applied sciences

MDPI

Article

Frequency-Aware Unsupervised Domain Adaptation for Semantic Segmentation of Laparoscopic Images

Huiwen Dong and Gaofeng Zhang *

School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China; dhuimin2023@lzu.edu.cn

* Correspondence: zhanggaof@lzu.edu.cn

Abstract

Semantic segmentation of laparoscopic images requires costly pixel-level annotations, which are often unavailable for real surgical data. This gives rise to an unsupervised domain adaptation scenario, where labeled synthetic images serve as the source domain and unlabeled real images as the target. We propose a frequency-aware unsupervised domain adaptation framework to mitigate the domain gap between simulated and real laparoscopic images. Specifically, we introduce a Radial Frequency Masking module that selectively masks frequency components of real images, and employ a Mean Teacher framework to enforce consistency between high- and low-frequency representations. In addition, we propose a module called Fourier Domain Adaptation-Blend, a style transfer strategy based on low-frequency blending, and apply entropy minimization to enhance prediction confidence on the target domain. Experiments are conducted on public datasets by jointly training on simulated and real laparoscopic images. Our method consistently outperforms representative baselines. These results demonstrate the effectiveness of frequency-aware adaptation in surgical image segmentation without relying on manual annotations from the target domain.

Keywords: fourier transform; laparoscopic surgery; semantic segmentation; unsupervised domain adaptation (UDA)

1. Introduction

Laparoscopic surgery is a widely used minimally invasive surgical technique. During the procedure, it is critical to accurately identify anatomical structures and surgical tools in laparoscopic images. Semantic segmentation serves as an effective solution to this task. With the rapid development of deep learning, various advanced CNN [1–5] and Transformer [6–9] architectures are introduced into semantic segmentation of laparoscopic images. These methods are effective in modeling the spatial semantic features of images and have achieved impressive performance on multiple public benchmark datasets. However, their success heavily relies on large quantities of high-quality pixel-level annotations. In the context of laparoscopic surgery, such annotations typically require frame-by-frame labeling and the involvement of domain experts, making the annotation process labor-intensive, time-consuming, and costly [10]. As a result, constructing large-scale datasets of real laparoscopic images remains a challenging and often impractical task.

As a viable alternative, researchers have increasingly turned to computer-generated simulated laparoscopic images to train models [11]. These simulated datasets offer significant advantages, including lower acquisition costs, clearly defined anatomical structures,

Academic Editor: Thomas Lindner

Received: 12 December 2025
Revised: 5 January 2026
Accepted: 7 January 2026
Published: 14 January 2026

and the automatic generation of accurate pixel-level annotations. Despite these benefits, models trained solely on simulated data often fail to generalize well to real surgical scenarios [12]. This is primarily due to the domain shift between simulated and real laparoscopic images, which manifests in differences in texture details, lighting, and background complexity. To address the challenge of domain shift, Unsupervised Domain Adaptation (UDA), which aims to improve target-domain performance by cooperatively utilizing labeled source data and unlabeled target data, has proven to be an effective solution [13,14].

Although recent UDA methods have achieved remarkable success in the semantic segmentation for laparoscopic images [11], there is still a noticeable performance gap compared to supervised training. A common problem is the confusion in classifying complex regions in the real images, due to the absence of ground truth in the target domain. For example, in Figure 1, the real laparoscopic surgery images demonstrate how factors such as shape deformation, lesions, blood vessels, and lighting conditions make visual recognition particularly challenging for the network. Several prior studies [11,15] leverage frequency information from images. For instance, as shown in Figure 1, complex backgrounds and rich textures represent the high-frequency components of real images, while lighting and style correspond to their low-frequency components. In contrast, simulated images with different styles often lack such rich details. This leads to a significant frequency-domain gap between simulated and real images.

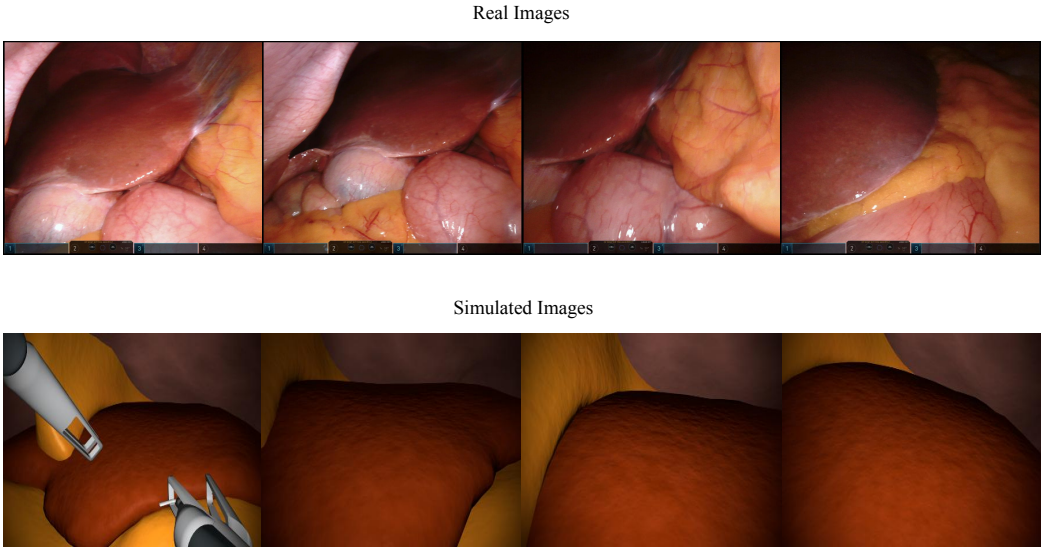

Figure 1. Real laparoscopic images are visualized in the first row, and simulated laparoscopic images in the second row.

Therefore, we propose a novel UDA framework for laparoscopic semantic segmentation, aiming to promote robust frequency-aware representation learning. Specifically, we introduce a plug-in module called Radial Frequency Masking (RFM) for UDA. RFM randomly masks out high-frequency components with high probability and low-frequency components with low probability in the frequency spectrum of real images. We train the network to predict the semantic segmentation result of the entire image, including the masked regions. In this way, the network is forced to infer the semantics from the missing frequency information. Since ground truth annotations are not available in the target domain, we adopt pseudo-labels generated by an Exponential Moving Average (EMA) teacher [16] network as supervision. The teacher takes the original (unmasked) target images as input, while the student network receives the frequency-masked versions. During training, different parts of the frequency spectrum are masked out in each iteration, compelling the network to learn robust representations across the frequency

domain, thereby improving its generalization ability and robustness on real laparoscopic images. On the other hand, to further mitigate the domain shift, we design a method to transform the source images before feeding them into the student network. Inspired by the success of Frequency Domain Adaptation (FDA) [15], we leverage the observation that semantic information is mainly preserved in the phase and high-frequency components of an image, while the low-frequency amplitude, although visually dominant, contributes little to semantic understanding. We propose FDA-Blend, an extension of Fourier Domain Adaptation (FDA), which modifies the style of source images by blending their amplitude spectra with those of target images using a mixing coefficient α. This fusion aims to align the low-frequency components between source and target domains, thereby achieving style-level adaptation. Unlike FDA, which directly replaces low-frequency components, FDA-Blend introduces a blending coefficient α to softly fuse the low-frequency spectrum of the source image with that of the target image, preserving domain-specific structures while promoting alignment. In this way, the semantic content of the source image is preserved, while its style and illumination are adapted to better match the target domain, facilitating more effective knowledge transfer during training. These transformed source images, along with their original ground-truth labels, are then used to supervise the student network. In addition, to reduce uncertainty in target-domain predictions, we apply pixel-wise entropy minimization [17] on target images, which encourages confident outputs.

The primary contributions of this paper are as follows:

- We creatively address the challenge of UDA for semantic segmentation of laparoscopic surgery images, where the model is trained using labeled computer-generated simulated images and unlabeled real images.
- We propose a novel UDA framework that enhances the model's robustness to both high- and low-frequency information. Specifically, we introduce a plug-in RFM module to selectively mask frequency components during self-training, and propose FDA-Blend to better align image styles between domains via low-frequency spectrum blending. In addition, we incorporate an entropy minimization loss to encourage more confident predictions.
- Extensive experiments demonstrate that our method surpasses existing UDA baselines and provides an effective solution for laparoscopic image segmentation without requiring real image annotations.

2. Related Work

2.1. Deep Learning for Medical Image Segmentation

In recent years, deep learning has significantly advanced medical image segmentation, with convolutional neural networks (CNNs) [1,3,5,18,19] emerging as the predominant architecture. One of the earliest milestones was the Fully Convolutional Network (FCN) [3], which enabled end-to-end pixel-level prediction and paved the way for subsequent developments. The U-Net [1] architecture introduced a symmetric encoder–decoder structure with skip connections to effectively integrate low-level spatial details and high-level semantic features. U-Net achieved remarkable performance in clinical applications and inspired a variety of variants with residual connections, dense pathways, and multi-scale fusion strategies, such as ResUNet [20], H-DenseUNet [21], UNet++ [22] and UNet3+ [23]. To handle volumetric data, 3D U-Net [24] extended convolution and upsampling operations to three dimensions, while V-Net [25] incorporated large-stride convolutions and residual learning to enhance 3D structural modeling. Across these CNN-based methods, expansion of receptive fields is typically achieved through deeper networks or multi-scale designs. However, these models still face limitations in capturing long-range dependencies and complex spatial relationships.

The recent emergence of Transformer-based architectures [6–8] has brought a paradigm shift to medical image segmentation. Powered by self-attention [9] mechanisms, Transformers capture global contextual information, addressing the inherent locality of CNNs. Early efforts such as TransUNet [26] integrated Vision Transformers (ViT) [6] into CNN backbones to improve semantic modeling. This line of research soon evolved towards more unified or purely Transformer-based designs. A representative example is the Swin Transformer [7], which employs hierarchical feature extraction, window-based self-attention, and cross-window connections to balance efficiency and global reasoning. Its adaptation to medical segmentation, such as in Swin-Unet [27], has shown promising results. More recent studies explore hybrid architectures that combine the local inductive bias of CNNs with the global modeling capability of Transformers. Representative examples like CoTr [28] adopt dual-branch or interleaved designs to capture both fine-grained textures and long-range semantics. Overall, Transformer-based models are driving a transition from structural augmentation to a fundamental rethinking of spatial reasoning and semantic representation in medical image segmentation.

2.2. Unsupervised Domain Adaptation in Semantic Segmentation

Unsupervised domain adaptation (UDA) [13,14] for semantic segmentation aims to transfer knowledge from a labeled source domain to an unlabeled target domain, despite significant shifts in data distribution and appearance. Over the years, UDA research has evolved into several methodological paradigms, each addressing the domain gap from a different perspective. Adversarial learning approaches align source and target distributions by introducing domain discriminators, typically applied in output or feature space. AdaptSegNet [29] and HRDA [30] are representative examples that achieve pixel-wise alignment through adversarial training at multiple scales. Appearance-level alignment methods, such as CyCADA [31] and recent diffusion-based techniques, operate at the pixel level by transforming source images to resemble target domain appearance, aiming to bridge the visual gap while preserving semantic content. Discrepancy-based methods, such as MCD [32], introduce multiple classifiers to expose ambiguous target regions by maximizing inter-classifier disagreement, which is then minimized through adaptation, providing a stable alternative to adversarial objectives. Self-training leverages high-confidence predictions as pseudo-labels for target data. CBST [33] and DACS [34] use filtering and mixing strategies. Mean Teacher [16] introduces temporal ensembling and consistency losses to stabilize pseudo-labels. Entropy-based strategies (e.g., ADVENT [17]) aim to minimize the uncertainty of model predictions in the target domain by reducing output entropy, thereby encouraging confident and compact output distributions. This process implicitly pushes the decision boundary away from high-density regions in the feature space, leading to improved domain discriminability and safer generalization. Fourier-based frequency-domain strategies have also shown strong effectiveness for medical image segmentation domain adaptation [35–37]. Frequency-domain alignment methods treat images as signals, replacing or blending low-frequency components (e.g., FDA [15]) to align domain-specific appearance while retaining high-frequency semantic structure. Despite their diverse formulations, these paradigms converge on the goal of improving generalization to unseen domains. However, challenges such as error accumulation in self-training, instability in adversarial objectives, and under-exploration of multi-modal cues remain open, highlighting directions for future research.

2.3. Masking Strategy

In natural language processing, predicting masked tokens in input sequences has proven to be a powerful self-supervised pretraining task [38,39]. Recently, this concept has

been successfully transferred to computer vision, where models learn visual representations by reconstructing masked regions of an image. For instance, Masked Autoencoders (MAE) [40] encode only the visible patches of an image using a ViT encoder and reconstruct the missing patches with a lightweight decoder, thereby enabling efficient self-supervised learning. Various masking strategies have been explored [40–44]. In contrast to prior work focusing on representation learning, we employ masking as a structured perturbation to facilitate domain adaptation. In semantic segmentation, previous UDA methods have applied masking in the pixel space to encourage models to capture contextual dependencies [45]. Some approaches have also explored frequency masking in the spectral domain [11]. Inspired by these spectral masking strategies, we further optimize the masking design. We propose the Radial Frequency Masking (RFM) module, which is specifically designed for UDA in laparoscopic image segmentation. To the best of our knowledge, RFM is the first to selectively mask frequency components based on their radial distance from the spectral center, thereby promoting consistency regularization across high- and low-frequency components to enhance domain adaptation.

3. Methodology

This paper addresses the task of domain-adaptive semantic segmentation of laparoscopic images. In unsupervised domain adaptation (UDA) for semantic segmentation, we are provided with a labeled source dataset $D^s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ where $x_s \in \mathbb{R}^{H\times W\times 3}$ denotes an RGB image, and $y_s \in \mathbb{R}^{H\times W}$ is the corresponding semantic label map. Similarly, $D^t = \{x_t^i\}_{i=1}^{N_t}$ is the target dataset, where the ground truth semantic labels are absent. Generally, the segmentation network trained on D^s will have a performance drop when tested on D^t due to the domain shift. Our objective is to train a model with robust semantic segmentation capability on the unlabeled target domain. To achieve this, we introduce a novel UDA framework comprising three key modules: Radial Frequency Masking (RFM), FDA-Blend, and Entropy Minimization.

3.1. Radial Frequency Masking (RFM)

3.1.1. Image Frequency Representation

Consider an RGB image $X \in \mathbb{R}^{H\times W\times 3}$. We compute the frequency-domain representation for each channel $c \in \{0, 1, 2\}$ of the image by applying a 2D discrete Fourier transform (DFT), as defined in Equation (1):

$$\mathcal{F}(\mathbf{X})_{(u,v,c)} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathbf{X}_{(h,w,c)} e^{-i2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)} \tag{1}$$

where $i^2 = -1$, and (u, v) and (h, w) denote the coordinates in the frequency domain and image domain, respectively.

To facilitate subsequent operations, the Fast Fourier Transform (FFT) output is rearranged by shifting negative frequency components to precede positive frequency components, thereby centering the low-frequency information. An inverse Fourier transform is then utilized to convert the spectral representation back to the spatial domain, reconstructing the original image:

$$\mathbf{X}_{(h,w,c)} = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} \mathcal{F}(\mathbf{X})_{(u,v,c)} e^{i2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)} \tag{2}$$

3.1.2. Radial Frequency Masking Strategy

As illustrated in Figure 2, we propose the RFM module, which masks the frequency information of target domain images. To achieve this, we define a mask map $\mathcal{M}$ that randomly covers the frequency spectrum, thereby eliminating parts of the frequency information. Specifically, a mask is defined as follows:

$$\mathcal{M}_{i,j} = \left[v > k\frac{r}{r_{max}} \right] \quad \text{with} \quad v \sim \mathcal{U}(0,1) \tag{3}$$

where $[\cdot]$ is the Iverson bracket, $\mathcal{U}$ is the uniform distribution, and (i,j) denotes the coordinates in frequency map, k is the ratio. After this procedure, the frequency data is randomly masked.

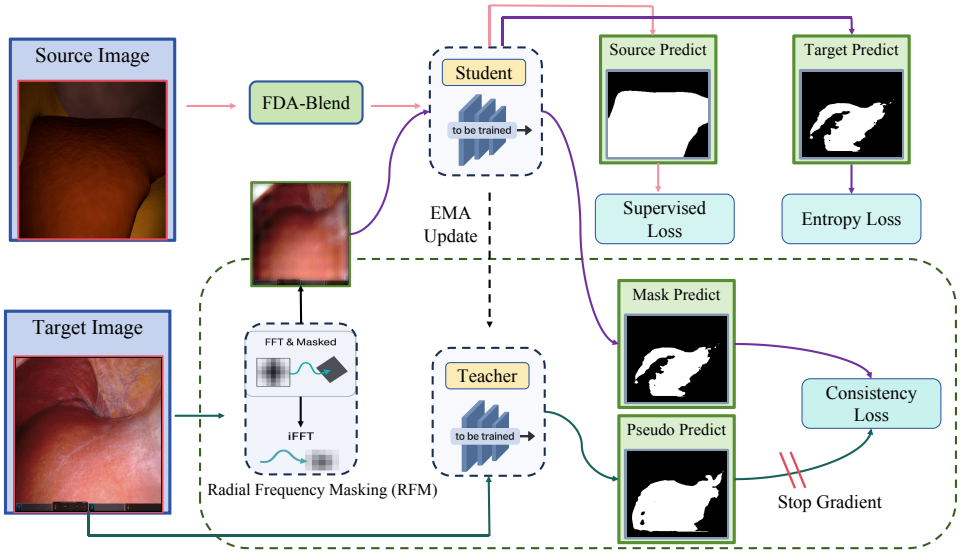

Figure 2. The overview of our proposed framework: Labeled source images are first transferred into the style of the target domain via FDA-Blend and then fed into the Student network. A supervised cross-entropy loss is applied using the ground truth labels of the source domain. The Teacher network receives the original unlabeled target images and generates pseudo-labels. Meanwhile, the Student network takes as input frequency-masked versions of the same images, which are obtained by applying the Radial Frequency Masking (RFM) module, and is trained using the pseudo-labels. The Teacher is gradient-free and updated using an exponential moving average (EMA) of the Student's parameters. In parallel, the Student also processes the masked target images and produces predictions. To improve the certainty of these predictions, an entropy loss is applied to the Student's predictions on the original target images, encouraging more confident outputs on the original target data.

While some existing methods [11] preserve the central low-frequency components as defined in Equation (4), we observe that enforcing this constraint in our masking strategy may adversely affect the model's performance. This finding will be further validated through comprehensive ablation studies.

$$\mathcal{M}_{H/2-h:H/2+h,\, W/2-w:W/2+w} = 1 \tag{4}$$

where h and w denote the size of the low-frequency information to be preserved.

We apply the mask $\mathcal{M}$ to the frequency spectrum, and then use the inverse Fast Fourier Transform (iFFT) to convert it back to the image domain as the network input.

The equation represents how the masked image is reconstructed by applying the frequency mask in the Fourier domain and then performing an inverse transform.

$$\mathbf{X}_m = \mathcal{F}^{-1}(\mathcal{F}(X) \odot \mathcal{M}), \tag{5}$$

where $\odot$ is the Hadamard product between the matrices. Moreover, with conjugate symmetry's disruption inhibiting the imaginary component's cancellation, we employ complex number magnitudes as outputs.

3.1.3. Consistency Regularization

We adopt the Mean Teacher [16] framework, where consistency regularization is utilized to encourage the student network to better capture masked frequency information. During training, the teacher takes the original target images as input, while the student network receives the frequency-masked versions. We adopt pseudo-labels generated by the teacher network as supervision. The weight θ_T of the teacher network is detached from gradient updates and progressively updated via the exponential moving average (EMA) of the student network's weight θ_S:

$$\boldsymbol{\theta}_T = \lambda_{EMA}\boldsymbol{\theta}_S + (1 - \lambda_{EMA})\boldsymbol{\theta}'_T, \tag{6}$$

where θ'_T represents the weight from the previous training step.

During training, the teacher network generates a set of pseudo-labels, which are used to supervise the learning of the student network. We employ the mean squared error (MSE) loss to measure the discrepancy between the predictions of the student network and pseudo-labels. Since the pseudo-labels may be inaccurate, as described in prior works [11,30,34,45], we only consider reliable pixels whose maximum predicted probability exceeds a confidence threshold τ. The overall consistency loss can thus be formulated as follows:

$$\mathcal{L}_C = q_T \odot \mathrm{MSE}(f_T(\mathbf{X}), f_S(\mathbf{X}_m)), \tag{7}$$

where f_T denotes the teacher network, f_S the student network, and q_T is the quality weight. The value of q_T is computed as the ratio of confident pixels across spatial dimensions, formulated as follows:

$$q_T = \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\Big[\max_c f_T(\mathbf{X})_{hwc} > \tau\Big]. \tag{8}$$

Here, $f_T(\mathbf{X})_{hwc}$ represents the predicted probability for class c at spatial location (h, w) from the teacher network, and the Iverson bracket $[\cdot]$ evaluates to 1 if the condition holds, and 0 otherwise.

3.2. FDA-Blend

We propose FDA-Blend, an extension of Fourier Domain Adaptation (FDA) [15], which linearly interpolates the amplitude spectra of source and target images using a mixing coefficient α. Let $\mathcal{F}^A, \mathcal{F}^P : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times 3}$ be the amplitude and phase components of the Fourier transform $\mathcal{F}$ of an RGB image. We apply a mask M, whose values are set to zero everywhere except for the central region, where $\beta \in (0, 1)$ controls the proportion of the preserved area.

$$M_\beta(h, w) = \mathbf{1}_{(h,w)\in[-\beta H:\beta H,\ -\beta W:\beta W]} \tag{9}$$

We assume that the center of the image is located at (0, 0). Note that β is a ratio coefficient and is independent of the image size or resolution. Given two randomly sampled images $x^s \sim D^s$, $x^t \sim D^t$, Fourier Domain Adaptation Blend can be formalized as follows:

$$\begin{aligned} x^{s\to t} = \mathcal{F}^{-1}\big(M_\beta \circ \Big(\alpha\mathcal{F}^A(x^t) + (1-\alpha)\mathcal{F}^A(x^s)\Big) \\ + (1 - M_\beta) \circ \mathcal{F}^A(x^s),\ \mathcal{F}^P(x^s)\big) \end{aligned} \tag{10}$$

where the low-frequency amplitude components $\mathcal{F}^A(x^s)$ of the source image are partially replaced by those from the target image x^t. The modified frequency representation of the source image x^s, with its phase component unchanged, is then mapped back to the spatial domain via the inverse Fourier transform, resulting in $x^{s\to t}$, which preserves the content of x^s while exhibiting the style characteristics of the target domain. The process is illustrated in Figure 3.

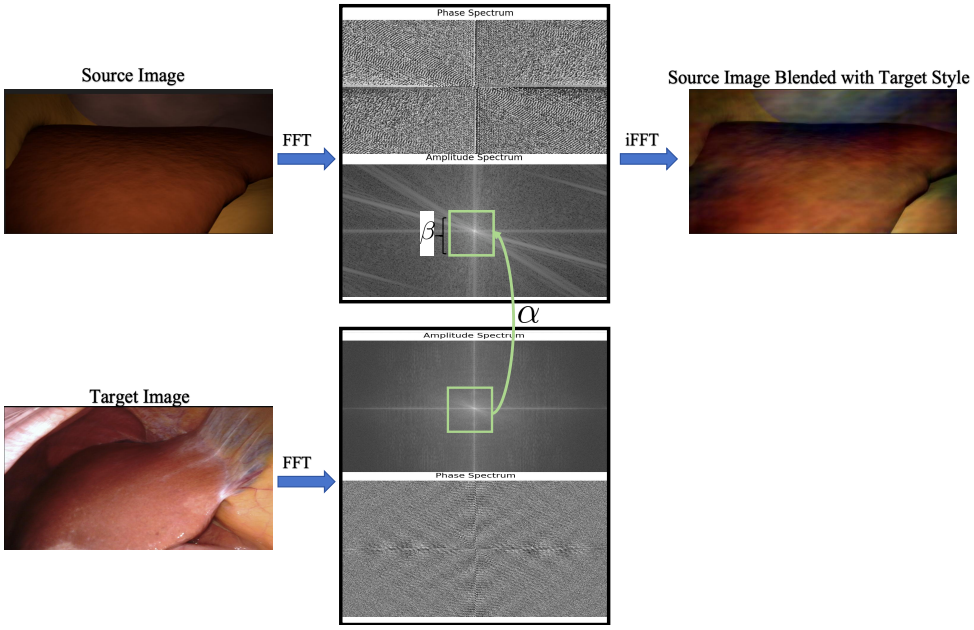

Figure 3. Illustration of the FDA-Blend procedure: FDA-Blend transfers the style of a target domain image into a source image via low-frequency amplitude blending. Specifically, both images are transformed using the Fast Fourier Transform (FFT) to obtain their amplitude and phase spectra. The central region of the amplitude spectrum, corresponding to low-frequency components, is extracted based on a ratio parameter $\beta \in [0, 1]$, which defines the proportion between the side length of the selected square and that of the full spectrum. This parameter is resolution-independent. Within the selected low-frequency region, the source amplitude is linearly blended with that of the target image using a mixing coefficient α, while the phase spectrum of the source image remains unchanged. The modified frequency representation is then transformed back to the spatial domain via inverse FFT (iFFT), resulting in an image that preserves the semantic content of the source while exhibiting the style characteristics of the target domain.

Given the adapted source dataset $D^{s\to t}$, we can train the student network by minimizing the following cross-entropy loss:

$$\mathcal{L}_{CE} = \mathrm{CE}(\mathbf{y}, f_S(X^{s\to t})) \tag{11}$$

where f_S denotes the student network, $X^{s\to t}$ denotes the image obtained by applying FDA-Blend to the source image X, and $\mathbf{y}$ denotes the ground truth corresponding to X.

3.3. Entropy Minimization Loss

To regularize the model, we propose minimizing the entropy loss over predictions on the target domain images. This encourages the model to produce confident predictions proactively and penalizes uncertainty in regions where data is dense and predictions are unstable (i.e., high entropy), which are prone to errors. Such a formulation helps to prevent the decision boundary from crossing densely populated regions of the feature space. However, as noted in [17], directly applying entropy as a penalty is ineffective in

low-entropy regions. Therefore, we apply a robust weighted entropy function to address this limitation.

$$\mathcal{L}_{ent} = \sum_i \rho\left(-\left\langle f_S(x_i^t), \log(f_S(x_i^t))\right\rangle\right) \tag{12}$$

where $\rho(x) = (x^2 + 0.001^2)^\eta$ is the Charbonnier penalty function [46]. When the parameter $\eta > 0.5$, the function imposes a stronger penalty on high-entropy predictions compared to low-entropy ones, as shown in Figure 4.

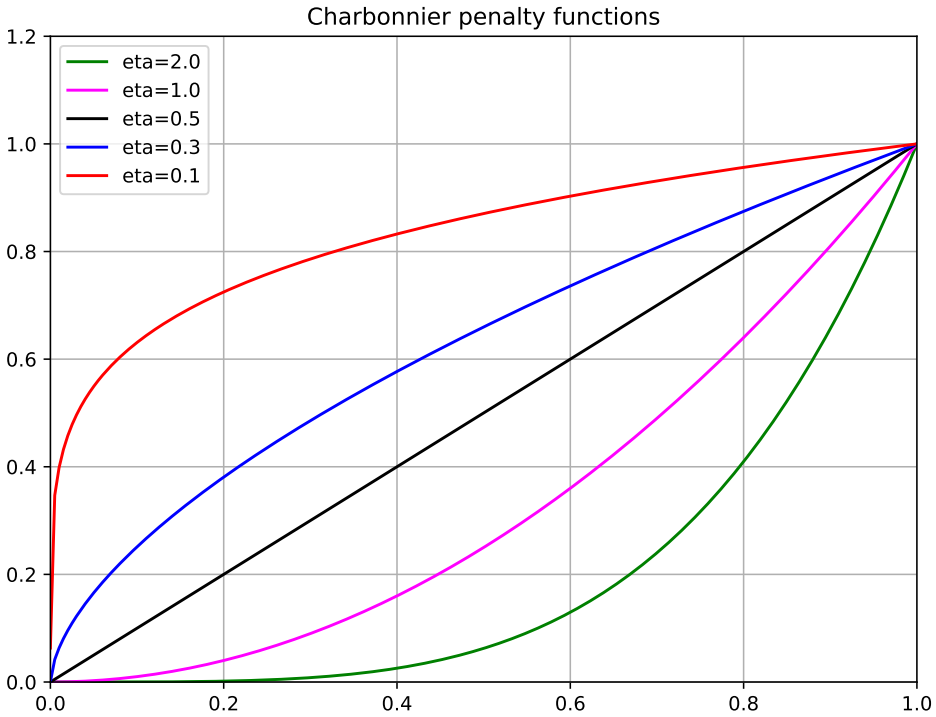

Figure 4. Visualization of the Charbonnier penalty for robust entropy minimization with varying η values.

The overall training objective of the proposed framework can be seamlessly integrated into a typical UDA optimization pipeline. The total loss function is defined as:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{CE} + \lambda_{ent}\mathcal{L}_{ent} \tag{13}$$

4. Experiment

4.1. Datasets and Implementation

4.1.1. Datasets

We evaluate our UDA framework based on two publicly available laparoscopic datasets that provide complementary domain characteristics: a large-scale synthetic-to-real benchmark with fully controllable rendering factors and a real clinical dataset with high-resolution images and expert pixel-wise annotations. Specifically, the Synthetic Laparoscopy dataset is generated from CT-derived anatomical meshes (from 3D-IRCADb 01) and contains 20,000 rendered RGB frames (2000 per virtual patient) with corresponding pixel-wise semantic masks and depth maps. In addition, the dataset authors release a style-translated synthetic set produced with an image-to-image translation pipeline, where each simulated frame is translated with multiple sampled styles (up to five variants per frame), resulting in 100,000 translated images while retaining the original pixel-level annotations. The Dresden Surgical Anatomy Dataset contains 13,195 real laparoscopic frames collected from 32 robot-assisted rectal surgeries at a resolution of 1920 × 1080, with pixel-wise labels

for 11 anatomical structures. To promote annotation diversity, the dataset is curated across multiple surgeries with a capped number of frames per surgery, yielding organ-specific subsets with consistent semantic definitions.

In this study, we utilize the Synthetic Laparoscopy dataset [47] and the Dresden Surgical Anatomy Dataset [48] as the source and target domains, respectively, to assess the effectiveness of our proposed unsupervised domain adaptation (UDA) method for real-world laparoscopic image segmentation. For our experiments, we specifically use the subset of 1023 liver-segmented images from 31 distinct patients to assess the domain adaptation performance on real clinical data.

To comprehensively validate generalization under different source-domain appearances, we construct two adaptation settings. (1) Simulated $\rightarrow$ Real: we use 2000 original synthetic images from a single virtual patient as labeled source data and adapt to unlabeled real images. (2) Translated Simulated $\rightarrow$ Real: we use the corresponding style-translated variants of the same 2000 samples (one translated image per original frame) as the labeled source domain, and adapt to the same unlabeled real target domain. These two settings share identical target data, enabling a controlled comparison of domain adaptation performance under different source-domain inputs.

4.1.2. Evaluation Metric

In this study, we adopt Intersection over Union (IoU) as the primary evaluation metric for laparoscopic semantic segmentation. IoU directly quantifies the region-level overlap between the predicted mask and the ground truth, and is widely used in medical image segmentation where class imbalance is common; compared with pixel accuracy, IoU penalizes both over-segmentation (false positives) and under-segmentation (false negatives) in a balanced manner. For a class c, IoU is computed as follows:

$$\mathrm{IoU}_c = \frac{|\hat{Y}_c \cap Y_c|}{|\hat{Y}_c \cup Y_c|} = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{14}$$

where $\hat{Y}_c$ and Y_c denote the predicted and ground-truth pixel sets of class c, and TP_c, FP_c, and FN_c are the numbers of true-positive, false-positive, and false-negative pixels, respectively. Since our evaluation focuses on the liver structure in the target domain, we report the class-specific IoU for the liver, i.e., $\mathrm{IoU}_{\mathrm{liver}}$.

4.1.3. Implementation Details

All models are implemented in PyTorch (V 1.2.0) and trained on a single NVIDIA Tesla A100 GPU. SegFormer [8] is adopted as the backbone network and initialized with ImageNet-pretrained weights. Training was performed using AdamW [49], with hyperparameters taken from previous works [8,30,45]. During training, input images are resized to 512 × 512 pixels, and the models are optimized for 40k iterations with a batch size of 16. We set $\alpha = 0.5$, $\beta = 0.09$ [15], $h = w = 8$ [11], $k = 0.9$, and $\lambda_{EMA} = 0.999$ [45]. The weight of the entropy loss is set to $\lambda_{ent} = 0.005$, and the exponent parameter $\eta = 2$ [17]. The confidence threshold for pseudo-labels is set to $\tau = 0.968$ [11,30,34,45].

4.1.4. Baselines

We benchmark our method against the current leading approaches, including representative UDA methods such as MIC [45], MFC [11], and FDA [15]. In addition, we include two reference settings for comparison: (1) a source-only baseline where the model is trained on the source domain and directly evaluated on the target domain in a zero-shot manner, and (2) a fully supervised setting where the model is trained with ground-truth labels from the target domain. These two settings serve as the lower and upper bounds, respectively,

for evaluating the effectiveness of our UDA method. For a fair comparison, all methods are implemented using the same semantic segmentation backbone, with identical initialization and optimizer settings.

4.2. Main Results

Table 1 summarizes the quantitative results of liver segmentation under two experimental settings: simulated images →, real images, and translated simulated images → real images. We compare our method with several strong baselines and observe consistent improvements across both settings, with our method outperforming all baselines. Following common practice in prior works [30,45,50], we include a fully supervised model trained on the target domain as a theoretical upper bound, providing a reference for the maximum achievable performance under ideal supervision. Notably, our method achieves substantial gains under the translated-to-real setting and demonstrates competitive performance compared to the fully supervised upper bound. Qualitative comparisons are presented in Figures 5 and 6, showcasing challenging scenarios such as complex backgrounds and varying illumination conditions. Previous methods failed to produce satisfactory results under these conditions. In contrast, our method produces segmentation outputs that align more closely with the ground truth, demonstrating superior robustness and accuracy.

Table 1. Comparison of liver segmentation under two domain adaptation settings: The evaluation is conducted on two domain adaptation settings: from simulated images to real images, and from translated simulated images to real images. The bolded data represents SoTA.

Methods	***Simulated* → *Real*** **IoU↑**	***Translated* → *Real*** **IoU↑**
Fully Supervised	68.29	68.29
Zero-shot (Source only)	23.27	30.27
FDA [15]	35.26	40.26
MIC [45]	38.07	42.88
MFC [11]	40.11	46.44
Ours	**47.27**	**55.23**

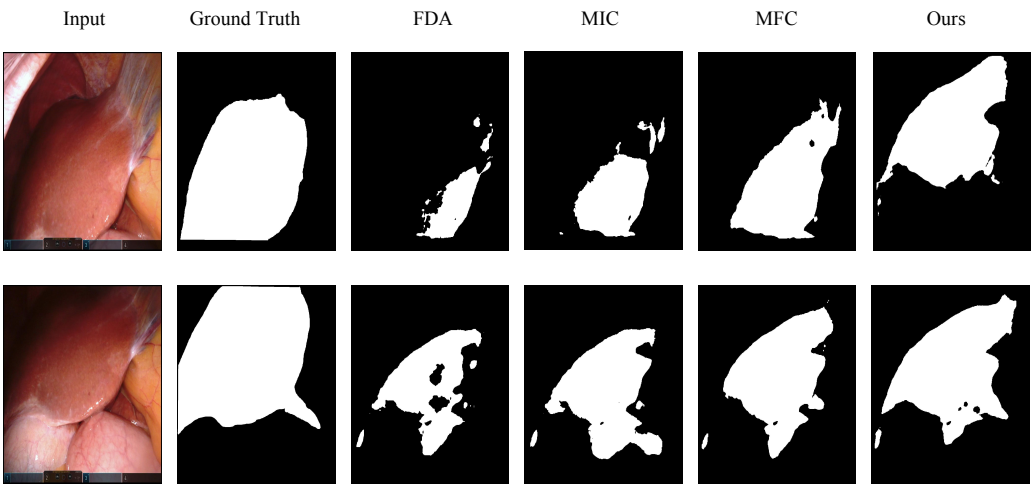

Figure 5. Qualitative results under the simulated → real domain adaptation setting.

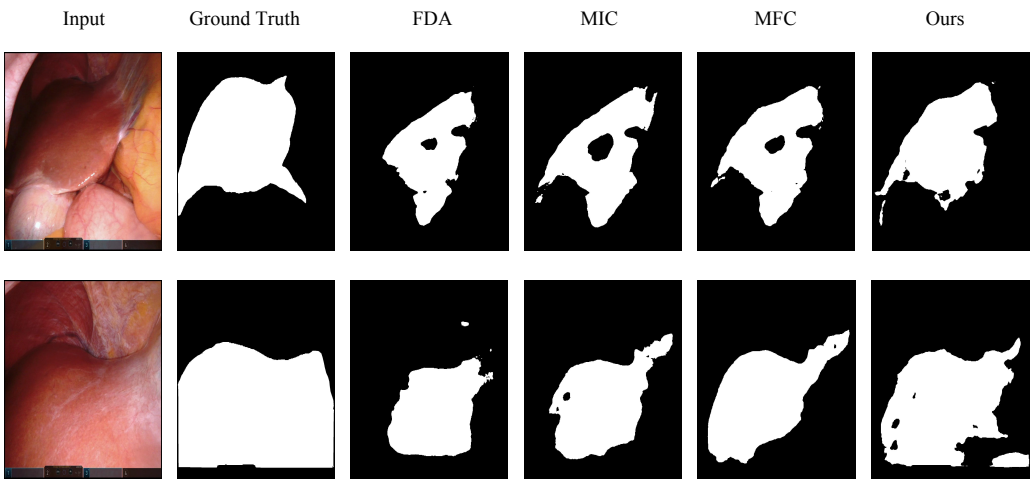

Figure 6. Qualitative results under the translated → real domain adaptation setting.

4.3. Ablation Studies

We conduct comprehensive ablation studies under two settings (simulated images → real images and translated simulated images → real images) to assess the contribution of each component in our framework. Quantitative results are reported in Table 2 and discussed in detail below.

Table 2. Ablation results under two domain adaptation settings: We evaluate the impact of each major component in our framework. w/o RFM removes the Radial Frequency Masking (RFM) module directly from the self-training pipeline. w/o FDA − Blend removes the FDA-Blend module from the supervised training stage. w/o Entropy Loss omits the entropy loss from the self-training pipeline. w/MFC module replaces the RFM with the masking strategy used in MFC. w/FDA replaces the FDA-Blend module with original FDA transfer method. w/Equation (4) preserves the central low-frequency components according to the definition in Equation (4). The bolded data represents SoTA.

Setup	**Simulated → Real IoU↑**	**Translated → Real IoU↑**
w/o RFM	40.23	46.27
w/o FDA-Blend	42.89	48.66
w/o Entropy Loss	43.22	50.19
w/MFC module	43.47	51.26
w/FDA	44.68	51.87
w/Equation (4)	45.66	52.11
full	**47.27**	**55.23**

4.3.1. Ablations on Radial Frequency Masking

To investigate the impact of Radial Frequency Masking (RFM), we compare the performance of our framework with RFM to a variant where RFM is replaced by simple Gaussian noise perturbation applied to the Student network's input images, following the setting described in the Mean Teacher framework [16]. As shown in Table 2, this substitution results in a significant drop in IoU (see w/o RFM in Table 2). These findings highlight the critical role of RFM in encouraging the model to recover information from masked frequency components and learn consistent representations across high- and low-frequency information.

To further investigate the impact of Radial Frequency Masking (RFM), we compare our full framework with a variant in which RFM is replaced by the masking used in MFC [11], applied to the student network's input images following the setting in [11]. To ensure fair comparison, all other components remain unchanged. This substitution (w/MFC module)

leads to a noticeable performance drop, suggesting that our RFM-based masking strategy is more effective for the UDA setting of laparoscopic surgical image segmentation.

Furthermore, previous studies have found that retaining low-frequency information can lead to better performance [11]. However, we demonstrate that low-frequency information also plays a role in our task, although high-frequency components contribute more significantly. Instead of adopting the full RFM-based masking strategy, which masks both high- and low-frequency components, we preserve the central low-frequency region as defined in Equation (4) [11]. This modification leads to a drop in IoU (see w/Equation (4) in Table 2). This indicates that low-frequency information should not be entirely preserved and masking low-frequency components plays a role in our UDA task, which is a key strength of our RFM design.

4.3.2. Ablations on FDA-Blend

To assess the impact of FDA-Blend on the domain adaptation process, we conduct an ablation study by removing FDA-Blend and directly feeding the source images into the Student network. As shown in Table 2, this variant results in a noticeable degradation in IoU. The performance drop indicates that, without frequency-domain alignment, the statistical discrepancy between source and target domains remains significant, making it difficult for the Student network to learn generalizable features during early training. In contrast, FDA-Blend mitigates this domain gap by replacing the amplitude components in the frequency domain, effectively transforming source images to better match the target domain distribution.

To further assess the impact of FDA-Blend on the domain adaptation process, we replace FDA-Blend with the original FDA [15] transfer method. As shown in Table 2, this modification (w/FDA) leads to a substantial drop in IoU. This can be attributed to the fact that full-amplitude substitution tends to cause severe color shifts under large inter-domain illumination differences, while partial blending offers a more stable and effective adaptation signal.

4.3.3. Ablations on Entropy Loss

We further conduct an ablation study to investigate the role of the Entropy Loss [17]. As reported in Table 2, removing this component results in a notable drop in IoU, indicating that it plays an important role in overall performance. The Entropy Loss introduces an information-theoretic regularization of the output probability distribution of the unlabeled target data, which constrains prediction entropy and discourages indecisive or ambiguous outputs.

5. Discussions

To address the substantial domain gap between synthetic and real images, we propose a frequency-domain perspective for unsupervised domain adaptation, incorporating three key modules: RFM, FDA-Blend, and Entropy Minimization. These modules collectively enhance the model's robustness. In the discussion, we further analyze the experimental results and provide a detailed examination of the advantages introduced by each module.

5.1. RFM

We propose a novel frequency-domain masking strategy and employ a consistency loss to promote regularization between high- and low-frequency representations learned from real images. Ablation results (Table 2) demonstrate that the RFM module plays a central role in the performance improvement of our framework among all modules. When replacing the RFM module with the MFC module (w/MFC module), a significant drop in IoU is observed. This can be attributed to our frequency masking strategy, which assigns

higher masking probabilities to high-frequency components based on their radial distance from the spectral center. In contrast, MFC [11] adopts uniform random masking while fully preserving low-frequency regions.

The advantage of our masking strategy lies in two main aspects. First, assigning higher masking probabilities to high-frequency components is particularly beneficial for laparoscopic surgery images, which often exhibit complex textures, illumination variations, and cluttered backgrounds. These characteristics are primarily reflected in the high-frequency components of the spectrum, to which deep neural networks are especially sensitive. Second, unlike MFC, which fully preserves low-frequency content, our strategy assigns a nonzero masking probability across the entire frequency spectrum, including the central low-frequency region. The ablation study in Table 2 (w/Equation (4)) shows that fully retaining low-frequency information leads to a performance drop, suggesting that selectively masking low-frequency components also contributes to enhancing the model's robustness to spectral variations.

In our work, we adopt the Mean Teacher framework [16] for self-training. While the student network receives frequency-masked inputs, the teacher network observes full-spectrum images and generates supervision signals. This setup encourages the student to recover consistent predictions despite incomplete frequency information. The exponential moving average (EMA) update of the teacher further stabilizes training and improves pseudo-label quality.

5.2. Sensitivity Analysis of the Mixing Coefficient α

As shown in Figure 7 and Table 3, as α increases from 0.1 to 1.0, the IoU first improves and then degrades, with the best performance achieved at $\alpha = 0.5$, indicating a clear unimodal dependency on the appearance mixing strength.

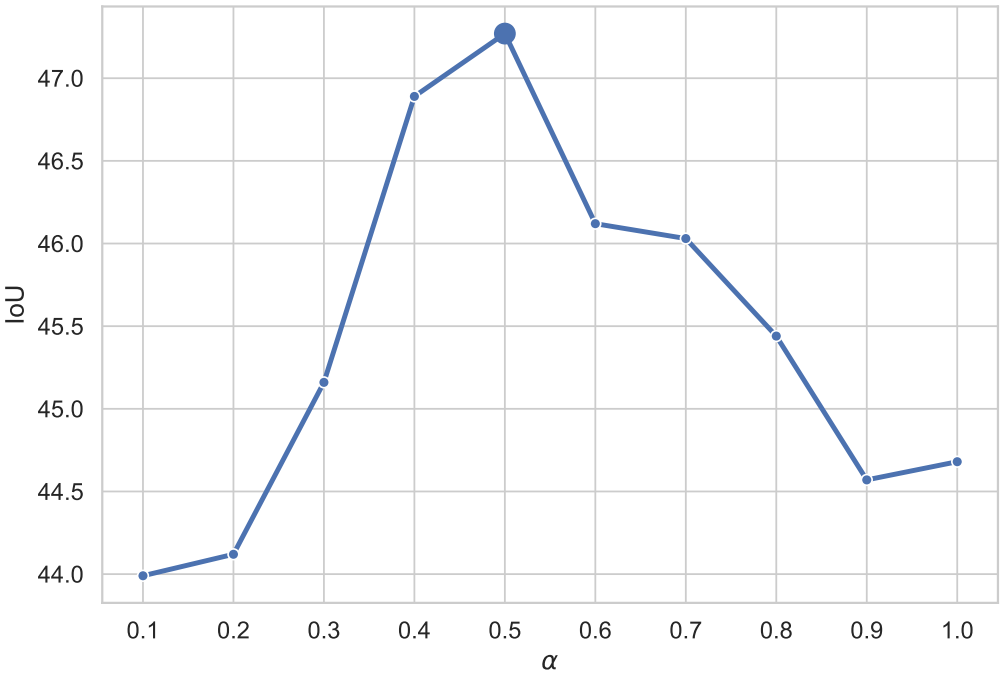

Figure 7. Sensitivity analysis of the mixing coefficient α: IoU under the simulated $\rightarrow$ real setting as a function of α. The performance first improves and then degrades as α increases, achieving the best result at $\alpha = 0.5$.

From a signal processing perspective, FDA-Blend performs interpolation only in the low-frequency amplitude space while preserving the phase information. Low-frequency amplitudes primarily determine global appearance attributes such as overall brightness, color distribution, and illumination structure, whereas phase information encodes geometric structure and semantic layout. As a result, the mixing coefficient α effectively controls the strength of appearance alignment toward the target domain. When α is small, the low-frequency amplitudes are still dominated by the source domain, and the generated samples remain visually distant from the target domain. Consequently, the model is exposed to only weak target-style perturbations, and the domain gap is insufficiently reduced, leading to limited performance gains.

As α increases, a larger portion of target-domain low-frequency information is incorporated, gradually shifting the global appearance of the training samples toward the target-domain distribution. This enables the model to adapt more effectively to target-domain imaging conditions and appearance statistics while maintaining stable semantic structure, resulting in improved performance. However, when α becomes too large and approaches 1.0, overly aggressive low-frequency replacement can substantially alter global illumination and color contrast, weakening the implicit consistency assumption between source-domain annotations and the input images. Although phase information remains unchanged, such extreme appearance shifts may be amplified by nonlinear components in the network, such as normalization layers and activation functions, thereby affecting boundary localization and local discriminative cues, which ultimately leads to training instability and performance degradation.

Therefore, the effect of α exhibits a clear unimodal trend, where performance first increases and then decreases as α grows. The optimal value of α corresponds to a balance between sufficient appearance alignment and semantic stability. Within this balanced regime, the model can effectively reduce appearance discrepancies between the source and target domains without introducing excessive appearance–label mismatch, yielding the best domain adaptation performance.

Table 3. Sensitivity analysis of the mixing coefficient α: We report the IoU under the simulated $\rightarrow$ real setting. The bolded data represents SoTA.

α	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
IoU	43.99	44.12	45.16	46.89	**47.27**	46.12	46.03	45.44	44.57	44.68

5.3. Entropy Minimization

In our setting, Entropy Minimization serves as an information-theoretic regularizer for unlabeled pixels. Penalizing high-entropy posterior distributions suppresses uncertain predictions, pushing the decision boundary toward low-density regions in feature space in accordance with the cluster assumption. This low-density separation stabilizes optimization, curbs the propagation of noisy pseudo-labels, and yields better-calibrated confidence estimates. As a result, the model exhibits enhanced robustness and improved generalization on the target domain.

5.4. Sensitivity Analysis of the Coefficient k in RFM

In Radial Frequency Masking (RFM), the hyperparameter k controls the overall masking strength by scaling the radius-dependent masking probability in the frequency domain. Intuitively, a larger k increases the masking rate for higher-frequency regions, which are more likely to contain domain-specific variations in laparoscopic videos (e.g., specular highlights, smoke, and subtle texture differences). Therefore, k mediates a trade-off be-

tween suppressing domain-specific high-frequency artifacts and preserving informative high-frequency cues that are also useful for accurate boundary delineation.

To examine the sensitivity of RFM to k, we conduct a sweep under the simulated→real setting (See Table 4). The results show a clear saturation trend: performance improves substantially when increasing k from 0.1 to 0.5, indicating that stronger high-frequency perturbations are beneficial for bridging the simulated-to-real gap. Beyond $k \geq 0.5$, the IoU becomes relatively insensitive to k (variation within 0.16 from 0.5 to 0.9), suggesting that RFM is stable within a reasonable parameter range. Meanwhile, an overly large value ($k = 1.0$) slightly degrades performance, which is consistent with the above trade-off—excessive masking may remove high-frequency details that are informative for precise liver boundaries and may also make the teacher–student consistency learning less reliable. Based on both stability and the best observed IoU, we use $k = 0.9$ as the default setting in our experiments.

Table 4. Sensitivity analysis of the coefficient k in RFM: We report the IoU under the simulated → real setting. The bolded data represents SoTA.

k	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**1.0**
IoU	45.11	46.79	47.11	47.21	**47.27**	46.67

6. Conclusions

This paper studies unsupervised domain adaptation (UDA) for semantic segmentation of laparoscopic images, aiming to bridge the domain shift between computer-simulated data and real intra-operative scenes without requiring pixel-wise annotations in the target domain.

We propose a frequency-aware UDA framework that combines (i) FDA-Blend for appearance alignment in supervised source training, (ii) RFM to introduce structured frequency perturbations for robust teacher–student consistency learning on unlabeled target images, and (iii) Entropy Minimization to encourage confident and stable predictions during self-training.

Scientific novelty. Differing from prior approaches that rely solely on spatial augmentations or global style transfer, our method explicitly leverages frequency-domain operations within a unified UDA pipeline. In particular, RFM provides a principled way to regularize target-domain learning by selectively masking frequency components, which promotes domain-invariant representations while preserving the semantic structure essential for surgical scene understanding. Together with FDA-Blend and entropy regularization, the proposed design yields consistent improvements over representative UDA baselines under both simulated →real and translated→ real settings.

Practical significance. Our framework offers an effective solution to laparoscopic image segmentation in real clinical data by primarily exploiting readily available simulated data, substantially reducing the dependency on costly expert annotations for each new surgical domain, device, or hospital. This makes the proposed method particularly appealing for scalable deployment and rapid adaptation in data-scarce clinical environments.

Prospects for real-world applications. The proposed approach can serve as a core component for downstream developments such as intra-operative scene understanding and organ boundary delineation, which may support surgical navigation, safety-critical region awareness, and workflow analysis. In addition, the frequency-aware adaptation strategy is compatible with practical deployment constraints (e.g., varying imaging devices and acquisition conditions) and can be extended to settings such as continual/online adaptation, semi-supervised adaptation with a few annotated target frames, and efficiency-oriented implementations for near-real-time inference.

Limitations and future work. A current limitation is that our experimental evaluation focuses on the liver class. In future work, we will extend the evaluation to additional anatomical structures and multi-class segmentation, and further investigate robustness across broader clinical scenarios (e.g., different procedures, centers, and imaging systems). Moreover, to provide a more rigorous robustness assessment, we will repeat experiments with multiple random seeds and report the mean and standard deviation, and we will conduct statistical significance tests (e.g., paired t-test or Wilcoxon signed-rank test) to quantify the consistency of the gains.

Author Contributions: Conceptualization, H.D.; methodology, G.Z.; software, H.D.; validation, H.D.; formal analysis, H.D.; investigation, H.D. and G.Z.; resources, G.Z.; data curation, H.D.; writing—original draft preparation, H.D.; writing—review and editing, H.D. and G.Z.; visualization, H.D.; supervision, G.Z.; project administration, G.Z.; funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the public availability of the dataset.

Informed Consent Statement: Patient consent was waived due to the public availability of the dataset.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
2. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
3. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
4. Twinanda, A.P.; Shehata, S.; Mutter, D.; Marescaux, J.; de Mathelin, M.; Padoy, N. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Trans. Med. Imaging* **2017**, *36*, 86–97. [CrossRef] [PubMed]
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
7. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [CrossRef]
8. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*; Morgan Kaufmann Publishers, Inc.: Burlington, MA, USA, 2021; Volume 34, pp. 12077–12090.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*; Uitgever NIPS Foundation: San Diego, CA, USA, 2017; Volume 30.
10. Silva, B.; Oliveira, B.; Morais, P.; Buschle, L.R.; Correia–Pinto, J.; Lima, E.; Vilaça, J.L. Analysis of Current Deep Learning Networks for Semantic Segmentation of Anatomical Structures in Laparoscopic Surgery. In Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, UK, 11–15 July 2022; pp. 3502–3505. [CrossRef]

11. Zhao, X.; Hayashi, Y.; Oda, M.; Kitasaka, T.; Mori, K. Masked Frequency Consistency for Domain-Adaptive Semantic Segmentation of Laparoscopic Images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2023*; Springer Nature: Cham, Switzerland, 2023; pp. 663–673.
12. Csurka, G. A Comprehensive Survey on Domain Adaptation for Visual Applications. In *Domain Adaptation in Computer Vision Applications*; Springer International Publishing: Cham, Switzerland, 2017; pp. 1–35. [CrossRef]
13. Patel, V.M.; Gopalan, R.; Li, R.; Chellappa, R. Visual Domain Adaptation: A Survey of Recent Advances. *IEEE Signal Process. Mag.* **2015**, *32*, 53–69. [CrossRef]
14. Wang, M.; Deng, W. Deep Visual Domain Adaptation: A Survey. *Neurocomputing* **2018**, *312*, 135–153.. [CrossRef]
15. Yang, Y.; Soatto, S. FDA: Fourier Domain Adaptation for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4084–4094. [CrossRef]
16. Tarvainen, A.; Valpola, H. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In *Advances in Neural Information Processing Systems (NeurIPS)*; Uitgever NIPS Foundation: San Diego, CA, USA, 2017; Volume 30.
17. Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2512–2521. [CrossRef]
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
19. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*; Morgan Kaufmann Publishers, Inc.: Burlington, MA, USA, 2012; Volume 25.
20. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
21. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.-W.; Heng, P.-A. H-DenseUNet: Hybrid Densely Connected U-Net for Liver and Tumor Segmentation from CT Volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [CrossRef] [PubMed]
22. Zhou, Z.; Siddiquee, M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–11.
23. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In Proceedings of the ICASSP 2020—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–8 May 2020; pp. 1055–1059. [CrossRef]
24. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 424–432.
25. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [CrossRef]
26. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net Architecture Design for Medical Image Segmentation through the Lens of Transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef] [PubMed]
27. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *Computer Vision—ECCV 2022 Workshops*; Springer Nature: Cham, Switzerland, 2023; pp. 205–218.
28. Xie, Y.; Zhang, J.; Shen, C.; Xia, Y. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*; Springer International Publishing: Cham, Switzerland, 2021; pp. 171–180.
29. Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; Chandraker, M. Learning to Adapt Structured Output Space for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7472–7481. [CrossRef]
30. Hoyer, L.; Dai, D.; Van Gool, L. Domain Adaptive and Generalizable Network Architectures and Training Strategies for Semantic Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 220–235. [CrossRef] [PubMed]
31. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 1989–1998.

32. Saito, K.; Watanabe, K.; Ushiku, Y.; Harada, T. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3723–3732. [CrossRef]
33. Zou, Y.; Yu, Z.; Vijaya Kumar, B.V.K.; Wang, J. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In *Computer Vision—ECCV 2018*; Springer International Publishing: Cham, Switzerland, 2018; pp. 297–313.
34. Tranheden, W.; Olsson, V.; Pinto, J.; Svensson, L. DACS: Domain Adaptation via Cross-Domain Mixed Sampling. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Virtual, 5–9 January 2021; pp. 1378–1388. [CrossRef]
35. Yang, C.; Guo, X.; Chen, Z.; Yuan, Y. Source Free Domain Adaptation for Medical Image Segmentation with Fourier Style Mining. *Med. Image Anal.* **2022**, *79*, 102457. [CrossRef] [PubMed]
36. Liu, S.; Yin, S.; Qu, L.; Wang, M.; Song, Z. A Structure-Aware Framework of Unsupervised Cross-Modality Domain Adaptation via Frequency and Spatial Knowledge Distillation. *IEEE Trans. Med. Imaging* **2023**, *42*, 3919–3931. [CrossRef] [PubMed]
37. Wang, Y.; Cheng, J.; Chen, Y.; Shao, S.; Zhu, L.; Wu, Z.; Liu, T.; Zhu, H. FVP: Fourier Visual Prompting for Source-Free Unsupervised Domain Adaptation of Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2023**, *42*, 3738–3751. [CrossRef] [PubMed]
38. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
39. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
40. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988. [CrossRef]
41. Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT Pre-Training of Image Transformers. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022.
42. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. SimMIM: A Simple Framework for Masked Image Modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9643–9653. [CrossRef]
43. Kakogeorgiou, I.; Gidaris, S.; Psomas, B.; Avrithis, Y.; Bursuc, A.; Karantzalos, K.; Komodakis, N. What to Hide from Your Students: Attention-Guided Masked Image Modeling. In *Computer Vision—ECCV 2022*; Springer Nature: Cham, Switzerland, 2022; pp. 300–318. [CrossRef]
44. Li, Z.; Chen, Z.; Yang, F.; Li, W.; Zhu, Y.; Zhao, C.; Deng, R.; Wu, L.; Zhao, R.; Tang, M.; et al. MST: Masked Self-Supervised Transformer for Visual Representation. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 13165–13176.
45. Hoyer, L.; Dai, D.; Wang, H.; Van Gool, L. MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 11721–11732. [CrossRef]
46. Bruhn, A.; Weickert, J. Towards Ultimate Motion Estimation: Combining Highest Accuracy with Real-Time Performance. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV), Beijing, China, 17–20 October 2005; Volume 1, pp. 749–755. [CrossRef]
47. Pfeiffer, M.; Funke, I.; Robu, M.R.; Bodenstedt, S.; Strenger, L.; Engelhardt, S.; Roß, T.; Clarkson, M.J.; Gurusamy, K.; Davidson, B.R.; et al. Generating Large Labeled Data Sets for Laparoscopic Image Processing Tasks Using Unpaired Image-to-Image Translation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*; Springer International Publishing: Cham, Switzerland, 2019; pp. 119–127.
48. Carstens, M.; Rinner, F.M.; Bodenstedt, S.; Jenke, A.C.; Weitz, J.; Distler, M.; Speidel, S.; Kolbinger, F.R. The Dresden Surgical Anatomy Dataset for Abdominal Organ Segmentation in Surgical Data Science. Dataset on Figshare. 2022. [CrossRef]
49. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**, arXiv:1711.05101. [CrossRef]
50. Hoyer, L.; Dai, D.; Van Gool, L. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9914–9925. [CrossRef]