Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Towards Better Laparoscopic Video Segmentation: A Class-Wise Contrastive Learning Approach with Multi-Scale Feature Extraction

Luyang Zhang[1] | **Yuichiro Hayashi**[1] | **Masahiro Oda**[2,1] | **Kensaku Mori**[1,2,3]

[1]Graduate School of In formatics, Nagoya University, Aichi, Japan

[2]Information and Communications, Nagoya University, Aichi, Japan

[3]Research Center of Medical Bigdata, National Institute of Informatics, Tokyo, Japan

Correspondence
Corresponding author Luyang Zhang.
Email: lzhang@mori.m.is.nagoya-u.ac.jp
Corresponding author Kensaku Mori,
Email: kensaku@is.nagoya-u.ac.jp.

Present address
1-4 Furocho, Chikusa Ward, Nagoya, Aichi 464-0814,Japan

Funding Information
This research was supported by the Text

JEL Classification: ejlje

Abstract

The task of segmentation is integral to computer-aided surgery systems, as it provides the shape of the organ and the location of the tools. However, the practical application of precise segmentation methods is limited due to the need for annotated data. Given the privacy concerns associated with medical data and the high cost of manual annotation, collecting a large amount of annotated data for training is challenging. Unsupervised learning techniques, such as contrastive learning, have shown powerful capabilities in learning image-level representations from unlabeled data. Additionally, compared to pixel-level segmentation labels for images, classification labels are easier to acquire. In this study, we leveraged classification labels to enhance the accuracy of the segmentation model trained on limited annotated data. Our method uses a multi-scale projection head to extract image features at various scales. We then improved the partitioning method for positive sample pairs to perform contrastive learning on the extracted features at each scale. This approach effectively represents the differences between positive and negative samples in contrastive learning. Furthermore, with the features extracted by the multi-scale projection head, our model is trained simultaneously with both segmentation labels and classification labels. This enables the model to extract features more effectively from each segmentation target class and further accelerates the convergence speed. Our method was validated using the publicly available CholecSeg8k dataset for comprehensive abdominal cavity surgical segmentation. Compared to select existing methods, our proposed approach significantly enhances segmentation performance, even with a small labeled subset (1%-10%) of the dataset, showcasing a superior Intersection over Union (IoU) score.

KEYWORDS
Self-Unsupervised Learning, Contrastive Learning, Multi-task Learning, Laparoscopic Video Segmentation

1 | INTRODUCTION

Semantic segmentation of anatomical structures in laparoscopic videos is a vital component in facilitating the development of novel computer-assisted systems, which are designed to augment surgical precision and support surgeons during operations. The complexity of this task is exacerbated by the presence of artifacts such as shadows, reflections, and occlusions, as well as the inherently similar visual characteristics of anatomical structures found in laparoscopic footage. Recent advancements in deep learning algorithms have shown promising results in the segmentation of anatomical structures in laparoscopic videos[1,2].

At present, researchers typically employ Convolutional Neural Networks (CNNs) for medical image segmentation tasks. A notable contribution in this regard is the U-Net model, introduced by Ronneberger and his team, which has found broad applications in the domain of medical imaging[3].

However, the development and optimization of these advanced technologies are not without challenges. Researchers typically employ CNNs for medical image segmentation tasks. Firstly, their method of inferring based on the integration of multi-scale feature information, along with their simple structure, makes them the most common model in segmentation tasks. However,

the majority of these CNN-based frameworks are heavily reliant on a supervised learning environment, which demands a large amount of manually annotated data. This is particularly the case for segmentation tasks that require pixel-level annotations, resulting in an enormous annotation workload and presenting considerable challenges for practical applications. This scarcity represents a significant barrier to progress, as the absence of sufficiently labeled datasets hampers the training and fine-tuning of algorithms that are integral to the continual improvement of computer-assisted surgical systems[4,1].

In order to improve the accuracy of training models with limited samples, self-supervised learning (SSL) enhances feature extractors by designing tasks that utilize the samples themselves rather than annotated data. This enables the model to learn intrinsic features within the samples. Contrastive learning has recently become a prevailing SSL method because of its superior performance[5,6,7]. In 2019, the field of machine learning saw significant advancements in contrastive learning methods. In the context of contrastive learning, the main idea is to compare and contrast pairs of samples in a representation space. The goal is to learn meaningful representations by pulling similar pairs closer together and pushing dissimilar pairs further apart. MoCo[8] was introduced, using a strong augmentation of the same image as a positive sample while treating other images as negative samples for contrastive learning. MoCo and MoCoV2[9] introduced a sequence data structure and a memory bank[10] to reduce the computational cost of contrastive learning, using an exponential moving average (EMA)[11] for smoother model training. In the same year, SimCLR[12] proposed a learnable projection head[13], eliminating the need for a memory bank, and introduced an improved loss function, significantly enhancing model performance. Later, Bootstrap Your Own Latent (BYOL)[14] was introduced, allowing for reinforcement learning using a single sample as both the positive and negative pair. MoCoV3[15] further refined these techniques by incorporating a transformer, significantly improving classification accuracy.

Multi-task learning (MTL) is a method that enables models to learn better representations for each category, which aims to improve the performance of multiple related tasks by learning them simultaneously. The underlying hypothesis in MTL is that tasks share some inherent associations that can be leveraged to boost the generalization performance of individual tasks. In the realm of medical image analysis, MTL has shown promising results by effectively utilizing various forms of auxiliary information[16]. One common application of MTL in medical image analysis involves using classification labels for improving segmentation performance. The central idea is that the auxiliary task of classifying an image (or a region of it) provides an additional source of gradient during training, which helps to regularize the learning process of the primary segmentation task. This setup has been shown to improve the robustness of the model against overfitting and often results in better generalization to unseen data. Particularly in situations where segmentation labels are costly to obtain, the use of classification labels, which are often easier to acquire, can significantly enhance the segmentation performance[17,18].

However, there are still considerable challenges in the task of endoscopic surgical image segmentation. Firstly, in the current contrastive learning methods, the definition of positive samples is typically an original image and the same image after image enhancement. This is indeed a very feasible method for dividing positive samples in large-scale multi-class datasets like ImageNet, but it is not applicable to endoscopic surgical images, because endoscopic surgical images have the following characteristics: 1. Images of the same category are extremely similar. This is due to the relatively fixed observation angle in endoscopic surgery, which will not easily change, and the viewpoints of images during the same surgery are almost all concentrated in the same parts, with similar backgrounds1. The surgical procedures are similar, and the relative positions of surgical tools and organs are very close. Therefore, it is not accurate enough to simply divide all different images into negative samples: the images taken during the same surgical stage, using the same instruments to perform the same operation, should contain similar information. In addition, current contrastive learning methods in training are aimed at classification problems, so they do not consider features extracted at multiple scales. But in segmentation models, whether it is classic structures like U-Net and DeepLab[19], or transformer-based models like Segformer[20], they all use the multi-scale features extracted by the encoder in the decoder stage through long connections. Lastly, the process of contrastive learning usually includes the pretrain-finetune stages. This lowers the efficiency of model training. Also, in the finetune stage, due to the characteristics of the segmentation model, the decoder also contains a large number of parameters. Therefore, training the encoder and decoder separately will result in the encoder not being able to extract enough features for the decoder to train for the segmentation task.

Accordingly, to address the aforementioned challenges, we propose an innovative training methodology in our research. This method leverages the intrinsic attributes of endoscopic images and segmentation models, employing a modest set of segmentation labels together with a wide array of readily accessible classification labels. We redefine the sample pairs in accordance with the classification and surgical phase labels of the images. The aim of this modification is to facilitate the extraction of similar features from images with analogous characteristics within the same batch. This strategy can better underscore the differences between positive and negative samples in contrastive learning, thereby enhancing the model's performance on smaller training datasets. To tackle the challenge of multi-scale feature extraction, we introduce a multi-scale projection head (MSPH). This component is

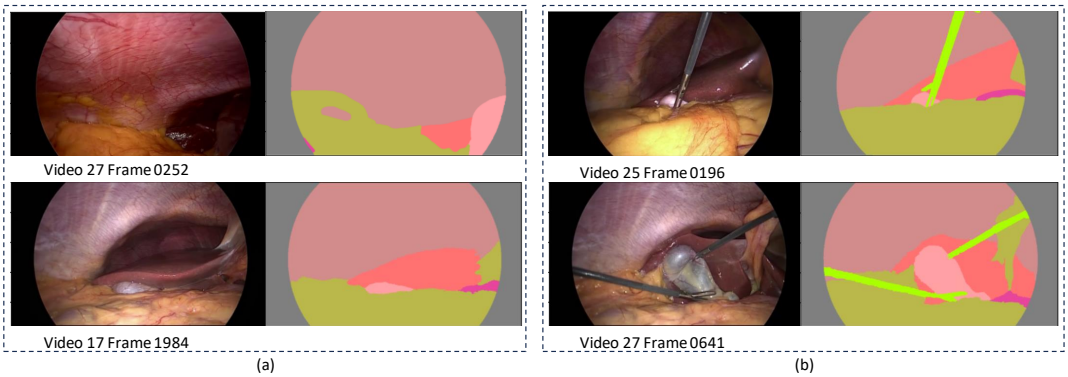

FIGURE 1 An example of positive sample pair division. In positive sample pair (*a*), the two images are taken from frame 252 of VID027 and frame 1984 of VID017 in the Cholecseg8k[4] dataset, respectively, in the same phase of preparation, containing the same types of organs: "Black Background", "Abdominal Wall", "Liver", and "Fat". In positive sample pair (*b*), the two images are taken from frame 196 of VID025 and frame 641 of VID027 in the Cholecseg8k dataset, respectively, in the same phase of preparation and containing the same types of tools and organs: "Black Background", "Abdominal Wall", "Liver", "Fat", "Grasper", and "Gallbladder".

designed to extract features from each encoder scale and optimize these features through deep contrastive learning. Moreover, we integrate the pretraining and fine-tuning processes by utilizing MTL. This simultaneous training for contrastive learning and segmentation tasks aims to improve model accuracy. To ensure that the encoder can extract more precise category representations, we incorporate a classification task during the training process. This step allows for a better representation of the differences between positive and negative samples in contrastive learning, leading to enhanced model performance on small training datasets. The primary objective of this approach is to improve the segmentation accuracy of models trained on a limited dataset of images with segmentation annotations. This innovation could significantly decrease the costs associated with image annotation.

In light of the above issues, the summary of novelty is as follows:

- Redefining positive pairs in contrastive learning based on class labels.
- Introducing MSPH for enhanced feature extraction at each scale.
- Implementing MTL to train for contrastive learning, classification, and segmentation tasks concurrently, thereby improving model accuracy.

2 | METHOD

2.1 | Class-Wise Positive Pairs Defining

In the context of endoscopic surgery images, surgical procedures typically occur within a confined region. Consequently, when comparing two surgical images that involve similar tools and organs, they tend to possess highly comparable information. As depicted in Fig. 1, the semantic content within two images sharing identical classification labels also exhibits substantial similarity. Nevertheless, in previous contrastive learning tasks, a conventional approach involves defining an original image x and its augmented counterpart $\hat{x}$ as a positive sample pair, while considering all other images as negative samples. Undoubtedly, such an approach may lack precision and specificity in this work.

Therefore, we define the set of positive samples Ω_i^+ as follows: for all samples $\{ \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N \}$ in the training dataset, along with their corresponding multi-class labels $\mathbf{Y} = \{ \mathbf{y}_i \mid i = 1, 2, \ldots, N \}$, where n shows the size of the dataset. If the multi-class labels $\mathbf{y}_i$ and $\mathbf{y}_j$ of two samples are equal, implying that they are images captured during the same stage of the surgical procedure using the same tools, then these two samples $\mathbf{X}_i$ and $\mathbf{X}_j$ are deemed as positive samples, and the positive sample pair $(\mathbf{X}_i, \mathbf{X}_j)$ is included in Ω_i^+. It should be noted that the multi-class labels in this context encompass both the target categories present in the surgical images and the surgical phase being performed at the time of image capture. During the process of contrastive learning, we aim for the model's encoder to extract similar features from positive sample pairs.

2.2 | Multi-Scale Projection Head (MSPH)

In this research, we pinpointed a significant gap in the realm of contrastive learning tasks, with particular reference to segmentation models. It was observed that extant studies primarily concentrate on the extraction of high-level semantic features, often neglecting or underemphasizing the importance of features at intermediate levels.

Given the original architecture of the U-Net model, it facilitates the direct transmission of features extracted at each scale to the decoder through long-skip connections. This strategy enables the decoder to employ a combination of features from diverse scales, thereby leading to enhanced segmentation predictions. In previous studies, it has been proven that the introduction of multi-scale feature contrastive learning is effective in training classification models[21].

Inspired by these observations, we put forth a hypothesis that paying attention to full-scale semantic features extracted from the encoder of the segmentation model, rather than focusing solely on high-level semantic features, is of paramount importance during the contrastive learning process. We postulate that this multi-scale contrastive supervision approach could potentially improve the model's comprehension of the data and enhance its performance.

We formulate a set of single projection heads, represented as $\mathbf{G} = \{ g_v(\cdot) \mid v = 1, 2, \ldots, L \}$, which we refer to as the scale-wise projection head. Each g was defined as a single projection shown in Fig. 2. Assuming the feature maps derived from each scale of the Encoder are denoted as $\mathbf{F} = \{ \mathbf{f}_v \mid v = 1, 2, \ldots, L \}$, we can obtain the preliminary features following the projection head, which are defined as $\mathbf{F}' = \{ \mathbf{f}'_v \mid v = 1, 2, \ldots, L \}$. L is the number of layers we used to extract the features by MSPH. The features garnered from this process retain consistent dimensions across different scales. These projection heads have the capacity to map the original scale-specific features onto a uniform representation space, thereby enabling us to utilize the features extracted at different scales more effectively for deep class-wise contrastive learning.

In conclusion, our proposed MSPH module encompasses a suite of projection heads designed to extract features from feature maps across multiple scales. These extracted features are then harnessed for subsequent deep class-wise contrastive Learning.

2.3 | Deep Class-Wise Contrastive Learning

Our novel Deep Class-Wise Contrastive Learning encompasses two primary aspects: contrasting the features obtained from the MSPH on feature maps at different scales across multiple images and performing multi-class classification with the features extracted by the encoder across all scales from a single image. In this subsection, we first introduce the contrastive learning formulation put forth in this study. Traditional supervised contrastive learning typically employs the softmax function as the loss function[22], which can be defined as

$$L_{i,+} = -\frac{1}{\left|\Omega_i^+\right|} \sum_{j \in \Omega_i^+} \log \frac{e^{(\mathbf{z}_i \cdot \mathbf{z}_j)/\tau}}{\sum_{k=1}^{B} \mathbb{I}_{i \neq k} \cdot e^{(\mathbf{z}_i \cdot \mathbf{z}_k)/\tau}}. \tag{1}$$

In this contrastive learning formulation, the set of positive sample pairs of a sample $\mathbf{X}_i$ is represented as Ω_i^+, and feature vectors extracted from various samples $\mathbf{X}_i$ within the sample set in a minibatch $\mathbb{X}^B$ are represented as $\mathbf{Z}$, while B stands for the batch size. The temperature parameter τ manages the magnitude of the loss. This formulation uses $e^{(\mathbf{z}_i \cdot \mathbf{z}_k)/\tau}$ to assess the dissimilarity between the feature vectors of two positive sample pairs and aggregates them to calculate the loss value. This method is commonly used in conventional contrastive learning, where positive sample pairs are typically derived from the same image through different augmentations, with the aim of extracting identical features.

However, in our proposed contrastive learning task, the positive sample pairs, defined under Class-Wise Positive Pairs, frequently consist of different images. Consequently, our objective is to extract similar, as opposed to identical, features from these positive sample pairs. As a result, we use cosine similarity between feature vectors from positive sample pairs as our measure.

The cosine similarity between two feature vectors, $\mathbf{z}_i$ and $\mathbf{z}_j$, is defined as

$$S(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\| \cdot \|\mathbf{z}_j\|}. \tag{2}$$

Therefore, for the entire set of positive sample pairs Ω^+, the overall Contrastive Learning (CL) loss function can be expressed as

$$L_+^{CL}(\mathbf{Z}) = -\sum_{i=1}^{B} \frac{1}{\left|\Omega_i^+\right|} \sum_{j \in \Omega_i^+} \log \frac{e^{\mathrm{S}(\mathbf{z}_i, \mathbf{z}_j)/\tau}}{\sum_{k=1}^{B} \mathbb{I}_{i \neq k} \cdot e^{\mathrm{S}(\mathbf{z}_i, \mathbf{z}_k)/\tau}}, \tag{3}$$

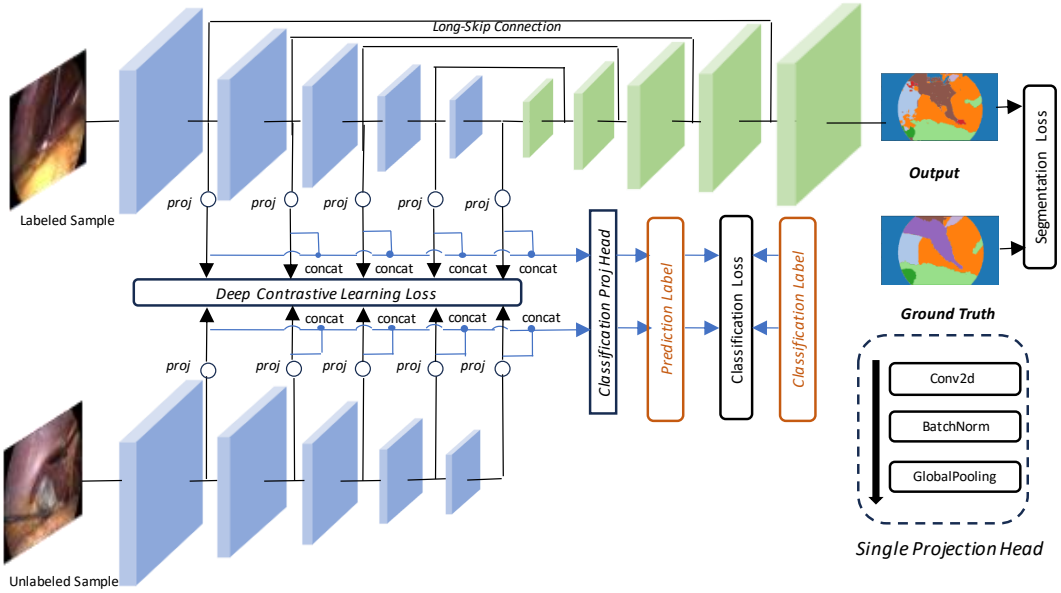

FIGURE 2 The structure of the purposed model. Initially, a set of images is collectively inputted into the system. While some of these images exclusively provide class labels, others supply both class labels and pixel-wise segmentation labels. A consistent encoder is utilized to extract features from this assortment of images, and a projection head is subsequently applied at each scale of the feature map to further extract features. Each "proj" in the diagram denotes a "Single Projection Head", wherein "Conv2d" signifies a 2D convolution with a stride of (1,1) and a kernel size of (1,1). The input dimension corresponds to the number of channels present in the feature map at that specific scale, while the output dimension is configurable. The features derived from distinct images at each scale are employed for contrastive learning. Subsequently, the features extracted at each scale from the same image are concatenated, and a classification projection head (a fully connected layer) is used to predict the target class encompassed within the image. Lastly, a decoder model is utilized to infer segmentation masks on the images within the set that offer segmentation labels. These three tasks are executed concurrently during the training phase. The segmentation model structure used in this experiment will be detailed in 3.2

where B is the minibatch size mentioned in Section 2.1. The MSPH module extracts features from L scales of feature maps. In deep class-wise contrastive learning, the feature vectors from positive sample pairs at the same scale are used for contrastive learning. For a minibatch of B images $\{ \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_B \}$, we apply stochastic data augmentation to each image, resulting in a batch of B images. Denote $\mathbf{Z} = \{ \mathbf{z}_i \mid i = 1, 2, \ldots, B \}$where the $\mathbf{z}_n$ shows the feature extracted by the MSPH for each image in this minibatch, while $\mathbf{z}_i = \{ \mathbf{z}_{i,l} \mid l = 1, 2, \ldots, L \}$ and $\mathbf{z}_{i,l} = g_l(\mathbf{X}_i)$ is the l-th layer's projection head outputs. The loss function for deep contrastive learning (DCL) can be defined as

$$L_{+}^{DCL}(\mathbf{Z}) = -\sum_{i=1}^{B} \frac{1}{|\Omega_i^+|} \sum_{j \in \Omega_i^+} \sum_{l=1}^{L} \log \frac{e^{S(\mathbf{z}_{i,l}, \mathbf{z}_{j,l})/\tau}}{\sum_{k=1}^{B} \mathbb{I}_{i \neq k} \cdot e^{S(\mathbf{z}_{i,l}, \mathbf{z}_{k,l})/\tau}}, \tag{4}$$

where l represents different scales. The Ω_i^+ is the set of indices of positive samples to each image $\mathbf{X}_i$.

In the classification task, we concatenate the multi-scale feature vectors extracted from the MSPH and employ a fully connected layer to predict the target classes present in the image. This portion of the process is depicted in the left half of Fig. 2, where the Classification Proj Head $c(\cdot)$ represents the fully connected layer for the feature vector $\mathbf{u}_i = \text{concat}(\mathbf{z}_{i,1}, \mathbf{z}_{i,2}, \ldots, \mathbf{z}_{i,L})$ extracted by multi-scale projection heads. The loss function for this part is defined as

$$L^{cls}(\mathbf{P}, \mathbf{y}) = -\sum_{i=1}^{B} \sum_{j=1}^{C} y_{ij} \log(p_{ij}) + (1 - y_{ij}) \log(1 - p_{ij}), \tag{5}$$

where C denotes the number of target classes, while B stands for the batch size. The variable $\mathbf{p}$ symbolizes the predicted confidence scores for each class, as produced by the classification projection head. $\mathbf{p}$ is the output from the classification

projection head in a minibatch, $\mathbf{P} = \{ \mathbf{p}_i \mid i = 1, 2, \ldots, B \}$ and $\mathbf{p}_i = c(\mathbf{u}_i)$, and the $p_{i,j}$ represents the probability that the $\mathbf{X}_i$ belongs to the class j, which $\mathbf{p}_i = \{ \mathbf{p}_{i,j} \mid j = 1, 2, \ldots, C \}$.

Simultaneously, $\mathbf{y}$ corresponds to the ground truth labels for the samples and $y_{i,j}$ represents if the $\mathbf{X}_i$ belongs to the class j. The execution of this classification task ensures that the encoder-derived feature vectors embody an adequate amount of information that is specific to the target classes.

2.4 | Segmentation Learning

During the segmentation learning phase, we employ the decoder structure of U-Net to merge the features extracted at diverse scales and infer the segmentation masks. To counteract the problem of class imbalance, we utilize the generalized dice loss and focal loss as loss functions for training our segmentation model[23,24].

The generalized dice loss (GDL), a loss function purpose-built for segmentation tasks, effectively manages class imbalance. Its formula is defined as

$$L^{GDL}(\mathbf{P}, \mathbf{g}) = 1 - \frac{2 \sum_{c=1}^{C} \mathbf{w}_c \sum_{q=1}^{O} p_{cq} g_{cq}}{\sum_{c=1}^{C} \mathbf{w}_c \sum_{q=1}^{O} p_{cq}^2 + \sum_{c=1}^{C} \mathbf{w}_c \sum_{q=1}^{O} g_{cq}^2 + \epsilon}, \tag{6}$$

where $L^{\mathrm{GDL}}(\mathbf{P}, \mathbf{g})$ denotes the Dice loss, C symbolizes the number of classes, and O is the number of pixels in the sample. Denote the segmentation model as $m(\cdot)$ (in this experiment, the segmentation model was set as a U-Net), and each prediction map for input image $\mathbf{X}_i$ should be denoted as $p_i = m(\mathbf{X}_i)$. $\mathbf{g}$ is the ground truth segmentation mask, while p_{cq} and g_{cq} denote the probabilities of pixel q belonging to class c in the predicted and ground truth segmentation masks, respectively. $\mathbf{P} = \{ \mathbf{p}_c \mid c = 1, 2, \ldots, C \}$, and $\mathbf{p}_c = \{ \mathbf{p}_{cq} \mid q = 1, 2, \ldots, O \}$. The weight assigned to class c is represented as $\mathbf{w}_c$, and ϵ is a small constant incorporated for numerical stability.

Focal Loss is another loss function we employ, which addresses the challenges of class imbalance and the varying difficulty of samples. Its formula is defined as

$$L^{Focal}(\mathbf{P}) = -\frac{1}{N} \sum_{q=1}^{O} \left[(1 - p_{cq})^{\gamma} \log(p_{cq}) \right], \tag{7}$$

where $L^{Focal}(\mathbf{P})$ symbolizes the Focal loss, and p_{cq} denotes the probability of pixel q belonging to class c in the predicted segmentation mask. γ is a tunable hyperparameter used to adjust the weight of difficult samples.

We utilize these loss functions as follows during the training of the segmentation model to gauge the discrepancy between the predicted segmentation mask and the ground truth segmentation mask, thereby guiding the optimization of model parameters. Therefore, our loss function is defined as

$$L^{all}(\mathbf{Z}, \mathbf{P}, \mathbf{g}, \mathbf{y}) = \alpha L^{cls}(\mathbf{P}_{cls}, \mathbf{y}) + \beta L_{+}^{DCL}(\mathbf{Z}) + L^{GDL}(\mathbf{P}_{seg}, \mathbf{g}) + L^{Focal}(\mathbf{P}_{seg}), \tag{8}$$

where α and β are hyperparameters, which were set to $\beta = 0.1$ and $\alpha = 0.03/L$ in order to harmonize the losses into a commensurate scale in this experiment. L represents the number of layers with the projection head added to the encoder. p_{seg} shows the prediction of the segmentation mask and the p_{cls} is the prediction of the classification task.

In summary, as shown in Fig. 2, we initially extract a number of images from samples with segmentation annotations. Simultaneously, we randomly select some images from the dataset that includes classification annotations to be used for comparative learning. Both groups of data use the Unet encoder with MPSH to extract features. The features extracted from the two groups of data are used for the deep contrastive learning task. At the same time, we concatenate these features and obtain prediction labels through a classification projection, which is used for the training of the classification task. Finally, the features extracted from the data with segmentation annotations are input into Unet's decoder, outputting segmentation predictions for the training of the segmentation task. It's important to note that the two groups of data share encoder weights, and these three tasks are conducted concurrently.

3 | EXPERIMENTS

3.1 | Dataset

We validated our proposal on the CholecSeg8k dataset[4]. The CholecSeg8k consists of 8080 laparoscopic images, each of the size 854 × 480 pixels, captured from 17 cholecystectomy surgery video clips. These images, originally based on the Cholec80 dataset[25], have been collected and annotated specifically for segmentation tasks. We split the training and test set by video VIDs. VID12 and VID52 were always used as the test set while the remaining data was utilized three times. We excluded some rarely occurring classes and conducted segmentation experiments on a reduced set of 8 classes: "Black Background", "Abdominal Wall", "Liver", "Fat", "Grasper", "Connective Tissue", "Blood", 'L-hook Electrocautery", "Gallbladder', and "Liver Ligament". The multi-class labels are derived from Cholecseg8k[4], while the phase labels are obtained from Cholec80[25] because the images in CholecSeg8k are a subset of Cholec80. We used the full training set and validation set for the classification task and contrastive learning task. Only a small part of the dataset (1%-10%) was used during the segmentation model training.

3.2 | Experimental Settings

3.2.1 | Model Definition

We used U-Net[3] as the segmentation model for all the experiments. U-Net is a traditional framework with an encoder-decoder structure designed for segmentation tasks, where the encoder can be most of the representation learning backbones, such as attention-based backbones[26]. In our experiments, we utilized a five-layer U-Net, with the latest 3 layers undergoing deep supervision for contrastive learning, so the number of multi-scale projection head layers $L = 3$. We conducted experiments utilizing the parameters employed for training the U-Net architecture. The channels of the U-Net were set as 32, 64, 128, 256, 512, and the length of the feature extracted from each projection head was set as 128.

3.2.2 | Data Augmentation

In line with related research[27], we implemented data augmentation during the training phase. Images were resized to dimensions of 427 × 240 pixels for the CholecSeg8k dataset, and a random scaling factor ranging from (-0.1, 0.1) was applied to the input. To inject greater variation into the input frames, we performed Gaussian blurring using kernel sizes of 3, 5, and 7 as the maximum Gaussian kernel size for blurring the input image, and for each kernel size, the sigma was set as 0.8, 1.1 and 1.4. We performed image flipping with a probability of 0.5. The images used for training were subsequently cropped randomly to a size of 256 × 256 pixels.

3.2.3 | Method Training and Evalutaion

Throughout the training process, the batch size was maintained at 200, and each projection head at every scale extracted features with a length of 128. In the loss function, $\gamma = 2$, $\tau = 0.07$[21]. The $\epsilon = 0.00001$ while the class number is $C = 8$ and the batch size N was set as 60. The α and β were set as 0.1 and $0.03/L$. During the inference phase, a sliding window strategy was adopted for segmentation, with a window size of 256 x 256 pixels. The overlay parameter was set to 0.25, indicating a 25% overlap between adjacent windows. For all experiments, Adam with decoupled Weight decay (AdamW)[28] optimizer was employed with an initial learning rate set at 0.01. To optimize the training process, we utilized a ReduceLROnPlateau strategy to dynamically adjust the learning rate based on the model's performance on the validation set. The training epoch for CholecSeg8k is set to be 200 for each percentage of labeled data. Eight NVIDIA A100 80G were used for conducting experiments. The evaluation was based on class-wise IOU. We used VID12 and VID52 from CholecSeg8k as the test set while the remaining 90% of the data was utilized for 3 times validation. We conducted experiments from 1% budget to 10% with a step of 5% on the training set, performed validation on the validation set during cross-validation, and evaluated the model on the test set.

TABLE 1 Class-wise IoU from 1% to 10% labeled data on CholecSeg8k. Results were averaged from 3 runs. The best performance is denoted using bold, and we also mark the best IoU scores for each class at each sample size with underscores. "Ours DCL" refers to the segmentation model training that utilized deep contrastive learning. "Ours, *cls*" signifies the segmentation model training experiment where multi-level features were classified via a class-wise projection head. "Ours DCL+*cls*" designates the experiment where both classification and contrastive learning were employed for segmentation model training.

	Samples	Background	Abdominal Wall	Liver	Fat	Grasper	Connective Tissue	L-hook Electrocautery	Gallbladder
Random [3]	1%	0.953 ± 0.004	0.519 ± 0.009	0.442 ± 0.036	0.761 ± 0.005	0.157 ± 0.006	0.007 ± 0.009	0.032 ± 0.005	0.241 ± 0.021
	5%	0.926 ± 0.034	0.664 ± 0.052	0.501 ± 0.045	0.796 ± 0.024	0.326 ± 0.112	0.265 ± 0.063	0.310 ± 0.123	0.360 ± 0.078
	10%	0.941 ± 0.008	0.612 ± 0.034	0.461 ± 0.023	0.767 ± 0.004	0.226 ± 0.042	0.091 ± 0.120	0.187 ± 0.132	0.335 ± 0.038
SimCLR [12]	1%	0.905 ± 0.012	0.482 ± 0.007	0.431 ± 0.023	0.742 ± 0.004	0.188 ± 0.030	0.040 ± 0.027	0.200 ± 0.057	0.210 ± 0.016
	5%	0.901 ± 0.008	0.558 ± 0.018	0.464 ± 0.023	0.762 ± 0.013	0.303 ± 0.014	0.009 ± 0.006	0.328 ± 0.017	0.324 ± 0.012
	10%	0.932 ± 0.002	0.633 ± 0.015	0.403 ± 0.008	0.780 ± 0.005	0.311 ± 0.030	0.342 ± 0.061	0.311 ± 0.008	0.393 ± 0.027
Ours DCL	1%	0.952 ± 0.005	0.542 ± 0.048	0.519 ± 0.017	0.755 ± 0.000	0.154 ± 0.020	0.096 ± 0.071	0.128 ± 0.148	0.252 ± 0.019
	5%	0.953 ± 0.003	0.606 ± 0.066	0.505 ± 0.038	0.797 ± 0.018	0.300 ± 0.092	0.246 ± 0.139	0.264 ± 0.188	0.374 ± 0.111
	10%	0.936 ± 0.004	0.631 ± 0.024	0.484 ± 0.007	0.792 ± 0.008	0.310 ± 0.059	0.305 ± 0.098	**0.420 ± 0.059**	0.421 ± 0.033
Ours *cls*	1%	0.955 ± 0.000	0.534 ± 0.034	0.494 ± 0.019	0.773 ± 0.016	0.192 ± 0.004	0.053 ± 0.047	0.001 ± 0.001	0.262 ± 0.028
	5%	0.925 ± 0.013	0.576 ± 0.057	0.444 ± 0.014	0.775 ± 0.022	0.245 ± 0.089	0.182 ± 0.079	0.356 ± 0.092	0.320 ± 0.076
	10%	0.945 ± 0.005	**0.657 ± 0.025**	0.485 ± 0.016	**0.796 ± 0.019**	0.341 ± 0.031	0.400 ± 0.133	0.417 ± 0.064	**0.458 ± 0.023**
Ours DCL+*cls*	1%	0.952 ± 0.006	0.614 ± 0.034	0.498 ± 0.019	0.762 ± 0.005	0.178 ± 0.005	0.060 ± 0.035	0.137 ± 0.019	0.287 ± 0.017
	5%	0.944 ± 0.012	0.624 ± 0.023	0.438 ± 0.035	0.798 ± 0.004	0.296 ± 0.070	0.336 ± 0.031	0.338 ± 0.046	0.365 ± 0.027
	10%	**0.948 ± 0.013**	0.638 ± 0.019	**0.485 ± 0.008**	0.791 ± 0.009	**0.407 ± 0.016**	**0.404 ± 0.042**	0.405 ± 0.032	0.447 ± 0.030

3.3 | Result

3.3.1 | Comparative Studies Using 1-Stage Methods

We conducted comparative studies using various contrastive learning methods. The studies included: (1) U-Net with Random Initialization [3], which is a U-Net model trained from scratch, and (2) SimCLR [12], adapted from contrastive learning for natural image classification tasks. SimCLR employs strong random transformations to define positive pairs and trains the encoder model using a contrastive loss. To validate the effectiveness of our proposed modules, we set up three experiments tailored to our techniques. The "Ours DCL" experiment emphasizes training the segmentation model using deep contrastive learning. The second experiment, dubbed "Ours, *cls*", targets the training of a segmentation model that classifies multi-level features via a class-specific projection head. Finally, the "Ours DCL+*cls*" experiment combines both classification and deep contrastive learning techniques for training the segmentation model. Results are presented in Table 1 and Fig. 3.

3.3.2 | Comparative Studies Using 2-Stage Methods

In addition to the primary experiments, we compared our approach with certain 2-stage contrastive learning methods to demonstrate the efficacy and generalizability of our proposed method. These strategies generally pre-train the encoder with unlabeled data and then fine-tune it using labeled data. The methods include: (1) SimCLR 2-stage [12], which employs SimCLR to pre-train the U-Net encoder and subsequently fine-tunes the U-Net using labeled data, and (2) BYOL [14], which utilizes a single sample as both the positive and negative pair for pre-training the U-Net encoder. The outcomes are available in Table 2. For comparison, we structured two experiments around our techniques. The "Ours DCL+*cls*" method amalgamates both classification and deep contrastive learning during the encoder's pre-training phase. Serving as the ablation study, the "Ours DCL" approach harnesses the MSPH module along with our unique definition of positive pairs to pre-train the encoder. The results indicate that our methodology delivers outstanding segmentation performance across multiple targets, substantiating the prowess of our approach.

3.4 | Disscussion

In the majority of categories, methods incorporating MSPH and Deep Class-Wise Contrastive Learning demonstrated superior performance, as evidenced in Table 1. It is clear that for categories with a higher frequency of occurrence, such as the background and abdominal wall, the improvements offered by our method are marginal compared to random selection. However, for

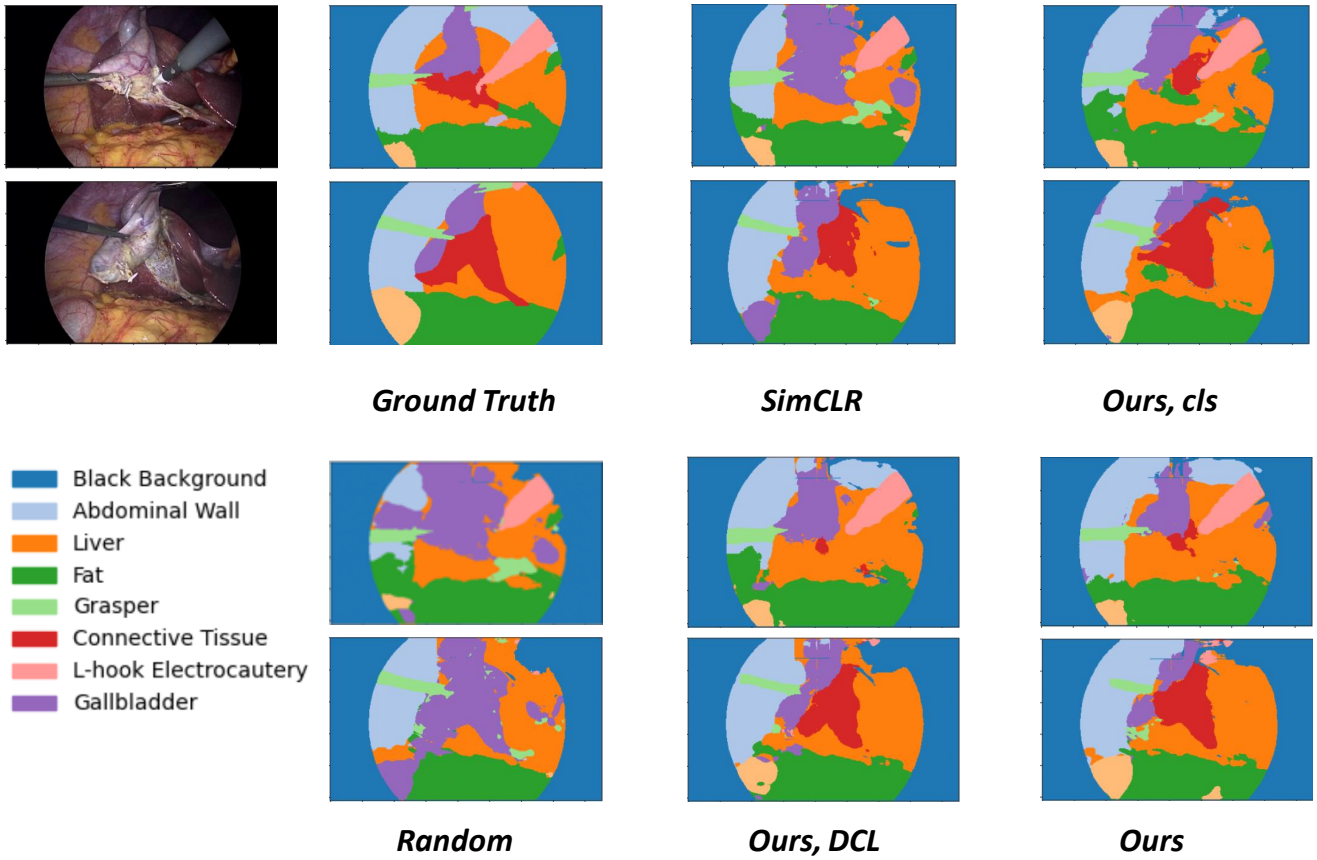

FIGURE 3 Visualization of the segmentation results on CholecSeg8k using 10% samples from the training set. Different colors indicate different classes. "Ours, DC" shows the results of multi-task training using only contrastive learning without the classification projection head. "Ours, cls" shows the results of multi-task training using only the classification projection head. "Ours" presents the results from using the class-wise positive pairs defining method mentioned in section 2.1 to divide the sample pairs and training with the deep class-wise contrastive learning method.

categories with lower occurrence frequencies, such as clamps, connective tissues, and L-hook electrosurgery, our approach delivers substantial advancement. This is mainly attributable to our method's capability to incorporate images with solely classification labels for contrastive learning, which allows us to optimize the encoder and the segmentation task simultaneously. This strategy enables the encoder to capture more similar features from images within the same target category. Simultaneously, through a class-wise projection head, we aim to guide the encoder to place greater emphasis on the extraction of features related to the target objects in our segmentation task.

Fig. 3 effectively demonstrates this point. In the sample image, related methods struggle to distinguish between connective tissue and gallbladder due to their high similarity and the limited sample size, often leading to the incorrect categorization of Connective Tissue. By incorporating a multi-scale projection head for deep contrastive learning, the accuracy of object boundaries has improved. However, it is evident that the amount of target information within each image is still not sufficiently explicit. However, by integrating the classification projection head, our encoder is able to differentiate between the features of Connective Tissue and Gallbladder. This issue is prevalent in the segmentation of endoscopic video images, a challenge that is widely recognized within the field. As evidenced in the relevant literature [1], some researchers have relied solely on the U-Net architecture for the segmentation task. In the absence of contrastive learning methodologies, and owing to the relatively small volume of training samples, the model can easily misidentify gallbladder structures and connective tissues. Despite the segmentation being imperfect in some pixel areas, our method can successfully segment these two closely related targets.

Furthermore, when compared with related methods, the results illustrate that approaches like SimCLR might achieve minor improvements with limited data (e.g., less than 1% of the entire dataset) due to their ability to acquire a few target features

TABLE 2 The Class-wise IoU by fine-tuning the U-Net using 5% of the labeled data. Results are averaged across 3 runs. The best performance is denoted using bold and under scores

	Background	Abdominal Wall	Liver	Fat	Grasper	Connective Tissue	L-hook Electrocautery	Gallbladder
SimCLR 2-stage [12]	0.921 ± 0.012	0.653 ± 0.005	**0.528 ± 0.021**	0.773 ± 0.013	0.250 ± 0.020	0.206 ± 0.038	0.360 ± 0.021	0.389 ± 0.012
BYOL [14]	0.916 ± 0.015	0.651 ± 0.027	0.520 ± 0.025	0.776 ± 0.011	0.250 ± 0.014	0.205 ± 0.037	0.359 ± 0.040	**0.419 ± 0.018**
Ours DCL	**0.933 ± 0.006**	0.657 ± 0.034	0.522 ± 0.010	0.779 ± 0.001	0.252 ± 0.015	**0.219 ± 0.037**	0.354 ± 0.026	0.413 ± 0.050
Ours DCL+*cls*	0.924 ± 0.005	**0.659 ± 0.022**	0.519 ± 0.015	**0.779 ± 0.014**	**0.266 ± 0.008**	0.212 ± 0.026	**0.368 ± 0.032**	0.407 ± 0.024

via contrastive learning across the entire dataset. However, as the volume of data used for training progressively increases, it becomes evident that the features learned by these methods do not significantly assist the training process. This is mainly because these methods can only generate positive sample pairs from a single image, thereby overlooking the similar semantic information contained within different images in laparoscopic surgery. This reaffirms our proposition in Section 2.1, which states that defining positive sample pairs as images with identical labels and evaluating the features extracted from these pairs with a cosine similarity function is beneficial. Ablation studies further reveal that during the contrastive learning process of positive sample pairs, the incorporation of a classification task can enhance the correlation between the features extracted by the encoder and our target categories, thereby boosting the accuracy of the overall segmentation task.

In order to substantiate the efficacy of our proposed novel positive sample pair partitioning method and the MSPH module, we orchestrated an additional experiment: initially, the encoder part was trained using a contrastive learning approach, and then the weights of the encoder were frozen. Following this, finetuning was performed using 5% of the original dataset that contained segmentation labels. The results are elucidated in Table 2. In this context, we employed the SimCLR and BYOL methods as comparative experiments. From the outcome, it is discernible that our approach generally outperforms traditional methodologies, especially in enhancing smaller target segmentation, where the improvements are more pronounced. This substantiates our hypothesis that, within datasets like laparoscopic surgical video images where inter-image similarity is high, utilizing images with analogous features as positive sample pairs proves to be more proficient in extracting generalized target features compared to using a single image as a positive sample pair. Concurrently, when classification labels are leveraged as auxiliary, the extracted features for various targets are more precise, imparting the model with increased robustness.

4 | CONCLUSION

In this study, we implemented several specific improvements to the process of training an endoscopic surgery image segmentation model with limited samples. Given the scarcity of samples, we designed a multi-task training method grounded in contrastive learning to maximize the utilization of available data. The strength of this approach lies in its ability to optimize the features extracted by the model's encoder. Since image classification labels are more readily available than segmentation labels, we utilized the model's classification labels to distinguish between positive sample pairs, facilitating contrastive learning. As these positive pairs do not stem from the same image, we incorporated the cosine similarity function to assess the similarity of features extracted from the positive samples. This method ensured that the encoder could accurately extract the features of the target object during contrastive learning. To enhance model accuracy, we used classification labels to categorize the features extracted from the encoder. Experiments conducted on the CholecSeg8k dataset have confirmed the effectiveness of these improvements.

- Due to the use of contrastive learning in the training structure, this method is significantly slower than direct training. Moreover, considering the training speed, we did not adopt the EMA approach to differentiate the encoder. Since our method divides positive sample pairs within the same minibatch for contrastive learning, a large batch size is required to provide a sufficient number of positive sample pairs. If we introduce a memory bank or similar feature storage for contrastive learning, it may alleviate this problem.
- The categories in the dataset are not diverse enough. The majority of annotated samples provided in the CholecSeg8k dataset only exist in the preparation phase, while the sample size for other stages of surgery is extremely insufficient. In subsequent experiments, manually annotating sample pairs or expanding labels through active learning might be a very effective approach.
- This method still requires a large number of manually annotated classification samples, which can increase annotation costs. Therefore, utilizing some self-supervised learning methods that can extract sufficient information from the image itself for segmentation, such as MAE [29], SimMIM [30], would be a promising strategy.

In future work, we aim to address the following:

- Optimize the contrastive learning training structure to accelerate the process.
- Increase the variety of surgical stages in our dataset through active learning or manual annotations.
- Explore self-supervised learning methods to lessen the need for manually annotated classification samples, thus reducing annotation costs.

Through these advancements, we seek to refine our endoscopic surgery image segmentation model.

ACKNOWLEDGMENTS

This work was supported by the JST Moonshot R&D grant number JPMJMS2033; the MEXT/JSPS KAKENHI under grant numbers 21K19898, 26108006, 17H00867; and the JST CREST grant number JPMJCR20D5, Japan.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

1. Silva B, Oliveira B, Morais P, et al. Analysis of Current Deep Learning Networks for Semantic Segmentation of Anatomical Structures in Laparoscopic Surgery. In: 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE. 2022:3502–3505.
2. Grammatikopoulou M, Sanchez-Matilla R, Bragman F, et al. A spatio-temporal network for video semantic segmentation in surgical videos. *International Journal of Computer Assisted Radiology and Surgery.* 2023:1–8.
3. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 Proceedings, LNCS 9351. Springer. 2015:234–241.
4. Hong WY, Kao CL, Kuo YH, Wang JR, Chang WL, Shih CS. Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on Cholec80. *arXiv preprint arXiv:2012.12453.* 2020.
5. Chen T, Kornblith S, Swersky K, Norouzi M, Hinton G. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029.* 2020.
6. Chen X, Fan H, Girshick R, He K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297.* 2020.
7. Misra I, Maaten Lvd. Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020:6707–6717.
8. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722.* 2019.
9. Chen X, Fan H, Girshick R, He K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297.* 2020.
10. Wu Z, Xiong Y, Yu SX, Lin D. Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:3733–3742.
11. Klinker F. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte.* 2011;58:97–107.
12. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: PMLR. 2020:1597–1607.
13. Gupta K, Ajanthan T, Hengel v. dA, Gould S. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491.* 2022.
14. Grill JB, Strub F, Altché F, et al. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems.* 2020;33:21271–21284.
15. Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057.* 2021.
16. Caruana R. Multitask Learning. 1998:95–133. doi: 10.1007/978-1-4615-5529-2$_5$
17. Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. 2015:1440–1448.
18. Zhang Z, Yang L, Zheng Y. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:9242–9251.
19. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2017;40(4):834–848.
20. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203.* 2021.
21. Zhang L, Chen X, Zhang J, Dong R, Ma K. Contrastive deep supervision. In: European Conference on Computer Vision, LNCS 13686. Springer. 2022:1–19.
22. Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. *Advances in Neural Information Processing Systems.* 2020;33:18661–18673.
23. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision, LNCS 10553. Springer. 2017:240–248.
24. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. 2017:2980–2988.
25. Twinanda A, Shehata S, Mutter D, Marescaux J, Mathelin MD, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging.* 2016;36.
26. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.* 2020.

27. Qiu J, Hayashi Y, Oda M, Kitasaka T, Mori K. Class-wise confidence-aware active learning for laparoscopic images segmentation. *International Journal of Computer Assisted Radiology and Surgery.* 2023;18(3):473–482.
28. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101.* 2017.
29. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. *arXiv:2111.06377.* 2021.
30. Xie Z, Zhang Z, Cao Y, et al. SimMIM: A simple framework for masked image modeling. In: International Conference on Computer Vision and Pattern Recognition (CVPR). 2022.