Multi-view Masked Contrastive Representation Learning for Endoscopic Video Analysis

Kai Hu
Xiangtan University
kaihu@xtu.edu.cn

Ye Xiao
Xiangtan University
yxiao@smail.xtu.edu.cn

Yuan Zhang*
Xiangtan University
yuanz@xtu.edu.cn

Xieping Gao*
Hunan Normal University
xpgao@hunnu.edu.cn

Abstract

Endoscopic video analysis can effectively assist clinicians in disease diagnosis and treatment, and has played an indispensable role in clinical medicine. Unlike regular videos, endoscopic video analysis presents unique challenges, including complex camera movements, uneven distribution of lesions, and concealment, and it typically relies on contrastive learning in self-supervised pretraining as its mainstream technique. However, representations obtained from contrastive learning enhance the discriminability of the model but often lack fine-grained information, which is suboptimal in the pixel-level prediction tasks. In this paper, we develop a **M**ulti-view **M**asked **C**ontrastive **R**epresentation **L**earning (M^2CRL) framework for endoscopic video pre-training. Specifically, we propose a multi-view masking strategy for addressing the challenges of endoscopic videos. We utilize the frame-aggregated attention guided tube mask to capture global-level spatiotemporal sensitive representation from the global views, while the random tube mask is employed to focus on local variations from the local views. Subsequently, we combine multi-view mask modeling with contrastive learning to obtain endoscopic video representations that possess fine-grained perception and holistic discriminative capabilities simultaneously. The proposed M^2CRL is pre-trained on 7 publicly available endoscopic video datasets and fine-tuned on 3 endoscopic video datasets for 3 downstream tasks. Notably, our M^2CRL significantly outperforms the current state-of-the-art self-supervised endoscopic pre-training methods, *e.g.*, Endo-FM (3.5% F1 for classification, 7.5% Dice for segmentation, and 2.2% F1 for detection) and other self-supervised methods, *e.g.*, VideoMAE V2 (4.6% F1 for classification, 0.4% Dice for segmentation, and 2.1% F1 for detection). [2]

1 Introduction

Video endoscopy is a crucial medical examination and diagnostic tool widely used for inspecting various tissues and structures (the digestive tract, respiratory tract, and abdominal cavity, *etc.*) [1]. In clinical practice, endoscopic video analysis usually relies on the experience and expertise of physicians, which is not only time-consuming and labor-intensive but also prone to subjective errors. Computer-aided medical analysis [2, 3, 4] can automatically and efficiently identify and classify

*Corresponding authors.

[2]Code is publicly available at: https://github.com/MLMIP/MMCRL.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

lesions, thereby assisting physicians in making more accurate diagnoses. In this paper, we focus on endoscopic videos with the aim of developing a robust pre-trained model for endoscopic video analysis to facilitate downstream tasks (*i.e.*, classification, segmentation, and detection).

Yann LeCun has mentioned "the revolution will not be supervised [5]" in multiple talks, emphasizing that the future development of artificial intelligence will increasingly rely on un-/self-supervised learning. Among them, self-supervised learning (SSL) aims to learn scalable visual representations from large amounts of unlabelled data for downstream tasks with limited annotated data. To learn meaningful representations for SSL, researchers crafted visual pretext tasks [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16], which are summarized into two main categories: contrastive and generative [17]. Contrastive methods, also known as discriminative methods, employ a straightforward discriminative idea that pulling closer representations from the same image and pushing away different images, *i.e.*, contrastive learning (CL) [6, 7, 8]. By utilizing image-level prediction with global features, CL can naturally endow pre-trained models with strong instance discriminability, which has been proven to be effective in classification tasks. However, CL also presents the challenge that downstream dense prediction tasks, such as segmentation and detection, are not fully considered.

Generative methods aim to reconstruct the input data itself by encoding the data into features and then decoding it, including AE [18], VAE [19], GAN [20], *etc*. Recently, masked image modeling (MIM) [13, 14, 16] has demonstrated the strong potential in self-supervised learning. MIM masks a substantial portion of image patches during training and utilizes an autoencoder [18] to reconstruct the original signal of the image, which, unlike CL, can enhance the ability to capture the pixel-level information. Following the success of MIM, some works have tried to extend this new pre-training paradigm to the video domain for self-supervised video pre-training [21, 22, 23]. Actually, mask techniques [24, 25, 26] are crucial for the success of mask modeling. Endoscopic videos, in particular, have higher dimensions and redundancy compared to static images. They also exhibit unstable inter-frame variations due to the manual manipulation by the doctor. Additionally, lesions in endoscopic videos often have low contrast and appear in obscured regions or with subtle variations. Therefore, simply applying random mask not only requires extensive pre-training time but also easily leads to poor performance.

To address the aforementioned issues, in this paper, we develop a Multi-view Masked Contrastive Representation Learning framework named M^{2}CRL. **First**, considering the characteristics of inter-frame instability and small inter-class differences in endoscopic videos, we propose a multi-view masking strategy. Specifically, we introduce a frame-aggregated attention guided tube masking strategy for the global views, which aggregates features from multiple frames to capture global spatiotemporal information. Simultaneously, a random tube masking strategy is employed from the local views, enabling the model to focus on local features. **Second**, to address the inadequacy of capturing pixel-level details in contrastive learning, we integrate multi-view masked modeling into contrastive approach, which not only encourages the model to learn discriminative representations but also forces it to capture more refined pixel-level features. Extensive experiments have verified that our M^{2}CRL significantly enhances the quality of endoscopic video representation learning and exhibits excellent generalization capabilities in multiple downstream tasks (*i.e.*, classification, segmentation and detection). Overall, our contributions are summarized as follows:

- We propose a novel multi-view masking strategy aimed at enhancing the capture of fine-grained representations in endoscopic videos by performing mask modeling on both global and local views. This strategy involves utilizing the frame-aggregated attention guided tube mask to capture global-level spatiotemporal contextual relationships from the global views, while employing the random tube mask to focus on local variations from the local views.

- We propose a multi-view masked contrastive representation learning framework that combines multi-view mask modeling with contrastive method to train endoscopic videos, which effectively addresses the limitation of contrastive method in capturing dense pixel dependencies by predicting the intensity of each pixel within masked patches.

- We conduct extensive experiments on 10 endoscopic video datasets to evaluate the performance of M^{2}CRL in comparison to other methods. M^{2}CRL achieves 94.2% top-1 accuracy on PolypDiag [27], 81.4% on CVC-12k [28], and 86.3% on KUMC [29], outperforming the state-of-the-art methods, *i.e.*, Endo-FM [30] by +3.5%, +7.5%, and +2.2%, and VideoMAE V2 by +4.6%, +0.4%, and +2.1%, respectively.

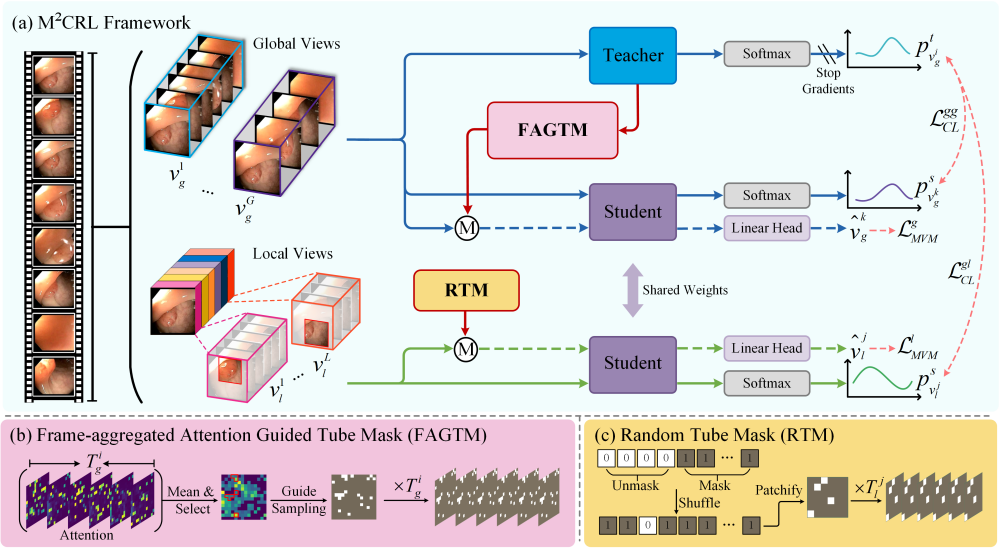

Figure 1: The pipeline of the proposed M^{2}CRL. For the generated global and local views with different frame rates and spatial sizes, we adopt two different mask strategies: the frame-aggregated attention guided tube mask and the random tube mask. These strategies are integrated with mask reconstruction and contrastive method, enabling the model to simultaneously learn both the pixel-level and discriminative features of the video.

2 Related Work

2.1 Self-supervised video representation learning

Self-supervised learning is a machine learning technique that mitigates the reliance on manual labeling by exploiting the inherent structure or properties of the data itself, and its common goal in computer vision is to learn scalable and generalisable visual representations. In SSL, the key lies in designing an appropriate pretext task, which involves generating suitable supervision using only visual signals. Pretext tasks in images [9, 10, 11, 12] have been widely explore, and motivated by images, there has been great interest in exploring SSL pretext tasks for videos. Some early works [31, 32, 33] focused on extending image SSL methods for video, however video has a specific temporal dimension. The temporal information can be leveraged for representation learning, including future prediction [34, 35, 36, 37, 38, 39, 40], temporal ordering [41, 42, 43, 44, 45], object motion [46, 47, 48, 49], and temporal consistency [50, 51]. In the last few years, contrastive learning [52, 53, 54, 55, 56, 57] has made great progress in SSL. However, SSL based on contrastive methods typically focuses on discriminative features of holistic information, while lacking the ability to focus on fine-grained features.

2.2 Masked visual modeling

Masked visual modeling is proven to be a simple and effective self-supervised learning paradigm [13, 14, 15, 16, 58, 21, 22] through a straightforward pipeline based on masking and reconstruction. This paradigm quickly expanded to the video domain, *e.g.*, VideoMAE [22] and ST-MAE [21], which extend image MAE to video. As with images, the choice of masking strategy affects self-supervised representation learning [24, 25, 26, 59]. VideoMAE [22] employs random tube mask to model the same locations across frames to ensure spatial discontinuity while maintain temporal continuity. In contrast, ST-MAE [21] generates a random mask for each frame independently, which are discontinuous in both space and time. Nevertheless, mask video modeling using randomly masked tokens to reconstruct is inefficient because the tokens embedded in a video frame are not equally important. Several studies [60, 61, 62, 63] have proposed various motion-based video mask strategies. However, our focus is on endoscopic videos, which lack a strong motion counterpart compare to the above works.

2.3 Self-supervised learning for endoscopy

In recent years, SSL has received increasing attention in the field of medical analysis, including endoscopy. Endo-FM [30] employs contrastive method to minimize the disparity in feature representations between different spatiotemporal views of the same video. Intrator [64] has proposed the use of contrastive learning to adapt video inputs for appearance-based object tracking. Hirsch [65] applies the SSL framework masked siamese networks (MSNs) to analyze endoscopic videos. While MSNs use a masking concept, it primarily serves as a data augmentation technique, with its essence still rooted in contrastive learning. Currently, most self-supervised pre-training for endoscopic videos relies on contrastive methods. While these methods have shown promise for endoscopic video pre-training, relying solely on them may not fully capture the fine feature expressions of endoscopic videos.

3 Method

To address the limitations of contrastive methods in fine-grained information perception in SSL for endoscopic videos, we propose a Multi-view Masked Contrastive Representation Learning (M^{2}CRL) framework for endoscopic video pre-training, as shown in Fig. 1. Here, we first review masked prediction in Section 3.1. Then, we present our proposed multi-view masking strategy in Section 3.2 and introduce our M^{2}CRL framework in Section 3.3.

3.1 Preliminary

Masked prediction is a prevalent representation learning technique in natural language processing (NLP) [66, 67, 68, 69], and many researchers have explore its application to images [14, 16, 70, 71, 72] and videos [21, 22, 23]. MIM endeavors to learn image representations by solving a regression problem, where the model is tasked with predicting the pixel values in randomly masked patches of the image. Specifically, an image $x \in \mathbb{R}^{H \times W \times C}$ is reshaped into $N = HW \big/ P^2$ flattened patches as $\{x_i\}_{i=1}^{N}$, where (H, W) is the resolution of image, C is the number of channels and P is the patch size. Each patch is represented with token embedding. MIM constructs a random mask $m \in \{0, 1\}^N$ to indicate the masked tokens that correspond to $m_i = 1$. In MAE [14], only the visible tokens $\{x_i \,|m_i = 0\}_{i=1}^{N}$ are fed into the vision transformer to obtain the latent feature, and then the decoder uses the latent feature and the masked tokens as inputs to predict $\hat{y}$. In SimMIM [16], visible and invisible tokens are fed into the encoder. The prediction loss is calculated as the loss between the normalized masked tokens and the reconstructed ones in the pixel space by:

$$\mathcal{L} = \frac{1}{N_m} \sum \left\| \hat{y}_m - x_m \right\|_p \tag{1}$$

where N_m is the number of masked tokens, x_m is the masked token, p is norm and its value is 1 or 2.

3.2 Multi-view masking strategy

Masked video modeling (MVM) [21, 22, 23] employs random mask strategies (*i.e.*, random, spatial-only, temporal-only) to capture meaningful representations from pre-training videos. Although these strategies are effective for general video datasets with well-curated and stable distributions, they do not account for the unique characteristics of medical data. We have summarized two key characteristics of endoscopic videos: (1) Instability of inter-frame variations is a prominent feature of endoscopic videos. These variations are driven by camera movement, instrument manipulation, and the uneven distribution of lesion areas, *e.g.*, variations can range from drastic to minor, as the camera navigates from the intestinal wall to specific lesion sites. (2) Endoscopic video exhibits characteristics of small inter-class differences. The lesion tissues typically resemble surrounding the normal tissues in color, texture, or shape, which complicates the model's ability to accurately identify the lesion area. Therefore, we propose a multi-view masking strategy by considering the above two points, the details are as follows.

3.2.1 Frame-aggregated Attention Guided Tube Mask

To address the challenge of instability between video frames, we propose a frame-aggregated attention guided tube masking strategy. We aggregate the attention of all frames of the video along the frame

dimension to generate a frame-aggregated attention map, which then dynamically guides the masking process. This way can capture the overall scene information in a video sequence from the global spatiotemporal information and ignores irrelevant spatiotemporal noise to some extent.

Semantic information extraction Our architecture consists of the teacher network f_t and the student network f_s. Our network employs a self-attention mechanism known as divided space-time attention mechanism [73], which enhances the learning ability of the network while reducing the computational complexity. Specifically, we take an endoscopic video, from which we sample the global views $\{v_g^i \in \mathbb{R}^{T_g^i \times 3 \times H_g \times W_g}\}_{i=1}^{G}$, where T_g is the number of frames in the sampled view. Each frame is then divided into $N = H_g W_g / P^2$ patches, which are mapped into patch tokens and fed into the transformer blocks of the teacher network. Thus, each encoder block processes N patch (spatial) and T_g temporal tokens. The network includes a learnable class token, $[cls]$, which represents the global features learned by the network along spatial and temporal dimensions. Given the intermediate token $e^u \in \mathbb{R}^{(N+1)\times D}$ from block u, the token in the next block is computed as follows:

$$\begin{cases} e_{time}^{u+1} = MSA_{time}\left(LN\left(e^u\right)\right) + e^u, \\ e_{space}^{u+1} = MSA_{space}\left(LN\left(e_{time}^{u+1}\right)\right) + e_{time}^{u+1}, \\ e^{u+1} = MLP\left(LN\left(e_{space}^{u+1}\right)\right) + e_{space}^{u+1} \end{cases} \tag{2}$$

where MSA, LN and MLP denote the multi-head self-attention, layer normalization, and multi-layer perceptron, respectively. Each block utilizes MSA layer to project and divide e into n_h parts. Each part contains the query Q_r, key K_r and value V_r for r = 1, 2, ..., n_h, where n_h denotes the number of heads. We can get the attention map of the last layer of blocks by calculating the correlation between the query embedding of class token Q^{cls} and key embeddings of all other patches K. It is averaged for all heads as follows:

$$A = \frac{1}{n_h} \sum_{r=1}^{n_h} Softmax(Q_r^{cls} \frac{K_r}{\sqrt{D_h}}) \tag{3}$$

where $D_h = D/n_h$, and A is T_g spatial attention maps. Although single-frame spatial attention integrates temporal information, it still only considers the spatial information of the current frame, neglecting the global spatiotemporal dependencies in the video sequence. Thus, we aggregate the attention of T_g frames to the mean value to obtain a simplified and holistic attention distribution by:

$$A_{agg} = \frac{1}{T_g} \sum_{t=1}^{T_g} A_t \tag{4}$$

This attention mechanism is capable of obtaining an approximation of the critical region of the video that are being attended to from both the temporal and spatial dimensions, while reducing the impact of excessive variations in individual frames or regions. We will further exploit this attention dynamic to guide the generation of tube masking to help the model perform the reconstruction task more appropriately.

Visible tokens sampling and masking The traditional random masking strategy treats critical and non-critical areas of the video equally in each iteration, which may lead to excessive masking of key video regions at a high masking ratio, thereby affecting the learning ability of the model. Therefore, we utilize a frame-aggregated attention mechanism to guide the generation of tube mask. By sampling some reasonably visible tokens from the model-focused areas and masking the rest, it allows our method to efficiently perform the reconstruction task even at a high masking ratio, while improving pre-training efficiency. Specifically, we begin by ranking tokens in descending order of attention scores and subsequently select a proportion of high-attention tokens based on the threshold γ. We randomly sample visible tokens to form a binary mask from selected tokens of attention, as depicted in Fig. 1(b). The number of sampled visible tokens N_v is determined by the predefined mask ratio $\rho \in (0, 1)$. Followed by SimMIM [16], we fill the mask tokens with learnable mask embeddings, which are subsequently fed together into the student network for feature mapping and finally into the prediction head for reconstruction.

In comparison to random sampling, which is inefficient in token allocation, our method selects visible tokens based on the global spatiotemporal information of the video sequence. This approach

Table 1: Comparison with other latest SOTA methods on 3 downstream tasks. We report F1 score (%) for PolypDiag, Dice (%) for CVC-12k, and F1 score (%) for KUMC, respectively.

Method	Venue	Year	Pretrain Time(h)	PolypDiag (Classificaton)	CVC-12k (Segmentation)	KUMC (Detection)
Scratch (Rand.init.)	-	-	N/A	83.5 ± 1.3	53.2 ± 3.2	73.5 ± 4.3
TimeSformer [73]	ICML	2021	104.0	84.2 ± 0.8	56.3 ± 1.5	75.8 ± 2.1
CORP [74]	ICCV	2021	65.4	87.1 ± 0.6	68.4 ± 1.1	78.2 ± 1.4
FAME [75]	CVPR	2022	48.9	85.4 ± 0.8	67.2 ± 1.3	76.9 ± 1.2
ProViCo [76]	CVPR	2022	71.2	86.9 ± 0.5	69.0 ± 1.5	78.6 ± 1.7
VCL [77]	ECCV	2022	74.9	87.6 ± 0.6	69.1 ± 1.2	78.1 ± 1.9
ST-Adapter [78]	NeurIPS	2022	8.1	84.8 ± 0.7	64.3 ± 1.9	74.9 ± 2.9
VideoMAE [22]	NeurIPS	2022	25.3	91.4 ± 0.8	80.9 ± 1.0	82.8 ± 1.9
Endo-FM [30]	MICCAI	2023	20.4	90.7 ± 0.4	73.9 ± 1.2	84.1 ± 1.3
DropMAE [79]	CVPR	2023	37.9	88.2 ± 0.8	80.9 ± 0.3	81.7 ± 2.6
VideoMAE V2 [23]	CVPR	2023	17.3	89.6 ± 1.4	81.0 ± 0.4	84.2 ± 1.0
M^{2}CRL	Ours	-	24.3	**94.2 ± 0.7**	**81.4 ± 0.8**	**86.3 ± 0.8**

minimizes redundant background area sampling, as these backgrounds have minimal impact on the significance of mask reconstruction. Furthermore, the attention-guided mask generation is dynamic and allows for adjustments during model training. This adaptability enables the model to continuously optimize its focus, adapting to complex or changing data characteristics.

3.2.2 Random Tube Mask

Global view mask modeling is primarily employed for a comprehensive understanding and spatiotemporal contextual awareness of the entire endoscopic video frame, which mitigates the effects of inter-frame variability instability by aggregating frame attention. However, due to the low contrast between lesions and normal tissues in endoscopic videos, global views may struggle to accurately capture local details. Hence, we apply random tube mask reconstruction on local views to learn more granular detail information, as shown in Fig. 1(c). Specifically, we obtain the local views $\{v_l^j = \mathbb{R}^{T_l^i \times 3 \times H_l \times W_l}\}_{j=1}^{L}$ $(T_l < T_g)$ by random cropping and uniform sampling at different frame rates. In local views, we also implement a high masking ratio ρ = 90% to reduce information leakage during the mask modeling process. By local view mask modeling with different frame rates and spatial cropping, it is possible to make the model proficient in capturing variations in time scale and spatial detail. Moreover, local view mask modeling focuses on specific regions in the video without interference from the global background, enables more targeted learning. This allows the model to finely capture local information within the video, enhancing its ability to recognize subtle differences between abnormalities and normal tissues.

3.3 Multi-view Masked Contrastive Representation Learning

The pipeline of our proposed M^{2}CRL is shown in Fig. 1(a), which introduces a multi-view masking strategy and combines multi-view mask modeling with contrastive learning to learn representations that possess fine-grained and discriminative capabilities simultaneously.

In DINO [55], self-distillation is proposed not from the posterior distribution but by a teacher-student scheme that extracts knowledge from the model's own past iterations. This self-distillation method of self-supervision is also considered a form of contrastive learning [80]. The contrastive learning part of our M^{2}CRL follows Endo-FM [30], which also employs self-distillation method to achieve representation learning. Given an endoscopic video, two types of views are created under random data augmentation (G global views $\{v_g^i\}_{i=1}^{G}$ and L local views $\{v_l^j\}_{j=1}^{L}$). The model is encoded by two encoders, a teacher network f_t and a student network f_s, which are respectively parameterized by θ_t and θ_s. It should be noted that the two student networks depicted in Fig. 1(a) actually represent

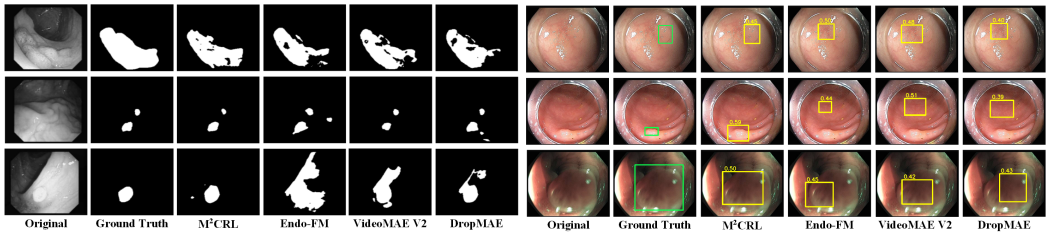

Figure 2: Qualitative results of segmentation and detection tasks. The segmentation results on the left are from the CVC-12k dataset, while the detection results on the right are from the KUMC dataset.

a single student network. We illustrate two student networks in the figure to more clearly convey the data flow. During pre-training, the global views are input into both the teacher and student networks, while the local views are only input into the student network. The network output f is normalized by a softmax function with a temperature τ to obtain the probability distribution $p = \text{softmax}(f/\tau)$. Subsequently, the cross-entropy loss function is used to compute the losses between the teacher's global views and the student's global views, as well as between the teacher's global views and the student's local views. The specific loss functions are as follows:

$$\begin{aligned} \mathcal{L}_{CL}^{gg} &= \sum_{i=1}^{G} \sum_{\substack{k=1 \\ k \neq i}}^{G} -p_{v_g^i}^t \log p_{v_g^k}^s, \\ \mathcal{L}_{CL}^{gl} &= \sum_{i=1}^{G} \sum_{j=1}^{L} -p_{v_g^i}^t \log p_{v_l^j}^s \end{aligned} \tag{5}$$

Contrastive learning utilizes class tokens [55] from global and local views to calculate matching losses, which offers strong capabilities for overall discriminative representation. However, this method overlooks the dense pixel dependencies that are crucial for dense prediction tasks like segmentation and detection. Therefore, we integrate the multi-view mask reconstruction pre-training task into the contrastive learning. In the multi-view mask modeling task, the video clips processed through masking are fed into the encoder to be mapped into the feature space. Subsequently, the prediction head utilizes the context of unmasked patches to regress the dense pixel intensities within masked patches. The losses for global and local view reconstruction are as follows:

$$\begin{aligned} \mathcal{L}_{MVM}^{g} &= \frac{1}{N_m^g} \sum_{i=1}^{G} \left\| \hat{v}_g^i - v_g^i \right\|, \\ \mathcal{L}_{MVM}^{l} &= \frac{1}{N_m^l} \sum_{j=1}^{L} \left\| \hat{v}_l^j - v_l^j \right\| \end{aligned} \tag{6}$$

M^{2}CRL exploits both contrastive loss and reconstruction loss in optimization, and the total loss is $\mathcal{L}_{total} = \mathcal{L}_{CL}^{gg} + \mathcal{L}_{CL}^{gl} + \mathcal{L}_{MVM}^{g} + \mathcal{L}_{MVM}^{l}$. The student network updates the student parameters θ_s by minimizing $\mathcal{L}_{total}$, and the teacher network θ_t is updated using the exponential moving average (EMA) of the student weights, and $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\,\theta_s$. Here λ denotes the momentum coefficient. By combining contrastive learning and masked video modeling, the model is encouraged to learn the holistic discriminative representation and detailed pixel information, effectively improving the learning ability of the model in complex visual data of endoscopic videos.

4 Experiments

4.1 Datasets and Experimental Settings

We conduct experiments on 10 publicly available endoscopic video datasets: Colonoscopic [81], SUN-SEG [82], LDPolypVideo [83], Hyper-Kvasir [84], Kvasir-Capsule [85], CholecTriplet [86], Renji-Hospital [30], PolypDiag [27], CVC-12k [28], and KUMC [29]. These datasets have a total of 33,231 videos with approximately 5,500,000 frames, covering 3 types of endoscopy examination protocols and 10+ different diseases. The videos are processed into 30fps clips with an average duration of 5 seconds. The first 7 datasets are used for pre-training and we sample $G = 2$ global views

and $L = 8$ local views, where $T_g \in [8, 16]$, $T_l \in [2, 4, 8, 16]$ and the spatial size set to 224×224 and 96×96, respectively. We take ViT-B/16 [87] as the backbone and perform 30 epochs of pre-training. In downstream tasks, we perform classification task on PolypDiag, segmentation task on CVC-12k, and detection task on KUMC, respectively. Our implementation is based on Endo-FM [30], and more experimental details can be found in § A.

4.2 Comparison with Prior Work

We compare our method with the recent state-of-the-art (SOTA) endoscopic video pre-training model, Endo-FM [30], which is the first pre-training model on a large scale across various endoscopic videos. The results of other methods are taken from the comparative method of Endo-FM, including TimeSformer [73], CORP [74], FAME [75], ProViCo [76], VCL [77], and ST-Adapter [78]. Additionally, we also compare the latest video self-supervised methods: VideoMAE [22], VideoMAE V2 [23] and DropMAE [79]. For a fair comparison, all methods are pretrained on the same union of 7 datasets as our M^{2}CRL. All experimental settings are referred to those documented in the original papers or in the released code.

Quantitative evaluation We observe that our method outperforms existing state-of-the-art methods, as shown in Table 1. Particularly, compared to Endo-FM [30], our M^{2}CRL achieves improvements of 3.5% F1 in classification (PolypDiag), 7.5% Dice in segmentation (CVC-12k), and 2.2% F1 in detection (KUMC) tasks. These improvements are attributed to our pre-training approach, which integrates multi-view mask modeling with contrastive method, substantially enhancing the model's ability for representation learning. Clearly, compared to other methods (TimeSformer [73], CORP [74], FAME [75], *etc.*), M^{2}CRL achieves considerable advantages across three downstream tasks. Although the performance gains on extremely dense task (*i.e.*, segmentation) is modest compared to the latest video self-supervised methods (*i.e.*, VideoMAE [22], VideoMAE V2 [23], DropMAE [79]), our M^{2}CRL makes great progress on classification task and reaches 94.2%. This shows that our M^{2}CRL not only focuses on pixel details but also enhances discriminative capabilities.

Qualitative evaluation We visualize segmentation and detection results in Fig. 2. Compared to other methods, our M^{2}CRL demonstrates superior visual results in segmenting both large and small polyp regions (1^{st} and 2^{nd} rows on the left in Fig. 2). Despite potential issues such as blurry boundaries or lens glare caused by camera movement, M^{2}CRL is still capable of accurately segmenting the target regions (3^{rd} row on the left in Fig. 2). We attribute these results to our multi-view mask modeling, which encourages the model to learn more precise detail information from videos. Similarly, M^{2}CRL also exhibits good performance in detection tasks, especially in detecting small target regions (2^{nd} row on the right in Fig. 2). Although there are some differences between the predictions of our method and the ground truth, we achieve a high degree of overlap with the ground truth. This demonstrates that our M^{2}CRL significantly improves the pre-training capability for endoscopic videos. See § D for more segmentation and detection visual comparison.

4.3 Ablation Studies

Multi-view mask Table 2 illustrates the impact of single-view mask and multi-view mask. All experiments use the same masking ratio. For the single-view, different masking strategies are employed for global and local views, respectively. The random tube mask [22] randomly samples masked tokens in the 2D spatial domain and then extends these tokens along the temporal axis. The random mask [21] randomly masks tokens in the spatiotemporal domain. Our approach samples visible tokens by leveraging frame-aggregated attention from global views, resulting in better results than above two strategies, particularly achieving 80.6% in segmentation task. Similarly, the random tube mask demonstrates some superiority on local views. However, from Table 2, it can be observed that the performance of solely conducting mask modeling on the single-view is inferior to that on multi-views. In multi-view mask, it is evident that employing the frame-aggregated attention guided tube mask on global views and the random tube mask on local views results in significant performance gains. These two complementary mask methods work together to learn richer details of video features.

Hyper-parameters of the FAGTM We conduct an experiment to investigate the impact of the hyperparameters of the frame-aggregated attention guided tube mask (FAGTM) for global views,

Table 2: **Multi-view mask**. We compare multiple different masking strategies on different views. FAGTM = Frame-aggregated Attention Guided Tube Mask. RTM = Random Tube Mask.

views	mask strategies		cla.	seg.	det.
	gloabl	local			
single-view	random	-	90.2 ± 1.5	78.6 ± 1.6	83.8 ± 1.9
	RTM	-	93.0 ± 0.8	77.5 ± 3.1	84.0 ± 1.0
	FAGTM	-	92.7 ± 0.4	80.6 ± 0.5	84.4 ± 1.4
	-	random	91.1 ± 0.7	76.1 ± 1.5	83.7 ± 0.7
	-	RTM	91.1 ± 0.5	77.5 ± 0.6	85.0 ± 0.4
multi-view	random	random	91.3 ± 0.3	77.7 ± 0.4	84.9 ± 1.0
	RTM	RTM	93.2 ± 0.4	80.2 ± 0.9	85.2 ± 1.3
	FAGTM	RTM	**94.2 ± 0.7**	**81.4 ± 0.8**	**86.3 ± 0.8**

Table 3: **Hyper-parameters** of the FAGTM.

γ	cla.	seg.	det.
0.5	91.6 ± 0.5	80.7 ± 0.7	84.8 ± 0.4
0.6	**94.2 ± 0.7**	**81.4 ± 0.8**	**86.3 ± 0.8**
0.7	93.7 ± 1.1	81.0 ± 0.1	85.9 ± 2.6
0.8	92.9 ± 0.7	79.0 ± 0.3	85.4 ± 0.8

Table 4: **The teacher's block** of the FAGTM.

blocks	cla.	seg.	det.
4	91.8 ± 0.7	76.6 ± 1.5	83.6 ± 1.1
8	92.5 ± 0.4	79.7 ± 2.2	84.8 ± 1.0
10	93.9 ± 1.0	80.6 ± 0.9	85.9 ± 1.7
12	**94.2 ± 0.7**	**81.4 ± 0.8**	**86.3 ± 0.8**

the results are shown in Table 3. We perform experiments by sorting each patch of the obtained frame-aggregated attention maps in descending order and selecting the top γ proportion of patches as candidate mask patches. Subsequently, visible tokens are random sampled in candidate mask patches. From Table 3, we can observe that the model performs best when γ is set to 0.6. A lower value indicates selecting visible patches from smaller high-attention regions, leading to excessive attention on non-critical areas during reconstruction, contradicting the setup of the self-supervised pre-text task. On the other hand, higher values of γ adversely affecting the learning efficacy of the model.

The teacher's block used for the FAGTM In our study, we use the last layer block of the teacher ViT-B for the FAGTM. The higher layer block incorporates lower-level features and object-level semantic information, offering comprehensive and abstract features that are essential for the model. Thus, it effectively guides the student network in masking. Table 4 shows that it is most beneficial for the FAGTM to utilize the last layer block of the teacher network.

Masking ratio The impact of different masking ratios is illustrated in Table 5. It can be observed that there is an improvement in results across three downstream tasks when the masking ratio increases from 75% to 90%. Due to the reundancy in videos, it shows that the 75% masking ratio utilized in ImageMAE is not suitable for videos. When the masking ratio in videos reaches 90%, the task becomes challenging due to the limited number of patches available for learning temporal correspondence, thus enhancing the learning capacity of the model. This observation is also validated in VideoMAE [22]. Compared to the optimal masking ratio, a higher masking ratio increases the difficulty of pre-training, hindering the model's ability to learn effective representations. Although our ablation experiments have shown that a masking ratio of 95% can achieve comparable performance, its effectiveness in three downstream tasks is lower than that of the 90% masking ratio. This suggests that at a masking ratio of 95%, the model is placed in a relatively unfavorable learning situation, resulting in suboptimal results.

Analysis of components As see from the first row in Table 6, although the performance of the contrastive learning framework on classification tasks is acceptable, its performance on pixel-level tasks, especially the segmentation task, is not very good. Similarly, within the single mask modeling

Table 5: **Masking ratio.** We choose a masking ratio of 90% for the FAGTM and RTM.

FAGTM (global)	RTM (local)	cla.	seg.	det.
95%	95%	94.0 ± 0.3	81.3 ± 0.4	85.1 ± 1.1
	90%	93.8 ± 0.9	80.7 ± 0.7	86.2 ± 1.3
	85%	93.2 ± 0.7	78.5 ± 0.6	84.9 ± 2.3
	75%	92.6 ± 0.4	77.4 ± 1.7	85.2 ± 2.1
90%	95%	93.8 ± 1.4	80.5 ± 0.7	85.8 ± 0.9
	90%	**94.2 ± 0.7**	**81.4 ± 0.8**	**86.3 ± 0.8**
	85%	93.8 ± 0.8	81.4 ± 1.7	85.6 ± 2.2
	75%	93.2 ± 0.9	78.5 ± 1.9	84.8 ± 1.3
85%	95%	93.2 ± 0.8	79.9 ± 0.4	83.1 ± 1.5
	90%	93.8 ± 0.2	81.2 ± 0.2	83.8 ± 0.9
	85%	94.0 ± 0.4	80.5 ± 1.0	85.1 ± 1.8
	75%	92.5 ± 1.2	79.6 ± 0.7	83.8 ± 2.5
75%	95%	91.7 ± 1.3	76.8 ± 1.8	84.2 ± 2.1
	90%	91.3 ± 0.3	79.0 ± 2.2	84.0 ± 1.3
	85%	91.8 ± 0.2	77.5 ± 1.9	83.8 ± 0.4
	75%	91.2 ± 0.7	74.6 ± 1.4	85.0 ± 0.8

self-supervised framework, there is an improvement in performance for the pixel-level task (*e.g.*, from 73.9% to 80.5% for segmentation), while there is a significant drop in performance for the discriminative task (*e.g.*, from 90.7% to 85.7% for classification). These interesting phenomena once again highlights the strong instance discrimination ability of contrastive learning and the robust ability of mask modeling to acquire local pixel-level information. Our approach integrates both methods and demonstrates consistently strong performance across both structural and pixel-level tasks.

Table 6: **Components analysis**. Our proposed M^{2}CRL combines masked video modeling with contrastive learning and has the best performance.

contrastive learning	masked video modeling	cla.	seg.	det.
✓		90.7 ± 0.4	73.9 ± 1.2	84.1 ± 1.3
	✓	85.7 ± 0.4	80.5 ± 1.2	83.5 ± 3.7
✓	✓	**94.2 ± 0.7**	**81.4 ± 0.8**	**86.3 ± 0.8**

5 Conclusion

In this study, we present a novel SSL approach called M^{2}CRL, which integrates multi-view mask modeling with contrastive learning for endoscopic video analysis. Our method aims to address the lack of pixel-level information extraction in existing SSL methods for endoscopic videos. We leverage the unique characteristics of endoscopy to propose a frame-aggregated attention guided tube mask that captures pixel-level relationships across spatial-temporal dimensions for global views. Additionally, we utilize random tube mask to complement the details of local features from local views. Notably, the attention guidance is derived from multi-head self-attention maps extracted from a teacher model, without incurring additional computational costs. By integrating contrastive learning, our method not only maintains performance in dense prediction tasks but also ensures effectiveness in discriminative tasks. The experimental results on 10 publicly avaiable datasets demonstrate that our M^{2}CRL outperforms the state-of-the-art methods across multiple downstream vision tasks (*i.e.*, classification, segmentation, and detection).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62272404 and 62372170, in part by the Natural Science Foundation of Hunan Province of China under Grants 2022JJ30571 and 2023JJ40638, and in part by the Research Foundation of Education Department of Hunan Province of China under Grant 23A0146.

References

[1] Mon Thiri Myaing, Daniel J MacDonald, and Xingde Li. Fiber-optic scanning two-photon fluorescence endoscope. *Optics Letters*, 31(8):1076–1078, 2006.

[2] Yuan Zhang, Yanglin Huang, and Kai Hu. Multi-scale object equalization learning network for intracerebral hemorrhage region segmentation. *Neural Networks*, 179:106507, 2024.

[3] Kai Hu, Xiang Zhang, Dongjin Lee, Dapeng Xiong, Yuan Zhang, and Xieping Gao. Boundary-guided and region-aware network with global scale-adaptive for accurate segmentation of breast tumors in ultrasound images. *IEEE Journal of Biomedical and Health Informatics*, 27(9):4421–4432, 2023.

[4] Tongtong Liu, Xiongjun Ye, Kai Hu, Dapeng Xiong, Yuan Zhang, Xuanya Li, and Xieping Gao. Polyp segmentation with distraction separation. *Expert Systems with Applications*, 228:120434, 2023.

[5] Atharva Tendle. “the revolution will not be supervised": An investigation of the efficacy and reasoning process of self-supervised representations. *Revolution*, 2021.

[6] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[7] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in Neural Information Pprocessing Systems*, 32, 2019.

[8] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[10] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[11] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[13] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[15] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Green hierarchical vision transformer for masked image modeling. *Advances in Neural Information Processing Systems*, 35:19997–20010, 2022.

[16] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[17] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021.

[18] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in Neural Information Processing Systems*, 6, 1993.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Kowledge and Data Engineering*, 35(4):3313–3332, 2021.

[21] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in Neural Information Processing Systems*, 35:35946–35958, 2022.

[22] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022.

[23] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[24] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021.

[25] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10375–10385, 2023.

[26] Zhengqi Liu, Jie Gui, and Hao Luo. Good helper is around you: Attention-driven masked image modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1799–1807, 2023.

[27] Yu Tian, Guansong Pang, Fengbei Liu, Yuyuan Liu, Chong Wang, Yuanhong Chen, Johan Verjans, and Gustavo Carneiro. Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2022.

[28] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[29] Kaidong Li, Mohammad I Fathan, Krushi Patel, Tianxiao Zhang, Cuncong Zhong, Ajay Bansal, Amit Rastogi, Jean S Wang, and Guanghui Wang. Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. *Plos One*, 16(8):e0255809, 2021.

[30] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023.

[31] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 179–189. IEEE, 2019.

[32] Yuqi Huo, Mingyu Ding, Haoyu Lu, Zhiwu Lu, Tao Xiang, Ji-Rong Wen, Ziyuan Huang, Jianwen Jiang, Shiwei Zhang, Mingqian Tang, et al. Self-supervised video representation learning with constrained spatiotemporal jigsaw. In *International Joint Conference Artificial Intelligence*, pages 751–757, 2020.

[33] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019.

[34] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, pages 843–852. PMLR, 2015.

[35] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 835–851. Springer, 2016.

[36] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016.

[37] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[38] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[39] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022.

[40] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European Conference on Computer Vision*, pages 391–408, 2018.

[41] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer, 2016.

[42] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3636–3645, 2017.

[43] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.

[44] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[45] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019.

[46] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.

[47] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 37–45, 2015.

[48] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[49] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[50] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

[51] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4086–4093, 2015.

[52] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[53] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[54] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[55] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[56] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[57] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[58] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[59] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318. Springer, 2022.

[60] David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector Santos-Villalobos, Rohith MV, and Xinyu Li. Motion-guided masking for spatiotemporal representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5619–5629, 2023.

[61] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14507–14517, 2023.

[62] Sunil Hwang, Jaehong Yoon, Youngwan Lee, and Sung Ju Hwang. Efficient video representation learning via motion-aware token selection. *arXiv preprint arXiv:2211.10636*, 2022.

[63] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13493–13504, 2023.

[64] Yotam Intrator, Natalie Aizenberg, Amir Livne, Ehud Rivlin, and Roman Goldenberg. Self-supervised polyp re-identification in colonoscopy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 590–600. Springer, 2023.

[65] Roy Hirsch, Mathilde Caron, Regev Cohen, Amir Livne, Ron Shapiro, Tomer Golany, Roman Goldenberg, Daniel Freedman, and Ehud Rivlin. Self-supervised learning for endoscopic video analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 569–578. Springer, 2023.

[66] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[67] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR, 2020.

[68] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

[69] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[70] Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. Pixmim: Rethinking pixel reconstruction in masked image modeling. *arXiv preprint arXiv:2303.02416*, 2023.

[71] Xue H, Gao P, and other. Stare at what you see: Masked image modeling without reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22732–22741, 2023.

[72] Yunjie Tian, Lingxi Xie, Zhaozhi Wang, Longhui Wei, Xiaopeng Zhang, Jianbin Jiao, Yaowei Wang, Qi Tian, and Qixiang Ye. Integrally pre-trained transformer pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18610–18620, 2023.

[73] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, volume 2, page 4, 2021.

[74] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949, 2021.

[75] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9726, 2022.

[76] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14711–14721, 2022.

[77] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Static and dynamic concepts for self-supervised video representation learning. In *European Conference on Computer Vision*, pages 145–164. Springer, 2022.

[78] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022.

[79] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14561–14571, 2023.

[80] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.

[81] Pablo Mesejo, Daniel Pizarro, Armand Abergel, Olivier Rouquette, Sylvain Beorchia, Laurent Poincloux, and Adrien Bartoli. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Transactions on Medical Imaging*, 35(9):2051–2063, 2016.

[82] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022.

[83] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 387–396. Springer, 2021.

[84] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020.

[85] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):142, 2021.

[86] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[87] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[88] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[89] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[90] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[91] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[92] Lingyun Wu, Zhiqiang Hu, Yuanfeng Ji, Ping Luo, and Shaoting Zhang. Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 302–312. Springer, 2021.

[93] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2016.

[94] Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, et al. Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis*, 88:102844, 2023.

[95] Xueying Shi, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition. *Medical Image Analysis*, 73:102158, 2021.

Supplementary Material

In this supplementary material, we first provide model implementation details for reproducibility in Section A. Next, we introduce additional ablation experiments in Section B. In Section C, we conduct another experiment on surgical phase recognition in downstream tasks. In Section D, we visualize more segmentation and detection results in downstream tasks and perform qualitative analysis. Subsequently, we provide the case study of frame-aggregated attention guided tube mask in Section E. Finally, we analyze the limitations and broader impacts of our work in Section F.

A Implementation Details

Pre-training We employ data augmentation techniques such as random horizontal flipping, color jitter, gaussian blur, and exposure adjustment, and also utilize temporally consistent spatial enhancements [88] to all frames within a single view. The decoder is a lightweight one-layer head [16] and the model performs learning using a simple ℓ_1 loss. The FLOPs of our model for a single execution are approximately 190G, with a parameter count of 121M. Pre-training are conducted on 4 Tesla A100 GPUs. The training parameters are shown in Table 7.

Table 7: **Pre-training settings.**

config	value
optimizer	AdamW [89]
optimizer momentum	β_1,β_2 = 0.9, 0.999
weight decay	$4e-2$
base learning rate	$2e-5$
learning rate schedule	cosine schedule [90]
warmup epochs	10
pretraining epochs	30
batch size	12
temperature parameters	$\tau_t, \tau_s = 0.04, 0.07$
mask rate	$\rho = 0.9$
attention areas threshold	$\gamma = 0.6$
momentum coefficient	$\lambda = 0.996$

Evaluation methodology For downstream fine-tuning, the procedure is as follows: (1) Classification: We use the PolypDiag [27] dataset, which includes 253 videos and 485,561 frames. Each video is annotated to indicate the presence or absence of a lesion. The dataset is divided into 71 normal videos without polyps and 102 abnormal videos with polyps for training, and 20 normal videos and 60 abnormal videos for testing. We sample 8 frames at a resolution of 224×224 from each video as input, utilizing a pre-trained model to initialize the backbone and appending randomly initialized linear layers for training 20 epochs. The SGD optimizer is employed, with the learning rate set to 1e-3, momentum to 0.9, and batch size to 4. (2) Segmentation: We use the CVC-12k [28] dataset, which includes 29 videos and 612 frames, with 20 videos allocated for training and 9 videos for testing. Each frame in the videos is annotated with ground truth masks (with a single class) to identify the regions covered by polyps. We employ TransUnet [91] as the segmentation decoder following the code of [91]. The AdamW optimizer is used to optimize the overall parameters by setting the learning rate as 1e-4, weight deacy as 5e-2 and the batchsize as 1. We resize the spatial size as 224×224 and fine-tune for 150 epochs. (3) Detection: We use the KUMC [29] dataset, which includes 53 videos and 19,832 frames. Each frame in each video is annotated with bounding boxes and polyp categories, with 36 videos allocated for training and 17 videos for testing. We employ STFT [92] to implement the detection task, fine-tuning for 24k iterations at a resolution of 640×640. The SGD optimizer is used to optimize the overall parameters by setting the learning rate as 2.5e-3, weight deacy as 1e-4 and momentum as 0.9. See the STFT [92] for more training details.

B Additional Ablations

Prediction target From Table 8, it is observed that M^{2}CRL achieves better performance when using RGB pixel values as the reconstruction target. The model, which utilizes feature distillation as the prediction target, demonstrates decent results in classification tasks but shows a decline in performance in segmentation and detection tasks. This once again highlights the significant advantage of incorporating a reconstruction task involving masked patches for pixel-level tasks.

Table 8: **Prediction target**. The effect of pixel regression is better.

prediction target	cla.	seg.	det.
feature distillation	94.2 ± 0.4	77.9 ± 1.6	84.5 ± 1.0
pixel regression	**94.2 ± 0.7**	**81.4 ± 0.8**	**86.3 ± 0.8**

Ablations on loss The mask modeling component of our model follows SimMIM [16], which employs the ℓ_1 loss function. The study demonstrates that different loss functions have minimal impact. To maintain consistency, we use the same loss function as SimMIM. Furthermore, we conduct ablation studies to demonstrate that different loss functions have a negligible effect on our results, as shown in Table 9.

Table 9: **Ablations on loss.**

loss	cla.	seg.	det.
ℓ_1	94.2 ± 0.7	81.4 ± 0.8	86.3 ± 0.8
ℓ_2	93.8 ± 0.7	82.0 ± 0.7	85.9 ± 1.5

Ablation on different architectures For a fair comparison, we use the weights for initialization as Endo-FM did. However, since this work does not provide weights for ViT variants, we are unable to conduct ablation experiments on different architectures with weight initialization. Consequently, we conduct a set of ablation experiments without weight initialization for the backbone, as shown in Table 10. We observe that the performance improvement is more pronounced with larger models due to their increased parameters and more complex structures, enabling them to capture more intricate features. In contrast, smaller models have limited feature extraction capabilities and cannot fully extract visual features. Although larger models exhibit stronger learning abilities, they are more prone to overfitting during training. Additionally, larger models require higher computational resources and longer training times. In conclusion, choosing ViT-B as the pre-trained backbone is a suitable compromise.

Table 10: **Ablation on different architectures.**

backbone	cla.	seg.	det.
ViT-T/16	93.4 ± 0.9	76.8 ± 1.2	76.3 ± 2.4
ViT-S/16	93.8 ± 0.4	78.2 ± 1.5	79.4 ± 0.7
ViT-B/16	93.4 ± 0.9	80.5 ± 0.5	83.4 ± 2.8
ViT-L/16	94.0 ± 0.9	83.2 ± 0.8	84.2 ± 2.0

Table 11: **Ablation on initialization status.**

initialization status	cla.	seg.	det.
random	93.4 ± 0.9	80.5 ± 0.5	83.4 ± 2.8
kinetics weights	**94.2 ± 0.7**	**81.4 ± 0.8**	**86.3 ± 0.8**

Ablation on initialization status We perform ablation experiments on the model initialization status. As shown in Table 11, the results indicate that M^{2}CRL without weight initialization performs slightly worse under the same pre-training conditions. This is because weight initialization accelerates model convergence and enhances model stability. However, for a fair comparison, we follow Endo-FM and use initialized weights.

C Surgical phase recognition

In the downstream tasks, we conduct surgical phase recognition experiments using the Cholec80 [93] dataset, which contains 80 complete laparoscopic cholecystectomy videos. This dataset is specifically designed to evaluate the performance of the model in automatic surgical phase recognition. To ensure a fair comparison with existing methods, we follow the data split and evaluation protocol described by [94], using 40 videos for training and 40 for testing. For evaluation, we adopt the relaxed boundary F1 score proposed by [95]. Since our backbone is based on Transformers, we do not use the CNN-based TCN [94] as a feature extractor during fine-tuning, so our result is based on single frames. As shown in Table 12, our method achieves superior performance. This experiment further verifies the robustness and effectiveness of our method across various endoscopic video tasks.

Table 12: **Surgical phase recognition.**

Methods	F1
DINO	77.6
MoCo v2	81.7
SimCLR	84.5
SwAV	86.1
Endo-FM	87.5
M^{2}CRL	**88.7**

D Qualitative Evaluation

Fig. 3 illustrates the segmentation results of our method and other self-supervised pre-training methods on the CVC-12k dataset. Polyps are characterized by their blurry boundaries and significant shape variations, which undoubtedly add complexity to the segmentation task. Nevertheless, our method consistently outperforms other state-of-the-art self-supervised methods. Specifically, when polyps are very close and overlapping (3^{rd} row of Fig. 3), our method produces smoother and more continuous edges, whereas other methods often result in blurred or fragmented edges. Additionally, some small polyps (4^{th} row of Fig. 3) occupy very few pixels and are easily missed, our method successfully segments these small polyps without omission. Fig. 4 shows the detection results of our method and other self-supervised pre-training methods on the KUMC dataset. As shown in the figure, our method performs well in boundary identification and localization for small polyps (1^{st} and 2^{nd} rows of Fig. 4). Furthermore, 3^{rd} and 4^{th} rows of Fig. 4 demonstrate that our method can also accurately identify and locate polyps even when they have low contrast and indistinct boundaries with surrounding tissues. Notably, compared to Endo-FM, a discriminative approach, our method significantly enhances the performance on pixel-level tasks.

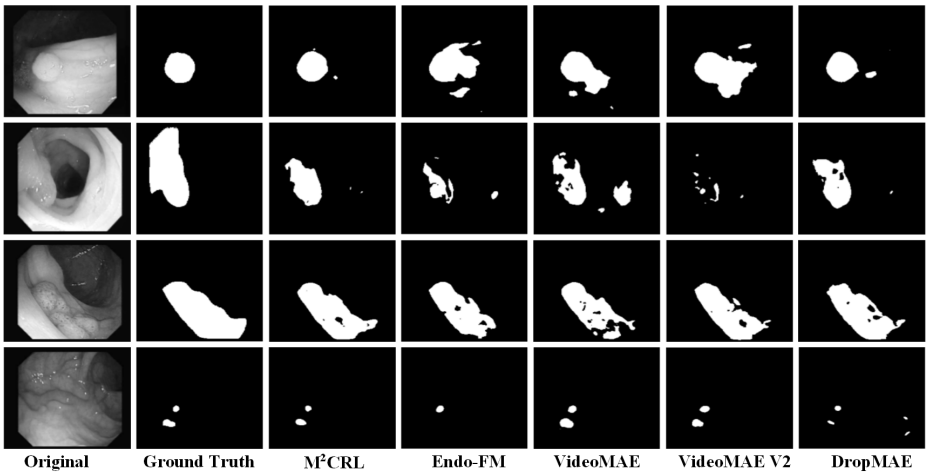

Figure 3: Qualitative results for segmentation task on the CVC-12k dataset.

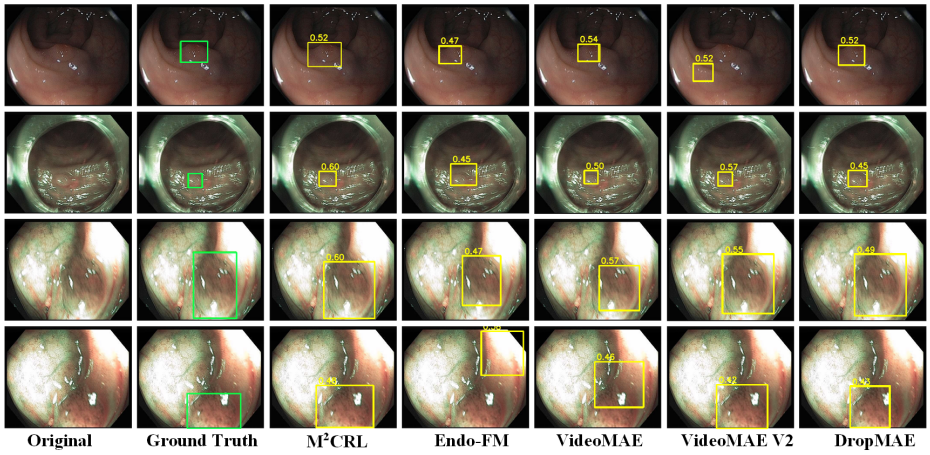

Figure 4: Qualitative results for detection task on the KUMC dataset.

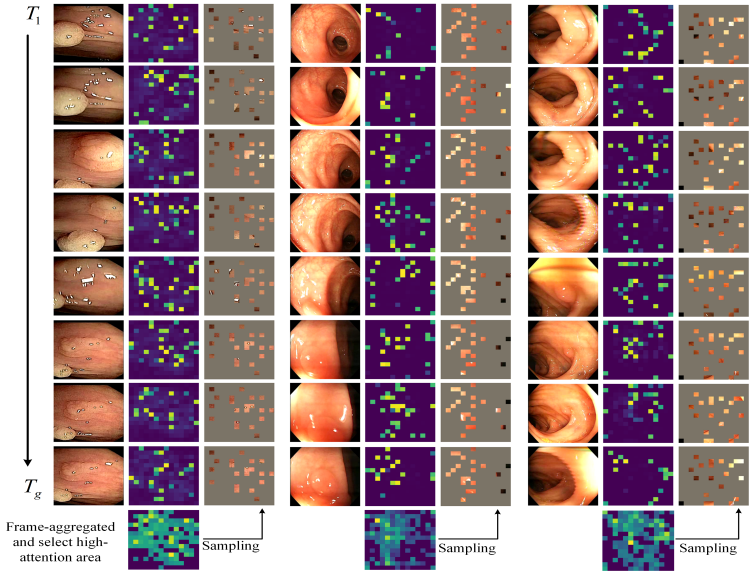

Figure 5: Illustration of our frame-aggregated attention guided tube masking strategy. We visualize spatial attention map with temporal information for each frame (2^{nd} column), then aggregate attention maps for all frames and select area of high attention. We sample visible patches in this area (3^{rd} column).

E Case Study

In Fig. 5, we visualize the attention maps of the ViT-B/16 model employing "the divided space-time attention" mechanism. We utilize the $[cls]$ token as a query and employ attention from the last transformer block. From the attention map of individual frames, we observe that the model does not distinctly learn the concept of endoscope object boundaries, attributed to the separate computations of attention in time and space. To enhance object concepts within endoscope video sequences, we aggregate attention across all frames of the video and then select a certain proportion of high-attention regions. This approach ensures that even in complex endoscopic video scenarios, key information within the video sequence is retained. Random masking at a high ratio in the video can obscure critical regions, thereby impeding the ability of the model to learn video representations. To address this issue, we select regions of high attention from the aggregated attention map for sampling visible tokens.

F Limitations and Broader Impacts

Limitations Our work presents a multi-view masked contrastive representation learning (M^2CRL) framework for endoscopic video analysis. However, our current self-supervised pre-training method only utilizes RGB video streams and does not incorporate additional audio and text streams. In the future, we expect that audio and text data can provide more information for self-supervised pre-training. Furthermore, our study requires extensive pre-training, leading to significant energy consumption and reliance on high-performance computing hardware (GPU). These negative impacts underscore the necessity of considering environmental protection and resource conservation. In future work, we will adopt more efficient training methods and optimization strategies to address these issues.

Broader impacts Our approach demonstrates the potential of SSL in endoscopic video analysis. By utilizing a large amount of unlabeled endoscopic video data for pre-training, we can reduce the dependence on costly annotated medical data, thus lowering healthcare expenses. Furthermore, our pre-trained models can be easily applied to various tasks such as classification, segmentation, and detection, thereby making valuable contributions to medical applications and enhancing the quality and efficiency of clinical disease diagnosis and healthcare services.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in our abstract and introduction accurately reflect the contributions and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our paper in § F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We introduce our method in detail in Section 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We present the information needed to reproduce the main experimental results of the paper in Section 4 and § A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We open-source the code. We do not need to release any data, as we use publicly available datasets.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and § A of our paper present all the training and test details necessary to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The reported result is in mean + std format, upon 3 independent runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We show the time required to run the experiment in Table 1 and the computing resources in § A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in our work conforms in all respects to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts of our paper in § F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There are no such risks in our paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all the relevant assets used in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We use publicly available datasets, so this issue does not arise.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We use publicly available datasets, so this issue does not arise.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.