applied sciences

MDPI

Article

Unsupervised Domain Adaptation for Automatic Polyp Segmentation Using Synthetic Data

Ioanna Malli [1], Ioannis A. Vezakis [1], Ioannis Kakkos [1,2,*], Theodosis Kalamatianos [2] and George K. Matsopoulos [1]

[1] Biomedical Engineering Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Zografou Polytechnic Campus, 15772 Athens, Greece; ivezakis@biomed.ntua.gr (I.A.V.); gmatsopoulos@biomed.ntua.gr (G.K.M.)
[2] Department of Biomedical Engineering, University of West Attica, 12243 Athens, Greece; tkalamatianos@uniwa.gr
* Correspondence: ikakkos@biomed.ntua.gr

Abstract

Colorectal cancer is a significant health concern that can often be prevented through early detection of precancerous polyps during routine screenings. Although artificial intelligence (AI) methods have shown potential in reducing polyp miss rates, clinical adoption remains limited due to concerns over patient privacy, limited access to annotated data, and the high cost of expert labeling. To address these challenges, we propose an unsupervised domain adaptation (UDA) approach that leverages a fully synthetic colonoscopy dataset, SynthColon, and adapts it to real-world, unlabeled data. Our method builds on the DAFormer framework and integrates a Transformer-based hierarchical encoder, a context-aware feature fusion decoder, and a self-training strategy. We evaluate our approach on the Kvasir-SEG and CVC-ClinicDB datasets. Results show that our method achieves improved segmentation performance of 69% mIoU compared to the baseline approach from the original SynthColon study and remains competitive with models trained on enhanced versions of the dataset.

Keywords: unsupervised domain adaptation; polyp segmentation; domain shift; mix vision transformer; self-training; synthetic medical data

Academic Editor: Bhanu Prakash KN

Received: 5 August 2025
Revised: 2 September 2025
Accepted: 4 September 2025
Published: 8 September 2025

Citation: Malli, I.; Vezakis, I.A.; Kakkos, I.; Kalamatianos, T.; Matsopoulos, G.K. Unsupervised Domain Adaptation for Automatic Polyp Segmentation Using Synthetic Data. *Appl. Sci.* **2025**, *15*, 9829. https://doi.org/10.3390/app15179829

1. Introduction

Colorectal cancer is a very common and dangerous type of cancer that begins in the areas of the colon or rectum, with the risk of developing it estimated to be 1 in 24 in men and 1 in 26 in women [1]. At the same time, early-stage detection dramatically improves the survival outcome of the patient [2]. Patients diagnosed at Stage 1 have a survival rates of approximately 90%, which falls to just 10% for Stage 4 [3].

Diagnosis is typically made by colonoscopy screening [4]. However, this procedure can be very time consuming and prone to errors. In [5], a false negative rate of 8% was reported in colonic screenings, with the majority of errors attributed to human error. Artificial intelligence-powered computer-aided detection and diagnosis (AI-CAD) systems have the potential to assist with real-time polyp detection and segmentation during endoscopy by significantly reducing miss rates and improving consistency across operators [6].

One of the main challenges in applying AI to medicine is data scarcity, as medical datasets often come from clinical trials or small-scale collaborations [7]. As a result,

they tend to be limited in size and lack the demographic diversity needed for broad generalization [8].

Obtaining high-quality annotated datasets is another major bottleneck [9], since it requires the contribution of expert healthcare professionals. Moreover, there are significant concerns regarding data privacy, particularly when working with real patient data [10], which are governed by comprehensive legal frameworks such as the EU's General Data Protection Regulation (GDPR) [11].

Lastly, real-world datasets are collected using different imaging devices and settings. This introduces domain shifts that degrade the performance of models trained on one dataset when applied to another [12]. These factors make it difficult to train robust segmentation models for real-world deployment.

Thus, our objective is to develop a robust segmentation pipeline capable of segmenting polyps in colonoscopy images while addressing the challenges described above. To this end, we hold the assumption that access to annotation is completely unavailable. To address the issue of domain shift, we employ a method with strong generalization capabilities. Additionally, by avoiding the use of labeled patient data, our approach mitigates concerns related to data privacy regulations.

Although several methods exist to improve model robustness across domains, including single domain generalization [13], semi-supervised learning [14], and data augmentation strategies [15], this paper explores unsupervised domain adaptation (UDA). The main advantage of UDA is that it achieves high performance in an unlabeled target domain, as is the case here.

We use Synth-Colon [16], a large synthetic dataset generated using computer graphics and refined via CycleGAN-style transfer [17], as our source domain. The target domain need not be fixed; in this work, we primarily evaluate our method on Kvasir-SEG [18], a widely used and publicly available endoscopic dataset. In this manner, we demonstrate that it is possible to leverage knowledge from a fully synthetic dataset, which can be significantly larger than a typical real-world one, and it comes with easily obtainable masks. This knowledge can be used to train a model on real-world datasets relevant to endoscopic image analysis, even in the absence of annotations.

The network and UDA framework are built on DAFormer [19], a state-of-the-art architecture for unsupervised domain adaptation. It combines a variant of the Vision Transformer (ViT) backbone with a confidence-based pseudo-label generation strategy. DAFormer has demonstrated strong performance in multiclass segmentation tasks, such as those related to autonomous driving.

Our work is organized as follows:

- In Section 2, we discuss recent advancements in Transformer architectures in computer vision, review various polyp segmentation methods, and give an overview of available UDA methodologies.
- In Section 3, we provide a detailed explanation of how we utilized the SynthColon and UDA methodologies.
- In Section 5, we evaluate DAFormer under different scenarios and compare it with other architectures and under different target datasets.
- In Section 6, we reflect on the insights gained through extensive experimentation.
- In Section 7, we summarize our findings and propose potential directions for future work.

To ensure the reproducibility of our research, our code and dataset splits are publicly available on Zenodo (https://doi.org/10.5281/zenodo.17007712, accessed on 30 August 2025).

2. Background and Related Work

2.1. Transformers

Transformers were introduced in [20] and revolutionized the field of Natural Language Processing, with the Multi-Head Self-Attention mechanism (MHSA). Recognizing that images could similarly benefit from modeling global relationships, researchers extended the self-attention mechanism to computer vision. The Vision Transformer (ViT) [21] architecture was the first successful attempt to apply a pure Transformer model to the task of image classification. While the Vision Transformer (ViT) [21] was largely successful in image classification, its flat, non-hierarchical structure and global attention mechanism made it less suitable for dense prediction tasks like semantic segmentation due to limited spatial resolution, high computational cost, and need for extensive pretraining. To address this, several hierarchical Transformer variants were proposed. Pyramid Vision Transformers (PVTs) [22] and Mix Vision Transformers (MiTs) [23] introduced multi-stage architectures with progressively reduced spatial dimensions, similar to CNNs, enabling multi-scale feature extraction. MiTs also adopted overlapping patches to preserve local continuity. Both PVTv2 [24] and MiTs integrated convolutional layers into their feed-forward networks to enhance spatial inductive bias. Swin Transformer [25] further reduced complexity by applying self-attention within local windows and enabling global interaction through a shifted window scheme.

2.2. Polyp Segmentation

Various architectures for the semantic segmentation of intestinal polyps have been proposed and evaluated on datasets like Kvasir-SEG [18] and CVC-ClinicDB [26]. Earlier networks utilized CNNs in an encoder–decoder design, including the popular U-Net [27] and its variants [18,28–30], with or without skip connections [31]. Further improvements came through specialized attention modules [32,33], which enhanced decoder performance in challenging regions like boundaries. Recently, Transformer-based models have been explored extensively—either in hybrid architectures [34] or as standalone feature extractors [35–38]. Vision Transformer variants are often chosen for their ability to capture global information early on via self-attention, allowing every patch to attend to every other [21]. In contrast, CNNs focus first on local patterns and gradually build up global context. While this global perspective gives ViTs an edge in generalization across domains, it also introduces challenges. Self-attention may over-mix unrelated regions, leading to uniformly distributed attention weights and degraded focus. Consequently, ViTs often require heavy pretraining or strong augmentation strategies to counteract the lack of built-in locality.

Foundational models have also begun to gain traction in polyp segmentation. For example, ref. [39] introduces a conditional mask loss that adapts to the type of annotation available for each training sample. Similarly, the Segment Anything Model (SAM) [40] has been extensively studied as a foundational approach for segmentation tasks. It shows strong performance on natural images, without fine-tuning, and it performs poorly in the medical domain, particularly on endoscopic data [41]. Most SAM-based approaches for polyp segmentation therefore either fine-tune the mask decoder on polyp datasets [42] or adapt the prompting strategy to better capture polyp boundaries [43,44]. In addition, zero-shot [45] and weakly supervised adaptations of SAM [46] have been explored. Despite their rising potential, SAM-based methods have not achieved state-of-the-art results in the task of polyp segmentation. Furthermore, they require labeled data or user-provided prompts in the form of bounding boxes and points. Thus, they do not address the fully automatic unsupervised domain adaptation setting considered in this work.

2.3. Unsupervised Domain Adaptation (UDA)

A model trained on one dataset often experiences a drop in performance when evaluated on another, even if the datasets are semantically similar. This is due to the domain shift between their underlying data distributions. In Unsupervised Domain Adaptation (UDA), the distribution the model is trained on is called the source domain—typically large and annotated—while the target domain is unlabeled, and UDA methods assume no access to its annotations.

UDA approaches can broadly be divided into two categories: adversarial methods and self-training methods. Adversarial methods employ Generative Adversarial Networks (GANs) [47] to align the distributions of the source and target domains. This alignment can occur at various levels, including the input level [48], feature level [49], output level [50], and patch level [51].

Synth-Colon, which is discussed in detail in a later section, is an example of input-level domain alignment. It transforms images from the source domain to resemble those of the target domain using a CycleGAN [17], a type of GAN that enables unpaired image-to-image translation.

Self-training methods rely on the generation of pseudo-labels [52] for the unlabeled target domain, which are then iteratively refined during training. The quality of these pseudo-labels is progressively improved through confidence-based filtering strategies. These strategies may include binary filtering [53,54], which discards low-confidence predictions; soft filtering, which assigns weights to predictions based on confidence scores [55]; or hybrid approaches that combine elements of both [56].

Another distinction in such methodologies is whether they are offline [57,58] or online. Online methods often employ teacher–student networks [59] in order to enforce consistency regularization [60]. In this setup, the teacher is a stable version of the student model that is updated through Expotential Moving Average and is tasked with generating pseudo-labels. A common loss is the Mean Squared Error (MSE) between the student and teacher output. In addition, mixing strategies are used, such as ClassMix [61] in DACS [62] and ClassDrop in [63].

3. Methodology

3.1. Problem Formulation

Let us define a source domain,

$$\mathcal{D}_S = \left\{ (x_s^{(i)}, y_s^{(i)}) \right\}_{i=1}^{N_s}, \quad x_s^{(i)} \sim P_S, \quad y_s^{(i)} \in \mathcal{Y} \tag{1}$$

where $x_s^{(i)} \in \mathcal{X}_S$ is an input image, $y_s^{(i)}$ is the corresponding ground truth segmentation mask, and P_S is the source domain distribution.

We also define a target domain,

$$\mathcal{D}_T = \left\{ x_t^{(i)} \right\}_{i=1}^{N_t}, \quad x_t^{(i)} \sim P_T \tag{2}$$

where $x_t^{(i)} \in \mathcal{X}_T$, but the corresponding labels $y_t^{(i)}$ are not available.

Our goal is the creation of a segmentation model,

$$f_\theta : \mathcal{X} \to \mathcal{Y} \tag{3}$$

that extracts knowledge from labeled source data $\mathcal{D}_S$ and uses them to either generalize or improve its generalization on an unlabeled target distribution P_T, despite the domain shift $P_S - P_T$.

In the absence of target domain labels, the naive approach would dictate we only calculate cross-entropy (CE) for the source domain. In our binary problem, for a single image i, this translates to

$$\begin{aligned}\mathcal{L}_S^{(i)} = -\sum_{j=1}^{H\times W} \Big[& y_S^{(i,j)} \log f_\theta(x_S^{(i)})^{(j)} \\ & + (1 - y_S^{(i,j)}) \log(1 - f_\theta(x_S^{(i)})^{(j)}) \Big]\end{aligned} \tag{4}$$

Here, binary cross-entropy loss is applied pixel by pixel, then summed over the whole image.

However, training using this loss function would result in a model overfitted to the source domain that cannot generalize well to the target.

Thus, most UDA methods suggest introducing an additional term $\mathcal{L}_T^{(i)}$ to incorporate information from the target distribution.

$$\mathcal{L}^{(i)} = \mathcal{L}_S^{(i)} + \mathcal{L}_T^{(i)} \tag{5}$$

In this manner, the model can learn from both the ground truth of the synthetic source dataset and its own confident guesses on the target dataset. This combination gradually transfers knowledge from the synthetic domain to the real one.

To compute $\mathcal{L}_T$, we opt to follow a self-training (ST) architecture that is state-of-the-art for the segmentation of street scenes, DAFormer.

In [19], a stable teacher network h_ϕ produces pseudo-labels using target domain data samples:

$$p_T^{(i,j)} = \left[\arg\max_{c' \in \{0,1\}} h_\phi(x_T^{(i)})^{(j,c')} = 1 \right] \tag{6}$$

where $[\dots]$ denotes the Iverson bracket:

$$[a = b] = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

So for the binary problem, $p_T^{(i,j)} = 1$ if the predicted class is foreground (polyp) or $p_T^{(i,j)} = 0$ if it is background (intestine). The teacher model follows an EMA model and its weights are not updated during backpropagation.

The pseudo-labels carry high uncertainty, so we require a confidence measure to guide training. In the original paper, for the multi-class case, the authors use the ratio of pixels whose softmax probability exceeds a threshold τ to the total number of pixels in the image. In our binary segmentation setting, we modify this by computing the ratio of predicted polyp pixels whose confidence exceeds τ to the total number of pixels predicted as polyp.

Thus, the confidence ratio $q_\tau(i)$ for image i, given a confidence threshold τ, is defined as

$$q_\tau(i) = \frac{|\{p \in P_i \mid s_i(p) > \tau\}|}{|P_i| + \varepsilon}$$

where ε is a small constant to avoid division by zero when no pixels are predicted as polyp and $s_i(p) \in [0, 1]$ denotes the confidence (sigmoid probability) of pixel p being a polyp.

Now we compute the unsupervised loss as the binary cross-entropy between the teacher's pseudo-labels and the student's predictions, weighted by the confidence score:

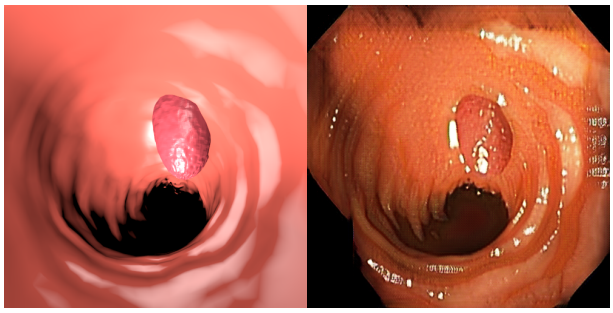
$$\mathcal{L}_T^{(i)} = -\sum_{j=1}^{H\times W} \sum_{c=1}^{C} q_T^{(i)} \, p_T^{(i,j,c)} \log f_\theta\left(x_T^{(i)}\right)^{(j,c)}. \tag{8}$$

where $H \times W$ is the total number of pixels in the image, C is the number of classes, $q_T^{(i)}$ is the confidence score for image i, and $p_T^{(i,j,c)}$ is the pseudo-label probability for class c at pixel j in image i.

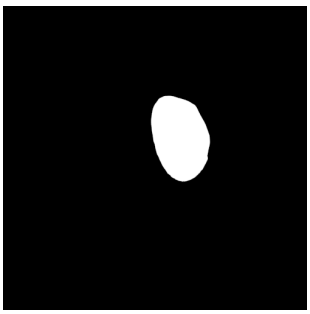
Self-training methods benefit greatly from augmentations being applied to the target dataset. In the DACS methodology we follow, the target dataset is augmented with color jitter, Gaussian blur and flip, as well as ClassMix.

3.2. Synthetic Dataset Generation

We distinguish between two types of synthetic datasets: (1) those generated from real datasets using generative models such as VAEs or GANs [64], which offer increased diversity and privacy preservation, and (2) fully synthetic datasets created from scratch using 3D modeling tools [65,66]. In this work, we utilize SynthColon [16], a fully synthetic dataset generated with Blender to simulate colonoscopy scenes. Synth Colon is such a synthetic dataset, created through the use of a 3D modeling software, specifically Blender. The process involves creating the colon geometry through distorting a cone to simulate the natural bumps and curves of the colon and adding a small ellipsoid to function as the polyp. Then, base colors are added in pink and red hues, and finally different types of lightning are introduced to simulate the light emitted by the endoscope and add depth to the images. To reduce the domain gap between synthetic and real images, SynthColon applies a CycleGAN-based image-to-image translation trained on the Kvasir dataset, enhancing realism by transferring textures, lighting, and dataset-specific artifacts. In Figure 1, we observe the added realism of the image (b) that is captured after the style-transfer step compared to the raw 3D image on the left. On the right, the binary mask is shown: as the polyp is placed by a computer software inside the polyp environment, its location and shape is known, and thus a mask can be created automatically.



(**a**) (**b**)



(**c**)

Figure 1. SynthColon dataset: (**a**) a raw 3D colocoscopy image sample, as produced by the modeling engine. (**b**) Image sample after the style-transfer step using CycleGAN. (**c**) The corresponding segmentation mask, where the polyp region is highlighted in white.

CUT-Seg [67] also proposes a style transfer approach for adapting synthetic colonoscopy data to real domains. While SynthColon relies on CycleGAN-based two-sided consistency to translate images, CUT-Seg replaces cycle consistency with patch-wise contrastive learning. CycleGAN's two-sided consistency is powerful in preventing the model from discarding important structure. However, it can also be overly restrictive, since the network may focus on learning a reversible mapping rather than producing realistic textures. In contrast, CUT-Seg replaces this constraint with a patch-wise contrastive objective, which encourages corresponding patches in the input and output to share similar representations while pushing apart unrelated patches. This is particularly effective in colonoscopy, where preserving small structures such as polyps and mucosal folds is critical. By removing the cycle constraint, the approach remains more flexible while still maintaining local structural fidelity.

Although CUT-Seg generally achieves stronger realism and segmentation performance, SynthColon is publicly released and provides a large, ready-to-use synthetic dataset with paired masks. This enables standardized and fair comparisons across methods and ensures our results can be easily reproduced by others.

3.3. Architecture

The architecture is based on [19], a state-of-the-art method for UDA. The encoder part is a Mix Vision Transformer (MiT) [23]. It processes the input using 4×4 patches and produces four feature maps at resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the original input, with each map having progressively higher channel dimensions. By setting the stride smaller than the kernel size, patches are overlapping, which preserves spatial continuity. MiT also features a self-attention mechanism optimized for efficiency by reducing the number of tokens used to compute the key K and value V matrix, using a sequence reduction ratio. Unlike standard Vision Transformers (ViTs) that rely on positional encodings, this architecture omits them and instead captures positional information using 3×3 convolutions within the MLP module. This design enhances the model's robustness to variations in input resolution during inference.

For the decoder, each of the four hierarchical feature maps is first projected through a 1×1 convolutional block to a shared embedding dimension, then upsampled to match the spatial dimensions of the highest-resolution feature map and subsequently stacked to form a single multi-scale representation. To fuse this information effectively, the decoder employs a smart aggregation strategy inspired by Atrous Spatial Pyramid Pooling (ASPP), similar to the design used in DeepLabV3. Here, the stacked features are passed through multiple parallel 3×3 depthwise separable convolutions with varying dilation rates, enabling the network to capture spatial context at multiple scales. The encoder contributes semantic depth, while the ASPP head refines the spatial understanding, selectively enhancing features across levels based on their utility for accurate segmentation at each spatial location.

For the Self-Training method, we closely follow [62]. In DACS, each pixel of a pseudo-label is weighted by the ratio of confident pixels in the image. In our binary problem, this translates to the ratio of pixels that are classified as polyp and whose prediction probability exceeds the threshold to all the pixels classified as polyp. Furthermore, the ClassMix strategy is also adapted to the binary problem: With 50% probability, we paste source polyp regions (from ground truth) onto the target background. Otherwise, we paste target pseudo-polyp regions onto the source background. ClassMix promotes feature-level consistency without explicit loss functions by requiring the model to learn from perturbed pseudo-labels. It also enforces object-level understanding across domains and mitigates source overfitting. The process is described in Figure 2.

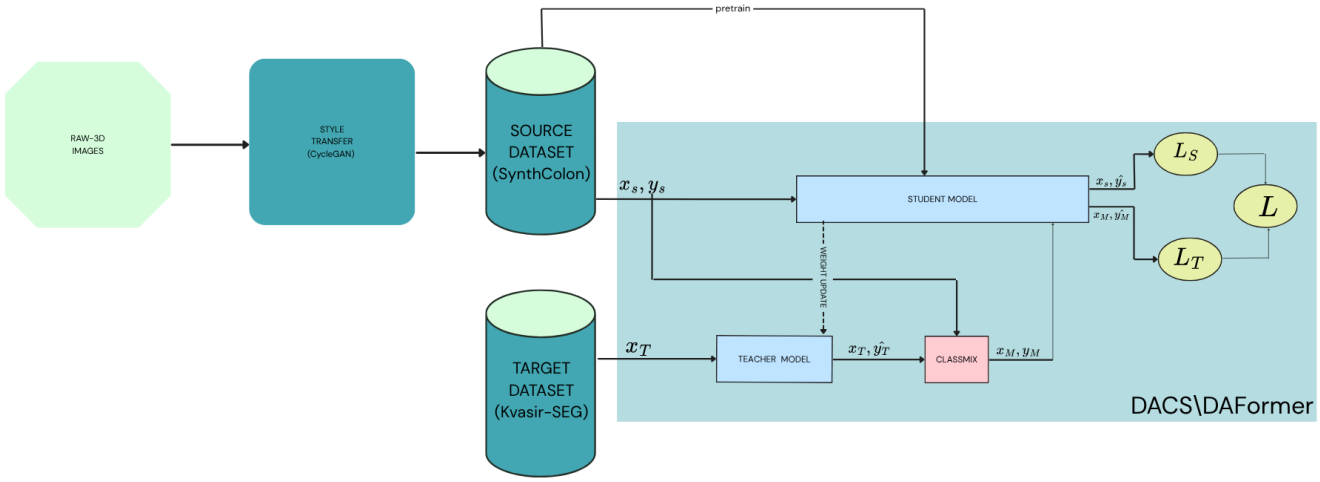


Figure 2. Visualization of the proposed UDA pipeline, which consists of two main stages. First, the raw 3D images are refined via CycleGAN-based style transfer to better match real-world imaging conditions. Then, a self-training loop is employed, along with the ClassMix operation.

3.4. Training Pipeline

We integrate the methods described above into our final training pipeline, aiming to train a segmentation model on a polyp dataset that lacks annotations in the target domain. The pipeline begins with preprocessing both the source and target datasets. This includes resizing images to match the input dimensions expected by the encoder, applying standard data augmentations such as horizontal and vertical flips, color jittering, affine and elastic transformations, and normalizing images using the ImageNet-1k mean and standard deviation.

Through experiments described in Section 5.2, we observe that pretraining the segmentor on the source dataset for 1–2 epochs provides a strong initialization for the UDA pipeline and consistently improves final performance on the target dataset. Therefore, we include a brief supervised pretraining phase on the source domain prior to commencing unsupervised domain adaptation.

Since the source dataset is larger than the target dataset, we construct a unified dataset by cycling through the target samples during training. Specifically, for each index in the source dataset, we retrieve the corresponding target sample using modular indexing. This ensures balanced sampling without discarding any target data. The EMA teacher model is initialized, with the same weights as the standard student model.

Each batch from the unified dataset is processed as illustrated in Figure 2. The source samples are passed through the student model to compute the supervised loss. The teacher model's weights are updated using an exponential moving average (EMA) of the student's weights, according to the following equation:

$$\theta_{\text{EMA}} \leftarrow \alpha \cdot \theta_{\text{EMA}} + (1 - \alpha) \cdot \theta \tag{9}$$

Target samples are passed through the teacher model to generate pseudo-labels, along with a single confidence score per image, representing the proportion of high-confidence pixels. This confidence score is then used to weigh the contribution of each pixel in the unsupervised loss calculation.

Subsequently, the source and target images, together with their ground truth and pseudo-labels, are used to generate mixed images and labels, as described in Section 3.3. The mixed image is passed through the student model to compute the unsupervised loss.

The supervised and unsupervised losses are then combined, and the student model is updated via backpropagation. This process is repeated for each training step.

4. Experiments

To optimize performance and identify the most effective model configuration, we conducted a series of experiments involving different encoder depths and an alternative architecture that incorporates convolutional layers. Additionally, we evaluated the impact of using both the raw and style-transferred versions of the source dataset. We also explored various pretraining strategies. To assess the effectiveness of our approach, we compared it against other architectures that utilize the Synth-Colon dataset and evaluated performance on different target datasets.

Experimental Setup

In our implementation, the Synth-Colon dataset serves as the source domain. Kvasir-Seg is used as the target domain, with CVC-ClinicDB included to evaluate the robustness of the method across datasets.

The images in the Synth-Colon dataset have a fixed resolution of 500 × 500 pixels, whereas the Kvasir-SEG dataset contains images with varying resolutions, ranging from 332 × 487 to 1920 × 1072 pixels. CVC-ClinincDB contains 612 images with their corresponding masks that come with a resolution of 384 × 288. As part of the preprocessing pipeline, all images from all datasets are resized to 512 × 512 pixels to match the input resolution expected by the encoder.

For pretraining, we apply a range of augmentations using Albumentations library to image samples from all datasets. The same pipeline is applied for both the source domain pretraining we describe later and the UDA adaptation step. Images and masks are resized to 512 × 512 with nearest-neighbor interpolation. We include random flips, color jittering, affine transformations, and elastic distortions to simulate variability in orientation, appearance, and geometry. Moreover, images are normalized using ImageNet statistics to match the pretraining of the encoder. Finally, the images are converted to tensors to be compatible with the Torch framework.

The original implementation of the DAFormer methodology is publicly available on github and based on the mmsegmentation framework. However, due to mmsegmentation having many dependency constraints that make it difficult to be deployed in cloud environments, we opt to create custom code based on standard pytorch. For the DAFormer architecture, we test three encoders from the MiT series, namely MiT-B2, MiT-B3, and MiT-B5. Although all variants share the same patch size, embedding dimensions ([64, 128, 320, 512]), and attention heads ([1, 2, 5, 8]), their depth of the attention block varies. Specifically, MiT-B1 uses [2, 2, 2, 2] (2 attention blocks per stage), MiT-B2 uses [3, 4, 6, 3], and MiT-B5 uses [3, 6, 40, 3]. All encoders are pretrained on ImageNet-1k.

All models are trained using the `AdamW` optimizer with a constant learning rate of 1×10^{-5} and weight decay of 1×10^{-4}. For the loss function, we use Binary Cross Entropy.

The confidence threshold for pseudo-labels is set to 0.968. The EMA model's smoothing factor α is set to 0.99, following the value set in the DAFormer official implementation.

The batch size used in UDA is 2, whereas in one-dataset (fully supervised) experiments, it is 4. All experiments are carried out in a NVIDIA Tesla T4 GPU that has a 16 GB RAM.

5. Experimental Results

5.1. Training on the Raw 3D-Generated Dataset

We conduct a generalization test by training on the source dataset Synth-Colon with and without the Cycle-GAN step. We evaluate on the target Kvasir-Seg. The results show

that the mIoU does not surpass 20% and the Dice score remains bellow 25% in the case of the raw dataset, showing that the model struggles severely to generalize and predict meaningful masks. This is evident in the resulting annotation masks, as shown in Figure 3.

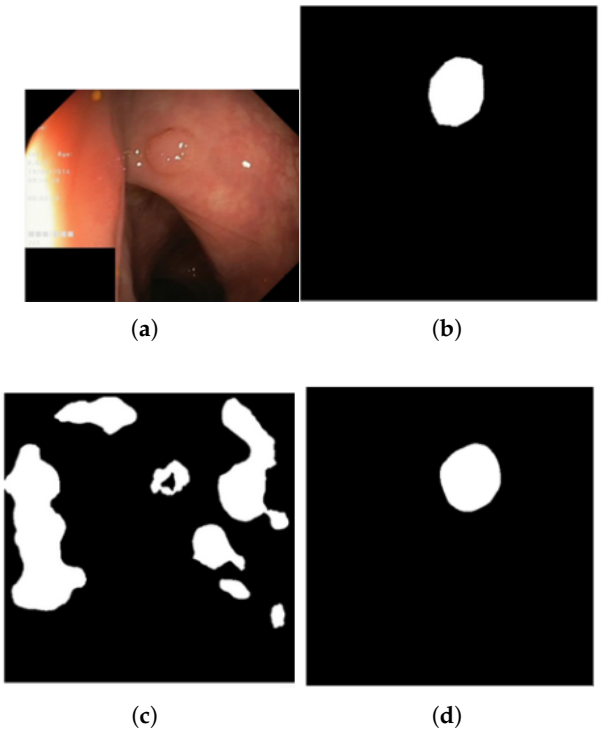


Figure 3. Difference between annotations produced by using the Raw 3D-designed datasets vs. the dataset that was refined using a CycleGAN. (**a**) Image from the target domain. (**b**) Ground truth segmentation mask. (**c**) Segmentation mask produced when source domain has not been refined with CycleGAN. (**d**) Segmentation mask produced when source domain has been refined with CycleGAN.

5.2. Effect of Pretraining on Source

In Table 1, we demonstrate the effects of model performance under two different pretraining strategies. In both of them, the encoder is pretrained on ImageNet as we already mentioned, but we also experiment with a second pretraining of the full segmentation architecture on the source dataset. Since we assume target annotations are unavailable, standard validation-based early stopping is not applicable. Instead, we find that pretraining for 1–2 epochs is sufficient, with source-domain mIoU exceeding 94% serving as a reliable heuristic for stopping. It should be noted that all subsequent experiments include this additional supervised pretraining step.

Table 1. Best model results based on different pretraining strategies. Bold font indicates the highest values.

Model	Encoder Pretrained on ImageNet	Segmentor Pretrained on Source Dataset	mIoU	Dice
DAFormer	✓	–	0.4520	0.5915
DAFormer	✓	✓	**0.7021**	**0.8157**

5.3. Comparison of Backbone Depth

We evaluate the performance of three MiT encoder variants of increasing depth (MiT-B2, MiT-B3, and MiT-B5) within our UDA pipeline. These are selected from the six MiT models available. The results are presented in Table 2. In this table, *Src.-only* refers to a

baseline model trained in a fully supervised manner on the source dataset and evaluated on the target validation set. *UDA* corresponds to our full unsupervised domain adaptation pipeline, which incorporates DACS as previously described. *Oracle* represents a fully supervised model trained and evaluated on the target dataset, serving as an upper performance bound that assumes access to target annotations. Lastly, *Rel.* (Relative) denotes the ratio between UDA and Oracle performance, indicating how closely each method approaches the upper-bound performance.

Table 2. Results from training models with different encoder sizes on the Kvasir-SEG target dataset. The "Rel." column denotes relative performance with respect to the Oracle. Results are reported in mIoU. Bold font indicates the highest values.

Encoder	# Params (M)	Src-Only (%)	UDA (%)	Oracle (%)	Rel. (%)
MiT-B2	27.9	49.28 ± 0.2	58.3 ± 3.0	82.49 ± 0.42	70.68
MiT-B3	47.7	50.23 ± 3.7	61.6 ± 2.6	84.43 ± 0.24	72.96
MiT-B5	**85.1**	**54.85** ± 2.8	**69.0** ± 0.9	**85.18** ± 1.18	**81.00**

5.4. Comparison of Different Network Architectures

We compare the performance of DAFormer on the SynthColon→Kvasir-SEG task with two different architectures that specifically use CNNs. We opt to use UNet as a baseline architecture since it is widely known and accepted in the community, EffiSegNet which is a state-of-the-art architecture on Kvasir-Seg dataset, and our DAFormer network. Their details are depicted in Table 3.

Table 3. Results for training our UDA pipeline with different segmentation architectures. Results are reported in mIoU. Standard deviations are reported only for our method, as they were not provided in the original documentation of the other methods.

Architecture	# Params (M)	Src-Only (%)	UDA (%)	Oracle (%)	Rel. (%)
UNet	3.1	25.84 ± 2.1	31.86 ± 0.3	74.6 [33]	41.55
EffiSegNet (B6)	40.7	40.03 ± 1.14	59.35 ± 0.67	90.56 [68]	65.55
DAFormer (mit-b3)	47.7	50.23 ± 3.7	61.6 ± 2.6	84.43 ± 0.24	72.96

It should be noted that the three architectures were trained under different training parameters, suitable to each architecture. For the sake of a fair comparison, we compare the version of EffiSegNet-B6 to that of DAFormer with the MiT-B3 encoder. These two networks have similar inputs (528 × 528 vs. 512 × 512) and the smallest difference in parameter count among all possible variant combinations of the two architectures (700K).

5.5. Comparison of Current Synth-Colon→Kvasir-Seg

Table 4 compares mIoU performance of existing methods trained on the SynthColon dataset and evaluated on Kvasir-SEG without access to target labels. SynthColon and CUT-Seg are style-transfer domain adaptation techniques that use GANs to enhance the realism of synthetic 3D images. Their reported segmentation results are obtained through fully supervised training on the translated datasets using HarDNet-MSEG as the segmentation backbone and evaluating performance on the target domain.

Importantly, these methods rely solely on the source data during training and do not incorporate any information from the target domain, which is used exclusively for evaluation. As such, they are categorized as *source-only* (Src-only) approaches.

Table 4. Comparison of mIoU and mDice (%) for methods using the SynthColon dataset and its variants as the source domain when adapting to Kvasir-SEG. No annotations from the target dataset are assumed.

Method	Src-Only		UDA		Oracle	
	mIoU	mDice	mIoU	mDice	mIoU	mDice
DAFormer	54.85 ± 0.2	73.2 ± 0.4	69.0 ± 0.9	80.67 ± 0.01	85.18 ± 1.1	89.04 ± 0.3
SynthColon [16]	52.7	75.9	–	–	85.70	90.4
CUT-Seg [67]	62.1	70.2	–	–	85.70	90.4
PL-CUT-Seg [69]	–	–	68.77	78.08	85.70	90.4

PL-CUT-Seg introduces self-training with pseudo-labels, aligning with our definition of *UDA*. For completeness, we also report *oracle* scores obtained by fully supervised training on Kvasir-SEG. For SynthColon and CUT-Seg, this corresponds to the performance of HarDNet-MSEG [70] trained directly on that dataset.

5.6. Synth-Colon→CVC-ClinicDB

To evaluate the robustness of our method on datasets different from the one used during its image-to-image translation preprocessing, we assess its performance on the CVC-ClinicDB dataset.

In Table 5, we explore the performance of two of our models under different experimental setups, all evaluated on the CVC-ClinicDB dataset. The column *UDA (SynthColon→Kvasir-Seg)* refers to a setup where the model was trained using UDA from SynthColon to Kvasir-Seg, as described in the previous sections. We then evaluate the resulting checkpoint on the CVC-ClinicDB dataset, even though it was not used during training or adaptation.

Table 5. Comparison of domain adaptation performance from SynthColon to CVC-ClinicDB. Results are reported in mean IoU (mIoU). DAFormer and EffiSegNet employ the proposed self-training method.

Method	Src-Only	UDA (SynthColon →CVC-ClinicDB)	UDA (SynthColon →Kvasir)	Oracle
SynthColon (baseline)	45.7 *	-	-	88.20 *
DAFormer (MiT-B5)	53.37 ± 1.5	60.70 ± 0.8	66.62 ± 0.2	85.18 ± 1.18
EffiSegNet (B6)	35.77 ± 1.1	54.04 ± 0.9	50.16 ± 0.7	89.50 *

* Standard deviation not provided in the original documentation.

5.7. Computational Efficiency

We measure wall-clock training time on a Google Colab runtime with NVIDIA Tesla T4 GPU with 16 GB VRAM. Pretraining on the synthetic source dataset (20k images, batch size 4) required 5000 iterations, with an average iteration time of 1.75 s, totaling ∼2 h 25 m for one epoch. During UDA fine-tuning (batch size 2), each iteration took ∼3.5 s, and model performance plateaued after ∼2000 iterations (∼2 h). In total, our method converges in ∼4 h 25 m of training. Unlike GAN-based approaches (e.g., CUT-seg, PL-CUT-seg), which additionally require adversarial training on the full dataset, our approach only trains the segmentation model and thus achieves substantially lower computational cost.

6. Discussion

6.1. Key Observations

Our experimental results revealed multiple performance-enhancing strategies that we incorporated into our final pipeline, as well as valuable insights.

6.1.1. Training on the Raw 3D-Generated Dataset

Firstly, we confirmed that the raw SynthColon images exhibit a domain gap that is too large for effective adaptation, leading us to use the CycleGAN-refined version of the dataset instead. While bypassing the style-transfer step would save computational resources and reduce the risk of introducing artifacts, our experiments (Section 5.1) showed that the raw samples differ significantly from real clinical data. This discrepancy arises not only from the lack of realistic textures but also from the inherent simplicity of the 3D modeling process. As a result, even advanced UDA architectures struggle to generalize effectively when faced with such a wide domain shift.

6.1.2. Effect of Pretraining on Source

In addition to ImageNet-1K pretraining of the encoder, we introduced a second pretraining stage for the full segmentation model on the source dataset. This decision stems from the fact that ImageNet contains no medical-specific features, such as tissues or lesions, limiting its usefulness to low-level representations like edges and textures. By pretraining the entire segmentor on the source domain, we provided a task-specific initialization better aligned with our downstream objective. As shown in Table 1, this additional step resulted in a substantial performance boost, increasing mIoU by 25% and Dice score by approximately 20%.

6.1.3. Comparison of Backbone Depth

Regarding the size of the encoder, it becomes evident from Table 2 that increasing the encoder depth consistently improves performance across all training scenarios: source-only, UDA, and Oracle. Notably, MiT-B5 outperforms the shallower variants MiT-B2 and MiT-B3 in every metric.

What is particularly interesting, however, is the trend in the *Oracle* column: increasing the number of parameters by a factor of 1.7 (MiT-b2→MiT-b3) yields an approximate 2% improvement in Oracle performance, while a three-fold increase (MiT-b2→MiT-b5) results in only a 2.7% gain. In contrast, in the UDA setting, the same parameter increases lead to gains of approximately 3.3% and 10.7%, respectively, which are significantly larger improvements. As a result, MiT-B5 achieves an 81% relative score.

This suggests that, although simply adding depth to the network may saturate its fully supervised (Oracle) performance, it substantially improves generalization in the unsupervised domain adaptation setting. We attribute this to the model being able to capture richer, domain-invariant representations.

6.1.4. Comparison of Different Network Architectures

We also investigated the choice of architecture, with a particular focus on the comparison between convolutional networks (CNNs) and transformer-based models. We present the results in Table 3. The two networks we tested achieve comparable performance on the UDA task, with DAFormer showing a modest 2% improvement. Notably, while EffiSegNet demonstrates superior performance in fully supervised settings, it falls significantly behind in the Source-only setting (by approximately 10%). This may be attributed to the generalization advantage of transformer-based architectures over CNNs. This advantage could be even more pronounced in our case, where the domain shift involves generalizing from a synthetic dataset generated with 3D graphics to a real-world dataset, rather than from a real dataset to another. The CNN's reliance on texture instead of shape may become a disadvantage when the synthetic images exhibit inaccurate or inconsistent textures due to imperfections in CycleGAN-based image-to-image translation.

6.1.5. Comparison of Current Synth-Colon→Kvasir-Seg

To further assess our pipeline, we compared it against other techniques that use the SynthColon dataset, as summarized in Table 4. In the Oracle setting, which reflects fully supervised training on Kvasir-SEG, DAFormer achieves performance comparable to HarDNet-MSEG [70], the segmentation backbone used in both SynthColon and CUT-Seg. In the source-only setting, DAFormer outperforms SynthColon by about 2% while using the exact same dataset. This indicates that DAFormer provides a stronger segmentation architecture than HarDNet-MSEG. However, DAFormer is slightly outperformed by CUT-Seg. Since CUT-Seg also relies on HarDNet-MSEG, its advantage comes not from a better segmentation network but from access to a more realistic translated dataset, which reduces the domain gap.

When we integrate our full UDA pipeline, however, our method clearly outperforms CUT-Seg, by approximately 7%, highlighting the effectiveness of our DACS-based self-training.

Most notably, our full DAFormer-based UDA pipeline surpasses PL-CUT-Seg by a small margin (69.0% vs. 68.77%). The latter combines self-training with a significantly refined version of the SynthColon dataset, yet our approach achieves comparable or better results. This suggests that our method exhibits strong generalization capabilities, even when faced with a larger domain shift.

While a standardized train/test split for SynthColon is not publicly available—limiting the precision of direct comparisons—the overall results indicate that our approach is at least competitive with current state-of-the-art methods. It is reasonable to expect that integrating the more refined dataset used by CUT-Seg into our pipeline could yield further improvements, potentially positioning our method as the new state of the art in this domain.

Regarding the computational complexity of each method, it should be noted that a direct comparison is difficult: the authors of SynthColon, CUT-Seg, and PL-CUT-Seg do not report training times or number of parameters of their methods.

Moreover, there are significant structural differences between these methods and ours. SynthColon uses an offline CycleGAN to pre-generate synthetic data and then trains the HarDNet-MSEG network on them. CUT-Seg jointly trains a generator, discriminator, and segmentation model in an online adversarial setup, with reported training schedules of up to 300 epochs. PL-CUT-Seg extends this by adding a self-training stage with pseudo-labels, further increasing complexity.

In contrast, our method fully decouples style transfer from segmentation, similarly to SynthColon. In addition, while we adopt a self-training scheme (DACS), the teacher network is an exponential moving average of the student, introducing no additional trainable parameters beyond the segmentation model itself. Decoupling style-transfer from segmentation significantly simplifies the training and lowers the computational overhead, without hurting performance as we have already showcased.

Overall, whereas CUT-Seg and PL-CUT-Seg involve multiple modules and hundreds of training epochs, our pipeline requires only one epoch of source pretraining and $\sim$2000 fine-tuning iterations before convergence. This translates to an order-of-magnitude fewer training iterations and a single-stage optimization process. Although our segmentation network is heavier than HarDNet-MSEG used in previous works, the streamlined training design makes our method more stable and computationally more efficient in practice.

6.1.6. Synth-Colon→CVC-ClinicDB

Finally, in Section 5.6, we present Table 5, where we test the performance of DAFormer and EffiSegNet when the target dataset is different from the one used in style transfer. Specifically, the source dataset used is CVC-ClinicDB. We assess these results by comparing them to what is reported in Tables 2 and 3.

- *Src.Only*: Both models exhibit a performance drop when trained excluseively on SynthColon and evaluated on CVC-ClinicDB. Specifically, DAFormer (MiT-b5) decreases slightly from 54.58% mIoU to 53.37%, while EffiSegNet drops more substantially from 40.03% to 35.77%. Although a decrease is expected since CVC-ClinicDB was not used during style-transfer, the more pronounced decline in EffiSegNet may be attributed to the texture sensitivity inherent to CNN-based architectures.
- *UDA (SynthColon→CVC-ClinicDB)*: Both models show reduced performance compared to their results under UDA from SynthColon to Kvasir-Seg (Tables 2 and 3). Specifically, DAFormer and EffiSegNet drop by 8.3% and 5.31% mIoU, respectively. However, they still outperform their Src.Only counterparts, meaning that our UDA method is capable of providing improved results even in the face of a larger domain shift.
- *UDA (SynthColon→Kvasir-Seg)*: In this setup, we utilize the best checkpoints of our SynthColon→Kvasir-Seg training and evaluate them on CVC-ClinicDB. Surprisingly, this setting results in a smaller performance drop on CVC-ClinicDB compared to the UDA (SynthColon→CVC-ClinicDB) setup. This is unexpected, as we assumed that directly adapting to CVC-ClinicDB would yield better results on that same dataset. This result can be intuitively explained by the fact that our CycleGAN-refined SynthColon dataset is visually closer to Kvasir-Seg due to the style transfer, and Kvasir-Seg in turn is naturally more similar to CVC-ClinicDB, as both are real-world datasets. However, SynthColon and CVC-ClinicDB are not necessarily close. Thus, adapting in two smaller steps, first from synthetic to Kvasir-Seg and then evaluating on CVC-ClinicDB, may generalize better than attempting to bridge the larger domain gap directly. In Figure 4, this improvement can be observed in two challenging samples. In the first image, multiple folds and discolorations may mislead the model, while in the second image, the boundaries are difficult to discern and the polyp texture closely resembles that of the surrounding intestinal walls. In both cases, the SynthColon→Kvasir-Seg→CVC-ClinicDB improves performance over SynthColon→CVC-ClinicDB, clearly showcasing this phenomenon.

To summarize our findings, it is evident that the quality of the source dataset plays a crucial role in achieving strong performance on the target domain. Notably, we demonstrate that without a style-transfer step, the model struggles to converge, underscoring the importance of bridging the domain gap early in the pipeline. This observation aligns with prior work by the authors of [67], who reported improved performance over their earlier approach [16] by incorporating a more advanced style-transfer technique. These results collectively indicate that even a robust domain adaptation method cannot fully compensate for a large domain gap; when the gap remains wide, performance inevitably suffers.

Despite being trained on a less realistic source dataset, our DAFormer-based method competes with PL-CUT-Seg, demonstrating the superior robustness of the DAFormer architecture. However, the greater realism of PL-CUT-Seg's source data keeps it competitive, highlighting that source domain quality remains a bottleneck. The importance of minimizing the domain gap is further illustrated by our SynthColon→CVC-ClinicDB experiment, where performance declined due to the style-transfer model being trained on Kvasir-SEG—a mismatch that introduced harmful artifacts during translation.

6.2. Limitations

Furthermore, it is important to highlight some limitations observed in the Synth-Colon dataset that extend beyond the style-transfer method employed. Specifically, the raw 3D renderings exhibit limited variability relative to the dataset size, particularly in terms of polyp positioning and viewing angles. Most images are captured along the central axis of

the intestinal tube, often showing polyps protruding laterally or, in some cases, seemingly not attached to any tissue, something anatomically unrealistic.

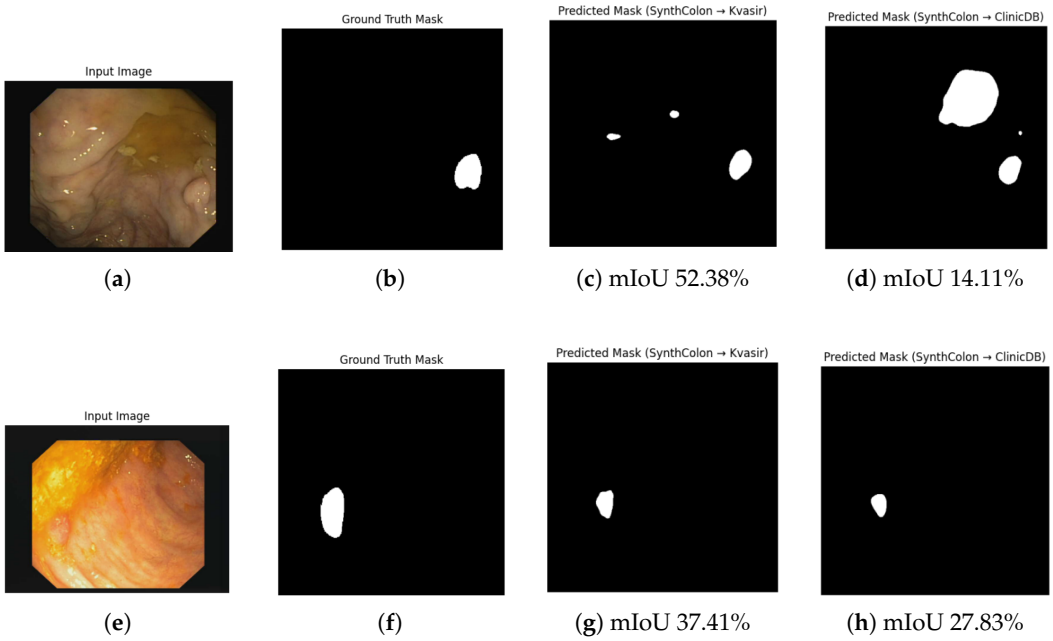


Figure 4. Hard, low-contrast examples. In the first column, images (**a**,**e**) correspond to sample images from the CVC-ClinicDB dataset. In the second column, images (**b**,**f**) correspond to ground-truth segmentation masks. In the third column, images (**c**,**g**) correspond to predictions of the *SynthColon→Kvasir-SEG→CVC-clinic* pipeline. In the fourth column, images (**d**,**h**) correspond to predictions by the *SynthColon→CVC-clinic* pipeline.

In contrast, real-world datasets like Kvasir-SEG include a broader range of camera angles, such as lateral views where only one side of the tissue is visible, as well as a more diverse array of polyp shapes beyond the primarily ellipsoidal forms seen in Synth-Colon. In UDA, model performance is closely linked to the quality and diversity of the source dataset, especially in teacher–student frameworks, where the source domain strongly guides the learning process on the unlabeled target domain.

These results suggest that further work should not only focus on improving the Image-to-Image translation step to produce more realistic texture but also increase the diversity in the raw images through more detailed 3D design.

Our method addresses key challenges in clinical applications of deep learning, such as data scarcity and privacy concerns. Nonetheless, synthetic data have inherent limitations when applied to clinical environments. In particular, they may fail to fully capture the complexity and variability of the true patient population, leading to potential biases and reduced generalization ability. Another risk is that deep learning algorithms primarily reconstruct patterns present in their input, without necessarily reflecting the underlying physiological mechanisms. This may cause the model to overlook clinically relevant cues or generate results that lack medical plausibility [71].

Finally, synthetic images and related methods are not yet regulated by established legislation or legal standards to guide their evaluation and application in clinical environments [72]. Even then, deploying such software would require extensive validation and the development of appropriate protocols to ensure that it does not negatively influence healthcare providers or patient outcomes.

6.3. Future Work

The results we present demonstrate the effectiveness of our method in adapting synthetic to real colonoscopy images for polyp segmentation. However, there remains great potential for improvement.

Firstly, as discussed in Section 6.2, the base 3D renderings of colon–polyp structures could be significantly improved to include more diverse polyp shapes, more anatomically detailed colon geometry, and a wider variety of camera angles. Enhancing these aspects would play a key role in narrowing the domain gap, particularly by introducing greater variation in features that are naturally domain invariant.

In addition, our results highlight the critical role of the image-to-image translation step. Therefore, future research should further investigate image-to-image translation approaches, not only refining the techniques in SynthColon and CUT-Seg but also exploring newer style transfer architectures such as Directional CycleGAN [73] and Structure-Preserving CycleGAN [74]. We hypothesize that integrating the improved CUT-Seg dataset into our DAFormer pipeline could result in a much clearer performance gain over the PL-CUT-Seg approach.

Another direction would be to evaluate different segmentation architectures. Our current results suggest an edge of Transformer-based backbones in generalization performance over CNNs. Still, there is room for experimentation, either by implementing different transformer backbones or introducing special attention modules to guide training in difficult-to-learn areas such as object boundaries [33].

Finally, our experiments show that UDA performance drops when adapting from SynthColon to CVC-ClinicDB. This suggests that a one-size-fits-all pipeline may not be optimal across datasets. We propose developing dataset-specific pipelines, where the unlabeled target dataset is utilized during the style-transfer step to accommodate variations in target domains.

7. Conclusions

In this work, we explored the problem of automatic polyp detection on unlabeled colonoscopy datasets. We proposed a method that leverages unsupervised domain adaptation from a fully labeled synthetic source, which is entirely synthetic, generated using 3D graphics techniques. We adopted DAFormer, a domain adaptation architecture featuring a Transformer-based encoder, and compared its performance to traditional CNN-based models. Our best-performing model, DAFormer with an MiT-B5 backbone, achieved a mean Intersection over Union (mIoU) of 69% when adapting from the synthetic SynthColon dataset to the real Kvasir-SEG dataset. This result not only improves upon the performance reported in the original Synth-Colon paper but also contests the results of their subsequent work, which combines self-training with a refined version of the synthetic dataset we used. This demonstrates that DAFormer has a better generalization ability and shows a clear advantage of Transformer-based architectures over CNN-based approaches for cross-domain medical image segmentation tasks. This paves a clear path for future research through the integration of a superior style-transfer method in our pipeline, which we expect to deliver greater results and show a clearer performance gain over the current methods, such as PL-CUT-Seg.

The method's ability to achieve competitive results while using only synthetic source data addresses critical challenges in medical AI deployment, including patient privacy concerns and the high cost of expert annotations. Future work should focus on improving the realism and diversity of synthetic datasets, exploring more sophisticated domain adaptation techniques tailored for medical imaging, and conducting clinical validation studies to assess the practical utility of such systems in real colonoscopy procedures.

The promising results demonstrated here provide a foundation for developing privacy-preserving, annotation-efficient solutions for computer-aided polyp detection systems.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. The source datasets used for training and validation are publicly available at https://enric1994.github.io/synth-colon/ (Synth-Colon), https://polyp.grand-challenge.org/CVCClinicDB/ (CVC-ClinicDB), and https://datasets.simula.no/kvasir-seg/ (Kvasir-SEG) (accessed on 1 August 2025). Further inquiries can be directed to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

UDA	Unsupervised Domain Adaptation
AI-CAD	Artificial Intelligence-Powered Computer-Aided Detection and Diagnosis
AI	Artificial Intelligence
ViT	Vision Transformer
MiT	Mix Vision Transformer
MHSA	Multi-Head Self-Attention mechanism
PVT	Pyramid Vision Transformer
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
MSE	Mean Squared Error
DACS	Domain Adaptation via Cross-Domain Mixed Sampling
ASPP	Atrous Spatial Pyramid Pooling
EMA	Exponential Moving Average
VAE	Variational Autoencoder

References

1. American Cancer Society. Key Statistics for Colorectal Cancer. 2024. Available online: https://www.cancer.org/cancer/types/colon-rectal-cancer/about/key-statistics.html (accessed on 21 April 2025).
2. Cheng, E.; Blackburn, H.N.; Ng, K.; Spiegelman, D.; Irwin, M.L.; Ma, X.; Gross, C.P.; Tabung, F.K.; Giovannucci, E.L.; Kunz, P.L.; et al. Analysis of Survival Among Adults with Early-Onset Colorectal Cancer in the National Cancer Database. *JAMA Netw. Open* **2021**, *4*, e2112539. [CrossRef]
3. Sikora, N.; Manschke, R.L.; Tang, A.M.; Dunstan, P.; Harris, D.A.; Yang, S. ColonScopeX: Leveraging Explainable Expert Systems with Multimodal Data for Improved Early Diagnosis of Colorectal Cancer. *arXiv* **2025**, arXiv:2504.08824.
4. Rex, D.K.; Boland, C.R.; Dominitz, J.A.; Giardiello, F.M.; Johnson, D.A.; Kaltenbach, T. Colorectal cancer screening: Recommendations for physicians and patients from the US Multi-Society Task Force on colorectal cancer. *Gastroenterology* **2017**, *153*, 307–323. [CrossRef]
5. Than, M.; Witherspoon, J.; Shami, J.; Patil, P.; Saklani, A. Diagnostic miss rate for colorectal cancer: An audit. *Ann. Gastroenterol.* **2015**, *28*, 94–98.
6. Takeda, K.; Kudo, S.E.; Mori, Y.; Misawa, M.; Kudo, T.; Wakamura, K.; Katagiri, A.; Baba, T.; Hidaka, E.; Ishida, F.; et al. Accuracy of diagnosing invasive colorectal cancer using computer-aided endocytoscopy. *Endoscopy* **2017**, *49*, 798–802. [CrossRef]

7. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [CrossRef]
8. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [CrossRef]
9. Schäfer, R.; Nicke, T.; Höfener, H.; Lange, A.; Merhof, D.; Feuerhake, F.; Schulz, V.; Lotz, J.; Kiessling, F. Overcoming Data Scarcity in Biomedical Imaging with a Foundational Multi-Task Model. *Nat Comput Sci* **2024**, *4*, 495—509. .. [CrossRef] [PubMed]
10. Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *npj Digit. Med.* **2020**, *3*, 119. [CrossRef]
11. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *L119*, 1–88. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 30 August 2025).
12. Yao, L.; Prosky, J.; Covington, B.; Lyman, K. A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging. Extended Abstract Track. In Proceedings of the Medical Imaging with Deep Learning (MIDL 2019), London, UK, 8–10 July 2019.
13. Hu, S.; Liao, Z.; Xia, Y. Devil is in Channels: Contrastive Single Domain Generalization for Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, 8–12 October 2023.
14. Ren, G.; Lazarou, M.; Yuan, J.; Stathaki, T. Towards Automated Polyp Segmentation Using Weakly- and Semi-Supervised Learning and Deformable Transformers. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023.
15. Basaran, B.D.; Zhang, W.; Qiao, M.; Kainz, B.; Matthews, P.M.; Bai, W. LesionMix: A Lesion-Level Data Augmentation Method for Medical Image Segmentation. In Proceedings of the Data Augmentation, Labelling, and Imperfections: Third MICCAI Workshop, DALI 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, 12 October 2023.
16. Moreu, E.; McGuinness, K.; O'Connor, N.E. Synthetic data for unsupervised polyp segmentation. In Proceedings of the 29th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2021), Dublin, Ireland, 7–8 December 2021.
17. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017.
18. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-SEG: A Segmented Polyp Dataset. In Proceedings of the 26th International Conference on MultiMedia Modeling (MMM 2020), Daejeon, Korea, 5–8 January 2020.
19. Hoyer, L.; Dai, D.; Gool, L.V. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
21. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New York City, NY, USA, 23–26 June 2021; pp. 45–67.
22. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
23. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Neural Information Processing Systems (NeurIPS), Virtual Conference, 6–14 December 2021.
24. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. PVT v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
25. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
26. Bernal, J.; Sánchez, F.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [CrossRef]

27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 5–9 October 2015.
28. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018.
29. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; de Lange, T.; Halvorsen, P.; Johansen, H.D. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019.
30. Jha, D.; Riegler, M.A.; Johansen, D.; Halvorsen, P.; Johansen, H.D. DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020.
31. 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.
32. Kim, T.; Lee, H.; Kim, D. UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation. In Proceedings of the 29th ACM International Conference on Multimedia, ACM, Virtual Event, 17 October 2021; pp. 2167–2175.
33. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In In Proceedings of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2020), Lima, Peru, 4–8 October 2020.
34. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2021), Strasbourg, France, 27 September–1 October 2021.
35. Dong, B.; Wang, W.; Fan, D.P.; Li, J.; Fu, H.; Shao, L. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *CAAI Artif. Intell. Res.* **2023**, *2*, 9150015. [CrossRef]
36. Rahman, M.M.; Marculescu, R. Medical Image Segmentation via Cascaded Attention Decoding. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 6211–6220. [CrossRef]
37. Shi, W.; Xu, J.; Gao, P. SSformer: A Lightweight Transformer for Semantic Segmentation. In Proceedings of the 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), Shanghai, China, 26–28 September 2022.
38. Fitzgerald, K.; Matuszewski, B. FCB-SwinV2 Transformer for Polyp Segmentation. *arXiv* **2023**, arXiv:2302.01027. [CrossRef]
39. Choudhuri, A.; Gao, Z.; Zheng, M.; Planche, B.; Chen, T.; Wu, Z. PolypSegTrack: Unified Foundation Model for Colonoscopy Video Analysis. *arXiv* **2025**, arXiv:2503.24108.
40. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023.
41. Li, H.; Zhang, D.; Yao, J.; Han, L.; Li, Z.; Han, J. ASPS: Augmented Segment Anything Model for Polyp Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI, Marrakesh, Morocco, 6–10 October 2024.
42. Li, Y.; Hu, M.; Yang, X. Polyp-SAM: Transfer SAM for Polyp Segmentation. In Proceedings of the Medical Imaging 2024: Computer-Aided Diagnosis, San Diego, CA, USA, 3 April 2024.
43. Rahman, M.M.; Munir, M.; Jha, D.; Bagci, U.; Marculescu, R. PP-SAM: Perturbed Prompts for Robust Adaptation of Segment Anything Model for Polyp Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, WA, USA, 17–21 June 2024.
44. Mao, X.; Xing, X.; Meng, F.; Liu, J.; Bai, F.; Nie, Q.; Meng, M. One Polyp Identifies All: One-Shot Polyp Segmentation with SAM via Cascaded Priors and Iterative Prompt Evolution. *arXiv* **2025**, arXiv:2507.16337. [CrossRef]
45. Mansoori, M.; Shahabodini, S.; Abouei, J.; Plataniotis, K.N.; Mohammadi, A. Polyp SAM 2: Advancing Zero shot Polyp Segmentation in Colorectal Cancer Detection. *arXiv* **2024**, arXiv:2408.05892. [CrossRef]
46. Zhao, Y.; Zhou, T.; Gu, Y.; Zhou, Y.; Zhang, Y.; Wu, Y.; Fu, H. WeakPolyp-SAM: Segment Anything Model-driven weakly-supervised polyp segmentation. *Know.-Based Syst.* **2025**, *322*, 113701. [CrossRef]
47. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS 2014), Montreal, QC, Canada, 8–13 December 2014
48. Gong, R.; Li, W.; Chen, Y.; Gool, L.V. DLOW: Domain Flow for Adaptation and Generalization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

49. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial Discriminative Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
50. 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 2018), Salt Lake City, UT, USA, 18–23 June 2018.
51. Tsai, Y.H.; Sohn, K.; Schulter, S.; Chandraker, M. Domain Adaptation for Structured Output via Discriminative Patch Representations. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
52. Lee, D.H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In Proceedings of the ICML 2013 Workshop: Challenges in Representation Learning (WREPL), Atlanta, GA, USA, 16–21 June 2013.
53. Sohn, K.; Berthelot, D.; Li, C.L.; Zhang, Z.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual Event, 6–12 December 2020.
54. Zou, Y.; Yu, Z.; Kumar, B.V.K.V.; Wang, J. Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In Proceedings of the 15th European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018.
55. Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.V.K.V.; Wang, J. Confidence Regularized Self-Training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019.
56. Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; Le, X. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
57. Sakaridis, C.; Dai, D.; Hecker, S.; Gool, L.V. Model Adaptation with Synthetic and Real Data for Semantic Dense Foggy Scene Understanding. In Proceedings of the 15th European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018.
58. Yang, Y.; Soatto, S. FDA: Fourier Domain Adaptation for Semantic Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
59. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. In Proceedings of the NIPS 2014 Deep Learning Workshop, Montreal, QC, Canada, 8–13 December 2014.
60. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
61. Olsson, V.; Tranheden, W.; Pinto, J.; Svensson, L. ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2021), Waikoloa, HI, USA, 5–9 January 2021.
62. Tranheden, W.; Olsson, V.; Pinto, J.; Svensson, L. DACS: Domain Adaptation via Cross-domain Mixed Sampling. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2021), Waikoloa, HI, USA, 5–9 January 2021.
63. Zhou, Q.; Feng, Z.; Gu, Q.; Cheng, G.; Lu, X.; Shi, J.; Ma, L. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *Comput. Vis. Image Underst.* **2022**, *221*, 103448. [CrossRef]
64. Diamantis, D.E.; Gatoula, P.; Koulaouzidis, A.; Iakovidis, D.K. This Intestine Does Not Exist: Multiscale Residual Variational Autoencoder for Realistic Wireless Capsule Endoscopy Image Generation. *IEEE Access* **2024**, *12*, 25668–25683. [CrossRef]
65. Barua, H.B.; Stefanov, K.; Wong, K.; Dhall, A.; Krishnasamy, G. GTA-HDR: A Large-Scale Synthetic Dataset for HDR Image Reconstruction. *arXiv* **2024**, arXiv:2403.17837.
66. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A.M. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3234–3243. [CrossRef]
67. Moreu, E.; Arazo, E.; McGuinness, K.; O'Connor, N.E. Joint one-sided synthetic unpaired image translation and segmentation for colorectal cancer prevention. *Expert Syst.* **2022**, *40*, e13137. [CrossRef]
68. Vezakis, I.A.; Georgas, K.; Fotiadis, D.; Matsopoulos, G.K. EffiSegNet: Gastrointestinal Polyp Segmentation through a Pre-Trained EfficientNet-based Network with a Simplified Decoder. In Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2024), Orlando, FL, USA, 15–19 July 2024.
69. Moreu, E.; Arazo, E.; McGuinness, K.; O'Connor, N.E. Self-Supervised and Semi-Supervised Polyp Segmentation using Synthetic Data. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023.
70. Huang, C.H.; Wu, H.Y.; Lin, Y.L. HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021), Toronto, ON, Canada, 6–11 June 2021.

71. McDuff, D.; Curran, T.; Kadambi, A. Synthetic Data in Healthcare. *arXiv* **2023**, arXiv:2304.03243. [PubMed]
72. Chen, J.; Chun, D.; Patel, M.; Chiang, E.; James, J.; Capobianco, J.; Lipson, J.; Hong, C.; Natarajan, K.; Cole, C.L.; et al. The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 44. [CrossRef] [PubMed]
73. Mathew, S.; Nadeem, S.; Kumari, S.; Kaufman, A. Augmenting Colonoscopy Using Extended and Directional CycleGAN for Lossy Image Translation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4695–4704. [CrossRef]
74. Iacono, P.; Khan, N. Structure Preserving Cycle-GAN for Unsupervised Medical Image Domain Adaptation. *arXiv* **2023**, arXiv:2304.09164. [CrossRef]