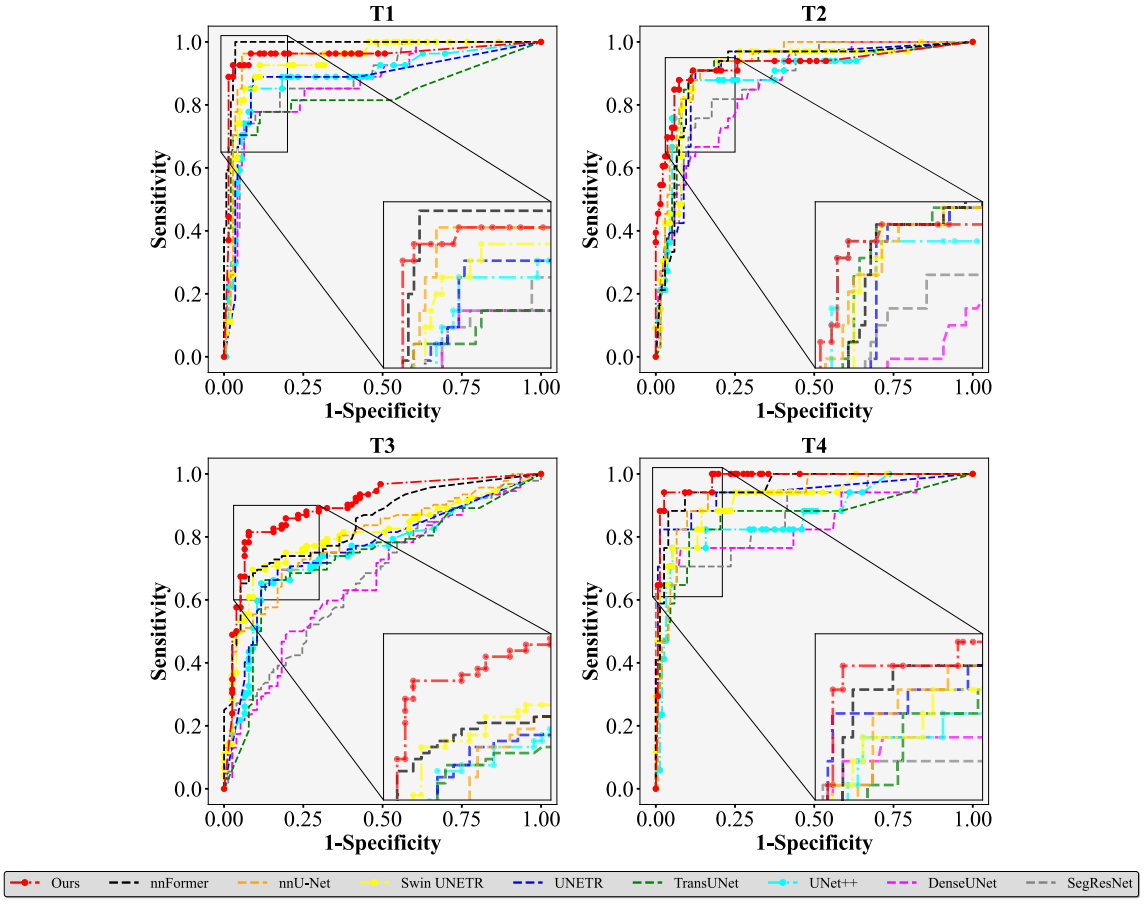


Fig. 9. The roc curves of nine methods on the classification of EC's T-stages are depicted for more intuitive comparisons. The AUC values reported in Table 8 can be calculated based on the above curves.

Table 11
Evaluation the effectiveness of four components in the proposed nn-TransEC on multi-organ segmentation. Performance of esophageal cancer (EC) and average value of all organs were reported.

Methods	#1	#2	#3	#4	*3D Dice* ↑		HD_{95} ↓		IoU ↑		ASD ↓	
					EC	Avg.	EC	Avg.	EC	Avg.	EC	Avg.
Baseline[a]					0.520	0.655	65.988	32.484	0.338	0.536	23.545	10.661
Ours	✓				0.726	0.776	19.277	11.670	0.555	0.680	7.770	3.146
	✓	✓			0.763	0.788	16.016	11.251	0.579	0.687	6.373	3.069
			✓		0.785	0.797	13.488	8.047	0.623	0.697	4.020	2.117
	✓	✓	✓		0.822	0.821	13.245	10.209	0.642	0.709	2.228	2.807
	✓	✓	✓	✓	0.844	0.825	8.497	8.205	0.685	0.713	1.447	1.900

#1 denotes deep supervision for adding auxiliary supervision to the segmentation network, described in the Section 3.2.1.
#2 denotes residual connections in the segmentation encoder, described in the Section 3.2.1.
#3 denotes the nnU-Net-based automatic configuration strategy for the segmentation network, described in the Section 3.3.
#4 denotes post-processing strategy, described in the Section 3.3.
[a] The same encoder and decoder of 3D U-Net are used without other components as baseline.

- When residual connection and attention enhancement were sequentially added to the classification baseline (an encoder of a 3D U-Net), the average values of six classification evaluation metrics were all significantly improved.
- The vanilla nnUNet-based strategy still achieved middle-high performance on the classification. After modifying the backbone using residual connection and attention enhancement, the average classification performance was further improved.
- The introduction of the KRT method yielded performance improvements in terms of six classification evaluation metrics. We could conclude that inputting the cropped CT patch further improved the performance of the baseline model, nnU-Net, and the integrated components, which was consistent with actual clinical diagnostic experience.
- The whole framework achieved the best performance after introducing transfer learning, demonstrating the effectiveness of all proposed components in the fine classification.

4.4. Segmentation reconstruction and volume quantization

Fig. 10 visualized the segmentation predictions of different methods in four views, i.e., axial view, sagittal view, coronal view and 3D reconstruction view. These results demonstrated that our segmentation predictions had the minimum discrepancy with the manual

Table 12
Evaluation the effectiveness of five components in the proposed nn-TransEC on classification of EC's T-stage. The same encoder of 3D U-Net and classification head are used without other components as baseline.

Methods	#1	#2	#3	#4	#5	Avg. AUC	Avg. Precision	Avg. Sensitivity	Avg. Specificity	Avg. NPV	Avg. F1
Baseline						0.538	0.333	0.297	0.777	0.770	0.271
	✓					0.579	0.602	0.335	0.780	0.829	0.337
	✓	✓				0.565	0.264	0.307	0.784	0.830	0.546
			✓			0.878	0.734	0.810	0.906	0.894	0.753
Ours				✓		0.641	0.369	0.380	0.802	0.799	0.371
			✓	✓		0.904	0.702	0.832	0.911	0.899	0.732
	✓	✓	✓			0.894	0.745	0.838	0.922	0.909	0.776
	✓	✓	✓	✓		0.925	0.769	0.851	0.923	0.911	0.782
	✓	✓	✓	✓	✓	0.941	0.845	0.890	0.947	0.938	0.865

#1 denotes residual connections in the classification encoder, described in the Section 3.2.1.
#2 denotes attention-aware cross-task enhancement between segmentation and classification, described in the Section 3.2.2.
#3 denotes the nnUNet-based automatic configuration strategy for network, described in the Section 3.3.
#4 denotes KRT-based method for cropping 3D CT volume as classification input, described in the Section 3.4.
#5 denotes that initialize the classification encoder with well-trained weights from segmentation encoder, described in the Section 3.2.3.

Table 13
This table reports the classification performance of the SOTA models on the EC169 dataset. Average AUC over T1–T4 and its 95% confidence interval (CI) were compared before and after introducing ROI tokenization method (KRT) for cropping CT patch as classification input. All results are based on the *5-fold cross-validation* with the same experimental protocol.

3D Models	Without KRT cropping	With KRT cropping
3D U-Net (Çiçek et al., 2016)	0.538 (0.488 - 0.589)	0.641 (0.591 - 0.691)
SegResNet (Myronenko, 2019)	0.772 (0.727 - 0.816)	0.829 (0.788 - 0.869)
DenseUNet (Li et al., 2018)	0.747 (0.701 - 0.793)	0.814 (0.772 - 0.855)
UNet++ (Zhou et al., 2019)	0.806 (0.763 - 0.848)	0.852 (0.814 - 0.890)
TransUNet (Chen et al., 2021)	0.795 (0.752 - 0.838)	0.837 (0.797 - 0.876)
UNETR (Hatamizadeh, Tang, et al., 2022)	0.812 (0.771 - 0.854)	0.873 (0.837 - 0.908)
Swin UNETR (Hatamizadeh et al., 2021)	0.840 (0.801 - 0.879)	0.895 (0.863 - 0.928)
nnU-Net (Isensee et al., 2021)	0.878 (0.843 - 0.913)	0.904 (0.873 - 0.936)
nnFormer (Zhou et al., 2023)	0.889 (0.855 - 0.923)	0.927 (0.899 - 0.954)
nn-TransEC (Ours)	**0.894** (**0.861** - **0.927**)	**0.941** (**0.916** - **0.966**)

annotations. Meanwhile, compared with other SOTA methods that generated misshapen predictions (e.g., UNet++ and UNETR), nn-TransEC generated more realistic and reasonable segmentation prediction. In addition, some SOTA methods (e.g., Swin UNETR and nnFormer) ignored the segmentation of small objects. For instance, the yellow region in Fig. 10 denoted lymph, which had the most errors and omissions. Our proposed method compensated for it and achieved promising lymph segmentation results. Above observations demonstrated the effectiveness of our proposed network in recognizing EC-related organs in the CT volume.

In addition to above qualitative visual comparison, this study further evaluated the ability of the proposed method in quantifying the volumes of EC and its surrounding related organs. Specifically, on the validation dataset ($N = 20$) of EC100, this section utilized the Pearson correlation coefficient (Sedgwick, 2012) to assess the consistency between the model-predicted volumes and the manual-labeled volumes. Eq. (5) defined the formulation of Pearson's r in this section.

Fig. 11 (A1–A5) depicted the results of measured Pearson correlation.[13] The p-values were also calculated to evaluate the statistical difference of the two variables, which was reported in the box at the top of subfigures. According to the previous research (Mukaka, 2012), results indicated that:

- Pearson's r were all greater than 0.9 with p-value <0.01 on the segmentation of aorta and pericardium, demonstrating the very high positive correlation between the predictions and the manual annotations.
- Pearson's r of esophageal cancer was greater than 0.8 with p-value <0.01, demonstrating the high positive correlation between the predictions and the manual annotations.
- Pearson's r were all greater than 0.5 with p-value <0.05 on the segmentation of esophagus and lymph, demonstrating the middle positive correlation between the predictions and the manual annotations.

Fig. 11 (B1–B5) depicted the Bland–Altman plots (Bland & Altman, 1986; Giavarina, 2015), which reported the distribution of the model-predicted volumes and the manual-labeled volumes. The horizontal coordinate was the average of the two volumes and the vertical coordinate was the difference between the two. Results indicated that the mean difference of volume between the ground truth and prediction was 1.911 cm^3 for the esophageal cancer, 1.124 cm^3 for aorta, 2.880 cm^3 for esophagus, 1.130 cm^3 for lymph, and –7.900 cm^3 for pericardium. As can be observed, most of the points fell within the 95% limits of agreement, i.e., within the two dashed black lines, and the mean difference was small and acceptable. Therefore, we could conclude that there was a good agreement between the two measured volumes so that the two methods could be replaced by each other.

4.5. Validation on the ROI tokenization (KRT) for cropping CT volume as input

In order to validate the generalization capability of the proposed KRT method for cropping 3D CT volumes as classification input, experiments were conducted to evaluate the performance of aforementioned state-of-the-art 3D models before and after ROI tokenization. Results were reported in the Table 13, which indicated that:

- If all 3D models did not receive cropped 3D CT volumes as input, the proposed nn-TransEC was still state-of-the-art and outperformed second-best model (nnFormer (Zhou et al., 2023)). After introducing ROI tokenization, nn-TransEC was still state-of-the-art and outperformed second-best model (nnFormer).

[13] Implemented by `scipy.stats.linregress` with version 1.10.1.

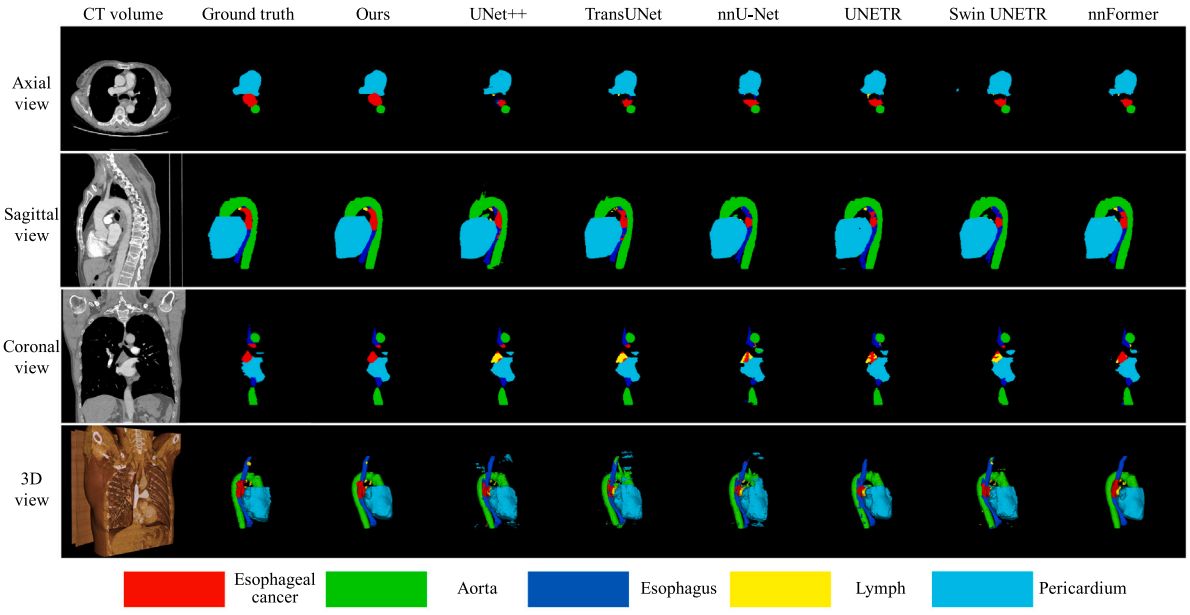


Fig. 10. Visualization of the ground truth annotated by doctors and the segmentation results predicted by state-of-the-art methods. From top to bottom, the results are presented in axial view, sagittal view, coronal view and 3D view. The third column to the last column are the segmentation results of different methods. Note that the 3D CT volume is reconstructed via 3D Slicer (Fedorov et al., 2012). For convenient observation, 3D view is colorized, and the same smoothing is added to predictions.

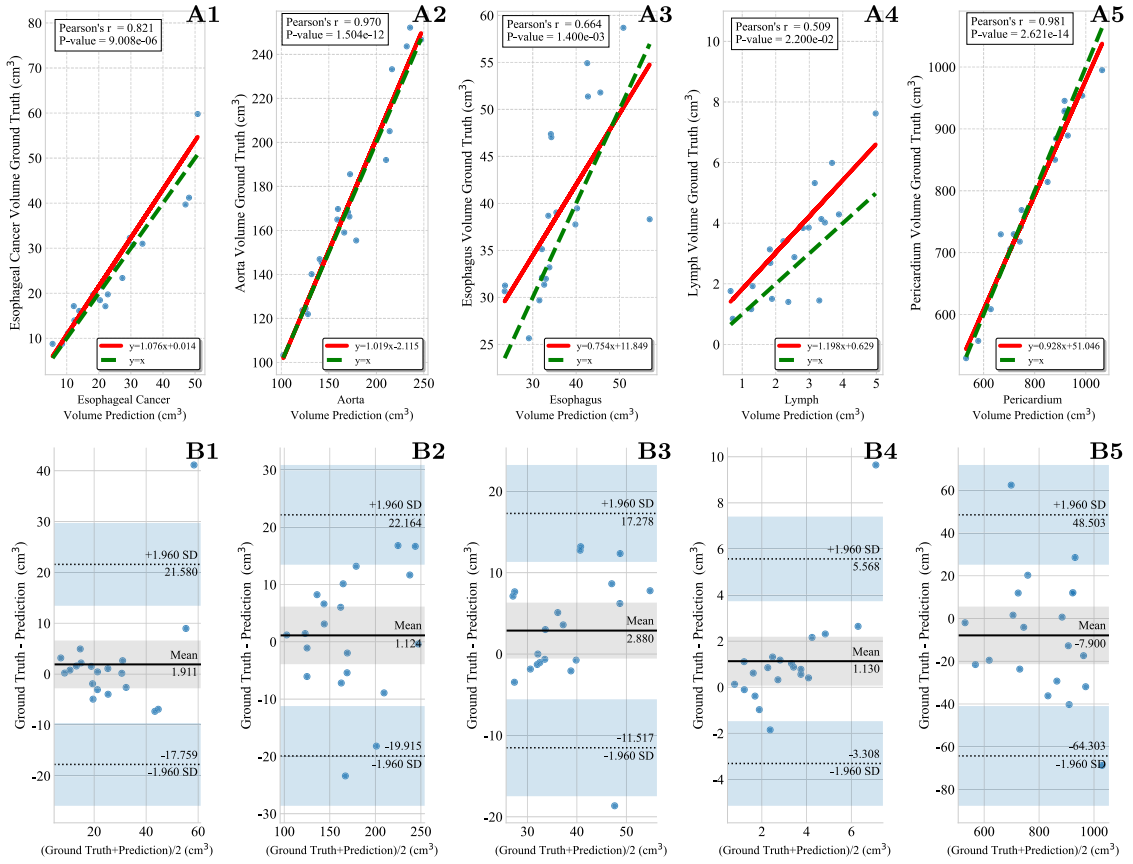


Fig. 11. Pearson Correlation (A1–A5) and Bland–Altman (B1–B5) plots of volume measured from ground truth and segmentation prediction. From left to right, these subfigures show the prediction of esophageal cancer, aorta, esophagus, lymph and pericardium, respectively. Note that each blue point in the above subfigures indicates one patient case. The results of the Pearson Correlation test are reported in the box at the top of subfigures A1–A5. The red lines in the subfigures A1–A5 denote the actual linear fitting results of all test cases, whose formulations are reported in the bottom boxes. The green dash lines in the subfigures A1–A5 denote the ideal situation. The black lines in subfigures B1–B5 indicate the mean difference between measured volumes, and the blue regions indicate the 95% limits of agreement.

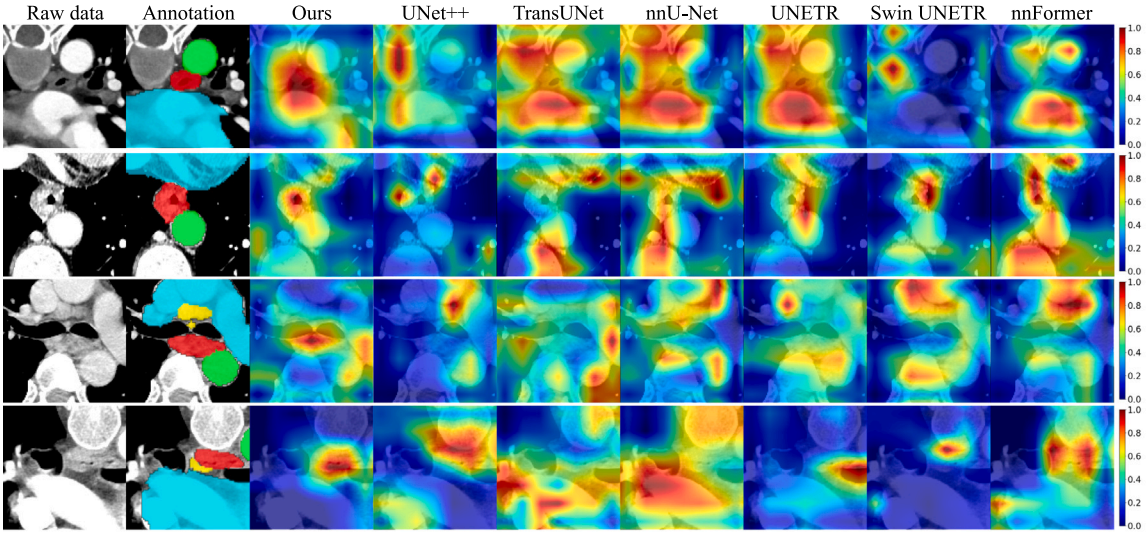


Fig. 12. Comparison of heatmaps generated by different methods during localizing EC-related organs. Note the red region represents the main concern of listed methods during classifying the T-stages.

- The proposed ROI tokenization was compatible to all listed cutting-edge models, whose performances improved after receiving
  cropped 3D CT volumes. Except for the baseline 3D U-Net model, the DenseUNet (Li et al., 2018) exhibited a noteworthy increase of 8.97 percentage points, elevating its AUC from 0.747 to 0.814.

Above observation demonstrated the generalization capability of the proposed ROI tokenization in improving the T-stage performance of the top models.

4.6. Model interpretation

In order to explore the interpretability of the network predictions, this section highlighted the regions in the CT volume, where the model considered to be closely related to the EC's T-stage. Specifically, this section first fed a single 3D CT volume into the well-trained network and then extracted the feature maps in the final convolutional layer after normalization. After that, the class activation mappings (Grad-CAM) (Selvaraju et al., 2017, 2020) were utilized to generate heatmaps as highlights. Finally, the heatmaps were upsampled to the size of the image and overlayed onto the image.

Fig. 12 exhibited four examples and compared the heatmap results of the seven methods. As can be observed, our work correctly localized the esophageal cancer and its surrounding tissues, and these regions received high attention from the model. We argued that this phenomenon benefited the classification network to accurately recognize the EC's T-stage. Meanwhile, such predictions generated by nn-TransEC are highly consistent with region of interest observed by doctors. In contrast, some methods misinterpreted the key to recognizing esophageal cancer. For example, Swin UNETR and UNet++ focused on irrelevant regions when classifying esophageal cancer T-staging. TransUNet and nnU-Net incorrectly expanded the regions surrounding esophageal cancer, resulting in decreased accuracy.

In conclusion, our proposed method owned model interpretability and its predictions were clinically interpretable.

5. Discussion

5.1. Advantages

Overall, nn-TransEC had made significant progresses towards alleviating the open issues in esophageal cancer diagnosis:

- The strong complementary relationship between segmentation of EC-related organs and classification of EC's T-stage is bridged by cross-task attention mechanism and transfer learning. These advanced techniques allow for the efficient sharing of esophageal cancer diagnosis representations between the two crucial and relevant tasks, enhancing the overall performance. On the one hand, cross-task attention enables the classification network to focus on relevant features from the segmentation task that are beneficial for T-stage classification. On the other hand, transfer learning leverages pre-trained segmentation network to carry out weight adaptation, facilitating more accurate and reliable T-stage classification.
- High-accuracy tumor T-staging from CT images has been prompted by tokenizing the coarse segmentation predictions. This innovative technique enables the precise determination of whether a tumor has invaded the surrounding muscle layer, a crucial factor in assessing the stage and prognosis of the EC. Through this approach, clinicians can obtain more reliable and detailed information about the tumor's extent and location, leading to improved patient outcomes through more targeted and effective treatment plans.

Therefore, nn-TransEC can be used as an automatic T-staging diagnosis system to segment the tumor, and certain the location, size, and shape of the EC and adjacent structures from CT images, and further predict the depth and range of tumor invasion, thereby improving the accuracy of T-staging diagnosis of EC and diagnostic efficiency. nn-TransEC not only reduces the burden of doctors, but also optimizes medical resource's utilization.

5.2. Methodological advances of nn-TransEC surpassing SOTA in esophageal cancer diagnosis

In Section 4.2, the proposed nn-TransEC was compared with the state-of-the-art models and achieved the best performance. This section analyzed the reason behind it from a methodological perspective.

As observed in Tables 7 and 8, the proposed nn-MTNet model in nn-TransEC framework outperformed its baseline model (nnU-Net) in both segmentation and classification tasks due to the optimizations detailed in Sections 3.2 and 3.3 for nnU-Net. Specifically:

- the deep supervision-based residual connections in the nn-MTNet enhanced the encoder's extraction of deep features.

- the cross-task attention gates in the nn-MTNet facilitated feature transfer between segmentation and classification tasks.
- the cross-task transfer learning in the nn-MTNet provided better initialization for the classification encoder.
- the modified nnUNet-based self-configuration strategy tailored nn-MTNet's model training parameters for esophageal cancer diagnosis. Notably, heuristic rules were instrumental in optimizing parameters for cross-task learning, and the loss function was specifically refined to mitigate pixel-level category imbalance prevalent in EC segmentation.

Collectively, these factors contributed to nn-TransEC's superior performance over nnU-Net and its variant nnFormer. The performance of other listed state-of-the-art models, including TransUNet, UNETR and Swin UNETR, also fell short of nn-TransEC in both segmentation and classification tasks. This gap may be attributed to the lack of multi-task learning optimizations.

As observed in Table 9, the proposed nn-TransEC surpassed other state-of-the-art multi-task learning (MTL) methods, attributable to its medically sound KRT method and advanced transfer learning strategy.

- Unlike other listed MTL methods, such as Zhou et al. (2021) and DSI-Net (Zhu et al., 2021), which learned from the original input images at a fixed resolution and were thus hindered by the majority of irrelevant areas, nn-TransEC mimicked the surgeon's T-staging diagnosis process for esophageal cancer. It also introduced a novel Knowledge-embedded ROI Tokenization (KRT) method. The KRT method utilized segmentation predictions to crop the input image to the region of interest, which significantly reduced the misguidance from irrelevant regions and enhanced learning for the subsequent classification of EC's T-stages, leading to improved performance.
- Compared to the proposed nn-TransEC, which employed nnU-Net-based self-configuration and cross-task transfer learning strategies, the cross-task learning mechanisms of other listed state-of-the-art MTL methods were less sophisticated and effective. For example, Zhou et al. (2021) directly applied high-level features extracted by V-Net (Milletari, Navab, & Ahmadi, 2016b) to multi-task learning. Although this approach achieved commendable performance in the segmentation and benign/malignant classification of 3D breast cancer, it encountered performance degradation when applied to segmenting esophageal cancer (EC) and its related organs, as well as in the classification of EC's T-stages. Another example was the ELNet (Wu, Ge et al., 2021), which utilized a dual-stream network for the classification of esophageal cancer lesions and a U-Net (Ronneberger et al., 2015) for the segmentation of the lesion, and it demonstrated prominence in the diagnosis of esophageal cancer from endoscopic images. However, the absence of cross-task transfer learning between the segmentation and classification networks resulted in performance degradation in the CT image-based esophageal cancer diagnosis tasks, which were the focus of the study.

Additionally, the ablation study in Tables 11 and 12 also proved the effectiveness of above components in the proposed nn-TransEC framework.

6. Conclusions

Aiming at the grand challenges for esophageal cancer T-stage diagnosis, with a primary focus on accurate tumor segmentation and T-stage classification, this paper developed a nnUNet-based 3D transfer learning framework (nn-TransEC). nn-TransEC synergized nnU-Net self-configuration strategy and cross-task transfer learning to enhance multi-task learning, i.e., joint multi-organ segmentation and fine-grained T-stage classification from CT images. nn-TransEC enabled complementary link between above two crucial and related tasks with a novel cross-task attention gate and cross-task transfer learning. Through the proposed ROI tokenization, nn-TransEC mimicked how doctors would diagnose esophageal cancer from CT images and transferred shared representations of the concerned organs from coarse ROI segmentation to fine T-stage classification. The nnUNet-based self-configuration strategy was modified to adapt the proposed multi-task learning framework, providing an advanced neural network for joint segmentation and classification.

Extensive experimental results indicated that: (1) nn-TransEC could mimic how doctors diagnoses EC's T-stages from CT images and bridge segmentation of EC-related organs and classification of EC's T-stage as a whole and provide better performance for multi-task learning by cross-task attention and transfer learning, (2) the proposed KRT and transfer learning methods contribute to aligning regions of interest with doctors in T-stage diagnosis, increasing AUC of T-stages to 0.941, thus providing a clinically interpretable enhancement for fine T-staging classification of EC, and (3) nn-TransEC could quantify the volumes of EC-related organs and had a good agreement with doctors' manually annotated results, which necessitated time-consuming pixel-by-pixel labeling. Therefore, nn-TransEC could serve as a relief from manual annotation in the esophageal cancer diagnosis Above results suggested that the key to sustaining reliable diagnosis of esophageal cancer from CT images lay with appropriate portraying a priori clinic knowledge in optimization.

However, there were still some limitations of the proposed nn-TransEC to be further explored:

- In the task of fine T-staging classification, there was no statistically significant difference in performance between nn-TransEC and the SOTA models when diagnosing the early esophageal cancer in T1 and T2 T-stages (see Table 8). Therefore, **future work** may focus on enhancing the optimization of multi-task learning models for fine-grained classification tasks.
- In the task of coarse multi-organ segmentation, performance of the proposed nn-TransEC was not statistically different from that of the listed models when segmenting certain adjacent large organs, such as aorta and pericardium (see Table 7). Therefore, **future work** could concentrate on designing better loss functions to balance the segmentation performance and generalization capabilities of the model within the context of multi-organ segmentation.
- As shown in Table 10, the proposed nnTrans-EC did not demonstrated a notable advantage in computational efficiency, such as FLOPs. Therefore, **future work** could focus on optimizing this aspect. This could involve updating the nnTrans-EC framework with advanced models, such as the VMamba (Ruan & Xiang, 2024; Wang, Zheng, Zhang, Cui, & Li, 2024; Yue & Li, 2024; Zhu, Liao, et al., 2024), or other lightweight models, to enhance computational efficiency.

CRediT authorship contribution statement

Chen Li: Conceptualization, Methodology, Writing – original draft. **Runyuan Wang:** Data curation, Software, Writing – review & editing. **Ping He:** Supervision, Validation. **Wei Chen:** Conceptualization, Project administration, Funding acquisition. **Wei Wu:** Resources, Project administration, Funding acquisition. **Yi Wu:** Project administration, Funding acquisition, Writing – review & editing.

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

Funding

This work was supported in part by the Chongqing Science and Technology Talent Project, China [grant number CQYC201905037]; the Chongqing Key Research and Development Project, China [grant number CSTB2022TIAD-KPX0181]; the Chongqing Science and Technology Project, China [grant number cstc2022ycjh-bgzxm0071]; the Science-Health Joint Medical Scientific Research Project of Chongqing, China [grant number 2022ZDXM018]; the National Key Research and Development Program of China [grant number 2018YFB0204301]; the Natural Science Foundation of Hunan Province of China [grant number 2022JJ30666]; the National Natural Science Foundation of China [grant number 31971113]; and the and University Funded Science and Technology Innovation Capacity Improvement Project, China [grant number 2019XYY14].

Appendix A and Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2024.125067.

References

Atlanta (2023). *Cancer facts & figures*. American Cancer Society.

Begon, F., Lockhart, A., Metreau, J., & Dhumeaux, D. (1979). A computer-aided system for the diagnosis of hepato-biliary diseases. A comparison with the performance of physicians. *Medical Informatics*, *4*(1), 35–42.

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Casella, A., Moccia, S., Paladini, D., Frontoni, E., De Momi, E., & Mattos, L. S. (2021). A shape-constraint adversarial framework with instance-normalized spatio-temporal features for inter-fetal membrane segmentation. *Medical Image Analysis (MedIA)*, *70*, Article 102008.

Chattopadhyay, S., Dey, A., Singh, P. K., & Sarkar, R. (2022). DRDA-Net: Dense residual dual-shuffle attention network for breast cancer classification using histopathological images. *Computers in Biology and Medicine*, *145*, Article 105437.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th international conference on machine learning*.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, URL https://arxiv.org/abs/2102.04306.

Chen, S., Wang, Z., Shi, J., Liu, B., & Yu, N. (2018). A multi-task framework with feature passing module for skin lesion classification and segmentation. In *2018 IEEE 15th international symposium on biomedical imaging* (pp. 1126–1129). IEEE.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical image computing and computer assisted intervention* (pp. 424–432). Springer.

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., et al. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In *IEEE international symposium on biomedical imaging* (pp. 168–172).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Enzinger, P. C., & Mayer, R. J. (2003). Esophageal cancer. *New England Journal of Medicine*, *349*(23), 2241–2252.

Fechter, T., Adebahr, S., Baltas, D., Ben Ayed, I., Desrosiers, C., & Dolz, J. (2017). Esophagus segmentation in CT via 3D fully convolutional neural network and random walk. *Medical Physics*, *44*(12), 6341–6352.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., et al. (2012). 3D slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, *30*(9), 1323–1341.

Ferreira, J., Domingues, I., Sousa, O., Sampaio, I. L., & Santos, J. A. M. (2020). Classification of oesophagic early-stage cancers: Deep learning versus traditional learning approaches. In *2020 IEEE 20th international conference on bioinformatics and bioengineering* (pp. 746–751). http://dx.doi.org/10.1109/BIBE50027.2020.00127.

Giavarina, D. (2015). Understanding bland altman analysis. *Biochemia Medica*, *25*(2), 141–151.

Guo, L., Xiao, X., Wu, C., Zeng, X., Zhang, Y., Du, J., et al. (2020). Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). *Gastrointestinal Endoscopy*, *91*(1), 41–51.

Han, L., Zhang, Y., Song, G., & Xie, K. (2014). Encoding tree sparsity in multi-task learning: A probabilistic framework. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1854–1860).

Harouni, A., Karargyris, A., Negahdar, M., Beymer, D., & Syeda-Mahmood, T. (2018). Universal multi-modal deep network for classification and segmentation of medical images. In *2018 IEEE 15th international symposium on biomedical imaging* (pp. 872–876). IEEE.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021). Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI brainlesion workshop* (pp. 272–284). Springer.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., et al. (2022). Unetr: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574–584).

Hatamizadeh, A., Xu, Z., Yang, D., Li, W., Roth, H., & Xu, D. (2022). Unetformer: A unified vision transformer model and pre-training framework for 3D medical image segmentation. arXiv preprint arXiv:2204.00631.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hosseini, F., Asadi, F., Emami, H., & Ebnali, M. (2023). Machine learning applications for early detection of esophageal cancer: A systematic. *BMC Medical Informatics and Decision Making*, *23*, 124.

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 1501–1510).

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*(2), 203–211.

Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., et al. (2024). Nnu-net revisited: A call for rigorous validation in 3D medical image segmentation. arXiv preprint arXiv:2404.09556.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference for learning representations*.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., & Heng, P.-A. (2018). H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging (TMI)*, *37*(12), 2663–2674.

Li, C., Tan, Y., Chen, W., Luo, X., He, Y., Gao, Y., et al. (2020). ANU-Net: Attention-based nested U-net to exploit full resolution features for medical image segmentation. *Computers & Graphics*, *90*, 11–20.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Lin, Q., Tan, W., Cai, S., Yan, B., Li, J., & Zhong, Y. (2023). Lesion-decoupling-based segmentation with large-scale colon and esophageal datasets for early cancer diagnosis. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 1–15. http://dx.doi.org/10.1109/TNNLS.2023.3248804.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Lozano, A. C., & Swirszcz, G. (2012). Multi-level lasso for sparse multi-task regression. In *Proceedings of the 29th international coference on international conference on machine learning* (pp. 595–602).

Malhotra, G. K., Yanala, U., Ravipati, A., Follet, M., Vijayakumar, M., & Are, C. (2017). Global trends in esophageal cancer. *Journal of Surgical Oncology*, *115*(5), 564–579.

McConnell, N., Ndipenoch, N., Cao, Y., Miron, A., & Li, Y. (2023). Exploring advanced architectural variations of nnUNet. *Neurocomputing*, *560*, Article 126837. http://dx.doi.org/10.1016/j.neucom.2023.126837.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016a). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision* (pp. 565–571).

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016b). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision* (pp. 565–571). http://dx.doi.org/10.1109/3DV.2016.79.

Mishra, S., Zhang, Y., Zhang, L., Zhang, T., Hu, X. S., & Chen, D. Z. (2022). Data-driven deep supervision for skin lesion classification. In *Medical image computing and computer assisted intervention* (pp. 721–731). Springer.

Misra, I., Shrivastava, A., Gupta, A., & Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3994–4003).

Mormont, R., Geurts, P., & Marée, R. (2020). Multi-task pre-training of deep neural networks for digital pathology. *IEEE Journal of Biomedical and Health Informatics (JBHI)*, *25*(2), 412–421.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69–71.

Myronenko, A. (2019). 3D MRI brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop* (pp. 311–320). Springer.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-Net: Learning where to look for the pancreas. In *Medical imaging with deep learning*.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *22*(10), 1345–1359.

Pennathur, A., Gibson, M. K., Jobe, B. A., & Luketich, J. D. (2013). Oesophageal carcinoma. *The Lancet*, *381*(9864), 400–412.

Profillidis, V., & Botzoris, G. (2019). Chapter 5 - Statistical methods for transport demand modeling. In V. Profillidis, & G. Botzoris (Eds.), *Modeling of transport demand* (pp. 163–224).

Rice, T. W., Ishwaran, H., Ferguson, M. K., Blackstone, E. H., & Goldstraw, P. (2017). Cancer of the esophagus and esophagogastric junction: An eighth edition staging primer. *Journal of Thoracic Oncology*, *12*(1), 36–42.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention* (pp. 234–241).

Ruan, J., & Xiang, S. (2024). Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491.

Rustgi, A. K., & El-Serag, H. B. (2014). Esophageal carcinoma. *New England Journal of Medicine*, *371*(26), 2499–2509.

Sang, Z., Li, C., Xu, Y., Wang, Y., Zheng, H., & Guo, Y. (2024). FCTformer: Fusing convolutional operations and transformer for 3D rectal tumor segmentation in MR images. *IEEE Access*, *12*, 4812–4824. http://dx.doi.org/10.1109/ACCESS.2024.3349409.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., et al. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis (MedIA)*, *53*, 197–207. http://dx.doi.org/10.1016/j.media.2019.01.012.

Sedgwick, P. (2012). Pearson's correlation coefficient. *British Medical Journal (BMJ)*, *345*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision (IJCV)*, *128*(2), 336–359.

Shen, C., Wang, C., Roth, H. R., Oda, M., Hayashi, Y., Misawa, K., et al. (2020). Spatial information-embedded fully convolutional networks for multi-organ segmentation with improved data augmentation and instance normalization. In *Medical imaging 2020: image processing: vol. 11313*, SPIE, Article 1131316. http://dx.doi.org/10.1117/12.2550496.

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, *73*(1), 17–48.

Strezoski, G., van Noord, N., & Worring, M. (2019). Learning task relatedness in multi-task learning for images in context. In *Proceedings of the 2019 on international conference on multimedia retrieval* (pp. 78–86).

Sui, H., Ma, R., Liu, L., Gao, Y., Zhang, W., & Mo, Z. (2021). Detection of incidental esophageal cancers on chest CT by deep learning. *Frontiers in Oncology*, *11*, Article 700210.

Sun, F., Chen, W., Fu, S., & Liu, N. (2023). TNM staging for adrenocortical carcinoma using SimCLR: A deep learning approach: SimCLR-based TNM staging for adrenocortical carcinomaa comprehensive deep learning approach for adrenocortical carcinoma TNM staging. In *Proceedings of the 2023 4th international symposium on artificial intelligence for medicine science* (pp. 642—-646).

Takeuchi, M., Seto, T., Hashimoto, M., Ichihara, N., Morimoto, Y., Kawakubo, H., et al. (2021). Performance of a deep learning-based identification system for esophageal cancer from CT images. *Esophagus*, *18*, 612–620.

Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., et al. (2022). Self-supervised pre-training of swin transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20730–20740).

Tang, S., Yu, X., Cheang, C.-F., Hu, Z., Fang, T., Choi, I.-C., et al. (2022). Diagnosis of esophageal lesions by multi-classification and segmentation using an improved multi-task deep learning model. *Sensors*, *22*(4).

Tang, S., Yu, X., Cheang, C. F., Liang, Y., Zhao, P., Yu, H. H., et al. (2023). Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images. *Computers in Biology and Medicine*, *157*, Article 106723.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, URL http://arxiv.org/abs/1607.08022.

Van Riel, S., Van Der Sommen, F., Zinger, S., Schoon, E. J., & de With, P. H. (2018). Automatic detection of early esophageal cancer with CNNS using transfer learning. In *2018 25th IEEE international conference on image processing* (pp. 1383–1387). http://dx.doi.org/10.1109/ICIP.2018.8451771.

Wang, X., Bi, J., Yu, S., & Sun, J. (2014). On multiplicative multitask feature learning. In *Advances in neural information processing systems (NeurIPS): vol. 27*, Curran Associates, Inc..

Wang, Z., Zheng, J.-Q., Zhang, Y., Cui, G., & Li, L. (2024). Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079.

Wu, Y.-H., Gao, S.-H., Mei, J., Xu, J., Fan, D.-P., Zhang, R.-G., et al. (2021). JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation. *IEEE Transactions on Image Processing (TIP)*, *30*, 3113–3126. http://dx.doi.org/10.1109/TIP.2021.3058783.

Wu, Z., Ge, R., Wen, M., Liu, G., Chen, Y., Zhang, P., et al. (2021). ELNet:Automatic classification and segmentation for esophageal lesions using convolutional neural network. *Medical Image Analysis (MedIA)*, *67*, Article 101838. http://dx.doi.org/10.1016/j.media.2020.101838.

Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., & Yu, S. (2021). A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, *69*, Article 101985.

Xie, Y., Zhang, J., Xia, Y., & Shen, C. (2020). A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Transactions on Medical Imaging (TMI)*, *39*(7), 2482–2493.

Yanase, J., & Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, *138*, Article 112821.

Yang, Z., Cao, Z., Zhang, Y., Tang, Y., Lin, X., Ouyang, R., et al. (2021). MommiNet-v2: Mammographic multi-view mass identification networks. *Medical Image Analysis*, *73*, Article 102204.

Yousefi, S., Sokooti, H., Elmahdy, M. S., Lips, I. M., Shalmani, M. T. M., Zinkstok, R. T., et al. (2021). Esophageal tumor segmentation in CT images using a dilated dense attention unet (DDAUnet). *IEEE Access*, *9*, 99235–99248.

Yousefi, S., Sokooti, H., Elmahdy, M. S., Peters, F. P., Shalmani, M. T. M., Zinkstok, R. T., et al. (2018). Esophageal gross tumor volume segmentation using a 3D convolutional neural network. In *Medical image computing and computer assisted intervention* (pp. 343–351).

Yu, X., Tang, S., Cheang, C. F., Yu, H. H., & Choi, I. C. (2022). Multi-task model for esophageal lesion analysis using endoscopic images: Classification with image retrieval and segmentation with attention. *Sensors*, *22*(1).

Yue, Y., & Li, Z. (2024). Medmamba: Vision mamba for medical image classification. arXiv preprint arXiv:2403.03849.

Zhang, Y., & Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, *5*(1), 30–43.

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6848–6856).

Zhao, Z., Li, M., Liu, P., Yu, J., & Zhao, H. (2022). Efficacy of digestive endoscope based on artificial intelligence system in diagnosing early esophageal carcinoma. *Computational and Mathematical Methods in Medicine*, *2022*.

Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., et al. (2021). Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Medical Image Analysis (MedIA)*, *70*, Article 101918.

Zhou, H.-Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., et al. (2023). nnFormer: Volumetric medical image segmentation via a 3D transformer. *IEEE Transactions on Image Processing (TIP)*, *32*, 4036–4045. http://dx.doi.org/10.1109/TIP.2023.3293771.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging (TMI)*, *39*(6), 1856–1867.

Zhu, M., Chen, Z., & Yuan, Y. (2021). DSI-Net: Deep synergistic interaction network for joint classification and segmentation with endoscope images. *IEEE Transactions on Medical Imaging (TMI)*, *40*(12), 3315–3325.

Zhu, W., Jin, Y., Ma, G., Chen, G., Egger, J., Zhang, S., et al. (2024). Classification of lung cancer subtypes on CT images with synthetic pathological priors. *Medical Image Analysis*, *95*, Article 103199.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417.