applied sciences

MDPI

Article

Polyp Generalization via Diversifying Style at Feature-Level Space

Sahadev Poudel [1] and Sang-Woong Lee [2,*]

[1] Department of IT Convergence Engineering, Gachon University, Seongnam 13120, Republic of Korea; sahadevp093@gmail.com
[2] Department of Software, Gachon University, Seongnam 13557, Republic of Korea
* Correspondence: slee@gachon.ac.kr

Abstract: In polyp segmentation, the latest notable topic revolves around polyp generalization, which aims to develop deep learning-based models capable of learning from single or multiple source domains and applying this knowledge to unseen datasets. A significant challenge in real-world clinical settings is the suboptimal performance of generalized models due to domain shift. Convolutional neural networks (CNNs) are often biased towards low-level features, such as style features, impacting generalization. Despite attempts to mitigate this bias using data augmentation techniques, learning model-agnostic and class-specific feature representations remains complex. Previous methods have employed image-level transformations with styles to supplement training data diversity. However, these approaches face limitations in ensuring style diversity due to restricted style sources, limiting the utilization of the potential style space. To address this, we propose a straightforward yet effective style conversion and generation module integrated into the UNet model. This module transfers diverse yet plausible style features to the original training data at the feature-level space, ensuring that generated styles align closely with the original data. Our method demonstrates superior performance in single-domain generalization tasks across five datasets compared to prior methods.

Keywords: polyp generalization; polyp segmentation; domain generalization; image segmentation

Citation: Poudel, S.; Lee, S.-W. Polyp Generalization via Diversifying Style at Feature-Level Space. *Appl. Sci.* **2024**, *14*, 2780. https://doi.org/10.3390/app14072780

Academic Editor: Stefan Fischer

Received: 26 February 2024
Revised: 22 March 2024
Accepted: 23 March 2024
Published: 26 March 2024

1. Introduction

In the development of effective convolutional neural networks (CNNs) for polyp segmentation, numerous approaches have been proposed and have demonstrated satisfactory performance over time [1–3]. Traditional deep learning models typically assume that training and testing data are identical and independently distributed. However, in real-world scenarios, especially when deploying segmentation models to predict polyps from entirely new centers, it becomes crucial to accurately segment them regardless of variations in styles, features, shapes, or illumination not present in the training data. Unfortunately, a significant challenge arises due to the domain shift problem, where the distribution of data in the testing environment differs from that of the training data, leading to a drop in the performance of trained CNN models [4,5]. Style discrepancy is one of the factors that can impede the generalization ability of deep learning models [6]. The discrepancy in style features between datasets exacerbates the domain shift problem, causing issues in real-time applications such as inaccurate polyp diagnosis and analysis, potentially impacting patient screening and treatment plans. Usually, style discrepancy refers to differences in visual characteristics, such as texture, color, or contrast, between datasets used for training and testing CNN models. These style features encompass various aspects of image appearance that may vary significantly across different sources or environments. For instance, variations in lighting conditions, imaging equipment, or image processing techniques can lead to distinct visual styles in the data, as shown in Figure 1. The presence of style discrepancies

between training and testing datasets can pose a significant challenge to CNN models, as they may struggle to generalize across these divergent visual styles, ultimately leading to performance degradation when deployed in real-world settings.

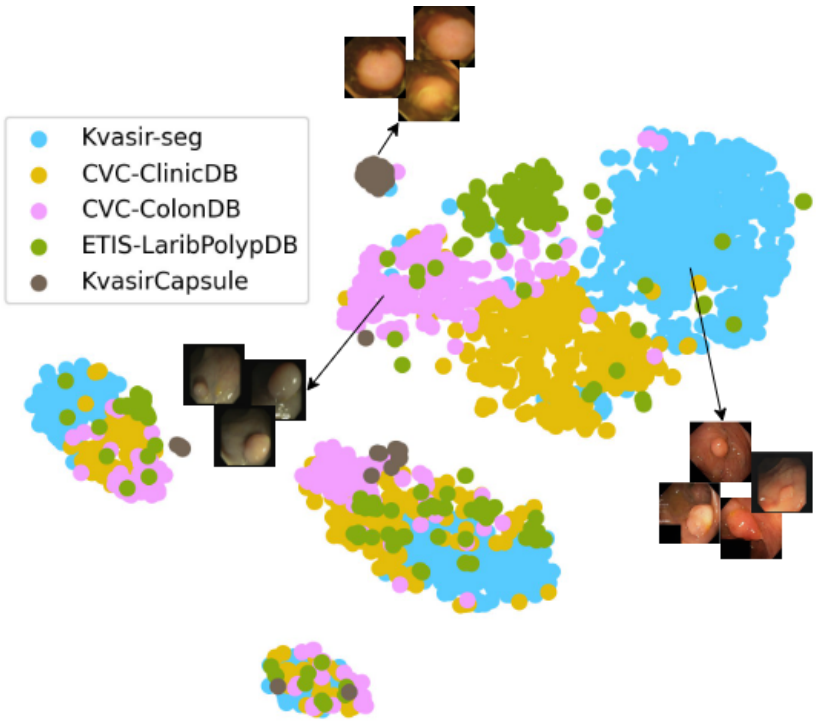

Figure 1. (Best viewed in color). Style feature statistics of five datasets. Note that each feature represents concatenation of mean and standard deviation from earlier layers of EfficientNet for all datasets.

To address the challenge of domain shift, extensive research has been conducted to develop a generalized model capable of performing effectively in novel environments. Domain adaptation is one approach that aims to align the feature distributions between training source data and target data in a domain-invariant setting [7]. However, this method typically requires access to target domain samples during training, which may not always be feasible in medical contexts. Alternatively, domain generalization (DG) involves incorporating multiple domains from various sources into a target domain without directly using target domain data. DG aims to train a model from one or more related domains to enable direct generalization to any unseen target domain without additional adjustments [8]. The primary objective of DG is to enhance the generalization ability of trained models across diverse domains and facilitate adaptation to new scenarios. Nonetheless, a common assumption in DG is that testing data share the same distribution as the training set, which may not always hold true in real-world medical applications.

Among recent advances in domain generalization (DG) models, those based on data manipulation show promising performance [9]. These models utilize data augmentation techniques with a learning-based approach to generate diverse data, complementing the original data and simulating unseen domains to enhance the learning process. Existing DG methods in data augmentation primarily focus on image-level manipulation in the source domain, such as translating images based on styles from auxiliary datasets [10] or converting images from different training domain styles [11]. In computer vision tasks, applying neural style transfer for image-level data augmentation has shown improvement in robustness against domain shift problems [12]. Geirhos et al. [6] highlighted that convolutional neural networks (CNNs) tend to be biased towards textural features rather than class-specific features, such as shape, leading to challenges in adapting to unseen

styles. They proposed stylized versions of datasets using style transfer to enhance accuracy and generalizability.

Similarly, image-level data augmentation using Generative Adversarial Networks (GANs) has been employed, but it cannot be applied universally across tasks [10]. In [13], Adaptive Instance Normalization (AdaIN) was employed to match the mean and variance of the content features with those of the style features. It takes both a content and style image as inputs, encoding them into the feature space at the encoder side. These encoded representations are then passed to an AdaIN layer, which adjusts the mean and variance of the content feature maps to match those of the style feature maps, thereby producing stylized feature maps. The final output is generated by a decoder from these stylized feature maps. In contrast, methods such as MixStyle [14] aim to boost style diversity by augmenting features directly at the feature level. This approach generates new styles by blending existing styles from seen source domains. However, existing style augmentation methods have limitations in fully representing real styles in unseen target domains, leading to reduced diversity in samples and potential performance decrements, particularly when significant differences exist between generated virtual styles and real unseen styles.

In this study, we introduce a novel style-based data augmentation module operating at the feature-space level, tailored for the task of polyp generalization. Our approach aims to address challenges stemming from differences in style distribution between source and target images, which can solve generalization in the polyp domain, as shown in Figure 2. To tackle this, we extract style statistics, such as mean and standard deviation, from the early layers of a convolutional neural network (CNN). The basic assumption behind the style transfer between the polyp is that the network may have confirmation bias towards the style features (color and textural information) while learning. Transferring such features facilitates learning style-agnostic representations, which eventually improves the generalization problem. Our proposed method employs a style-aware encoder–decoder UNet network, integrating style information in the feature space. We utilize Adaptive Instance Normalization (AdaIN) to transfer or generate diverse style features, thereby reducing the style gap between training and unseen testing sources of polyp images while preserving original semantic features for segmentation. Through experiments, we demonstrate the effectiveness of transferring style features from unseen target images to source images during training, enhancing model generalizability. Additionally, increasing style diversity by mixing style features of two images improves model performance. We also introduce a novel approach to generating synthetic yet plausible styles to ensure minimal deviation in generated style features. Our primary objective is to mitigate domain shifts in polyp segmentation tasks while transferring knowledge from one source domain to multiple unseen domains. This scenario can be viewed as a single-domain generalization problem, wherein the segmentation model is trained on a single polyp dataset and applied to multiple unseen datasets. Extensive experiments conducted on five public polyp datasets validate the efficacy of our proposed method. Additionally, we evaluate our method by comparing it with the style augmentation technique conducted at the image level, as demonstrated in the study by Yamashita et al. [15].

The contributions are listed below:

- We introduce a novel style-aware UNet approach for the task of polyp generalization. This method enables the model to learn diverse style features from target style source images, thereby enhancing its generalization ability and effectiveness in unseen target sources.
- We propose a novel style synthesis module (NSSM) aimed at generating diverse yet plausible style features dynamically during training, while also constraining the transfer of unnecessary and highly deviated styles to the source features.
- Our evaluation encompasses five public polyp datasets: Kvasir-SEG [16] (used for training), CVC-Clinic [17], CVC-COLONDB [18], ETIS [19], and KvasirCapsule-SEG [20] (utilized for testing). The experiments conducted demonstrate the effectiveness of our proposed method in the generalization task.

- Finally, we conduct qualitative analysis and quantitative studies to validate the efficacy of our method.

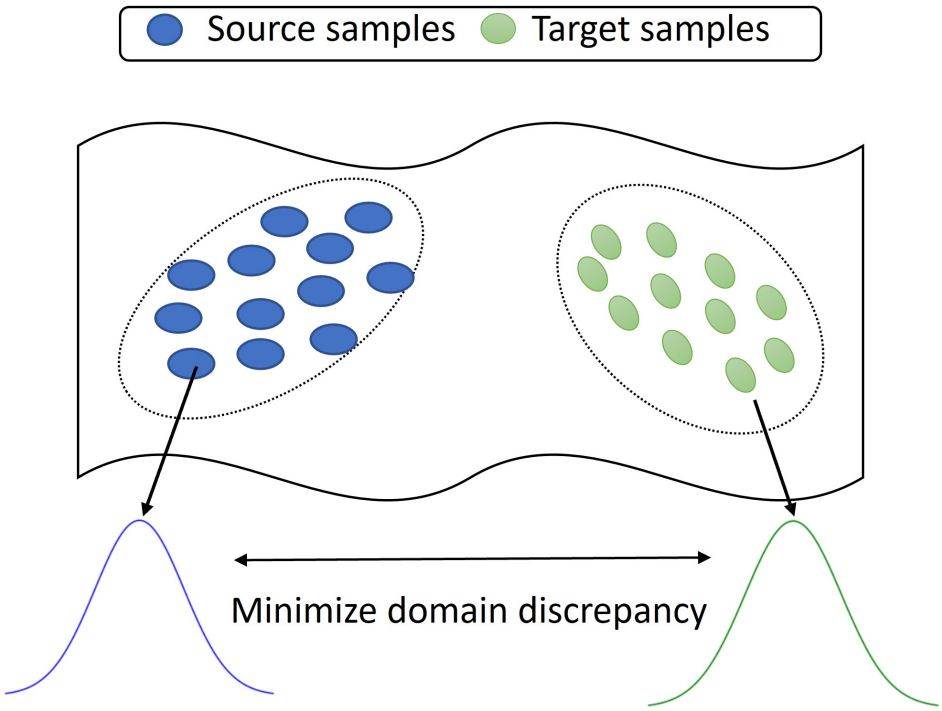

Figure 2. Our proposed method attempts to minimize the domain discrepancy between two polyp domains.

2. Related Work

2.1. Polyp Segmentation

The encoder–decoder-based U-Net architecture [21] is widely recognized for its effectiveness in medical segmentation tasks. Recently, numerous U-Net variants have emerged to improve segmentation performance [22–25]. UNet++ [22] introduced a redesigned skip connection path to minimize the semantic gap between decoder and encoder networks. PraNet [23] proposed a parallel reverse attention network to address the diverse size, color, and texture variations of polyps. Similarly, ACSNet [26] proposed a local and global context attention module to handle polyps of varying sizes. HarDNet-MSEG [24] is a lightweight network incorporating the HarDNet68 [27] module as an encoder and a cascaded partial decoder to enhance accuracy. DCRNet [28] was devised to capture both intra-image and inter-image contextual information. MKDCNet [29] introduced multiple kernel dilated convolution to expand the spatial field of view at deep layers for improved feature representation. Despite their advancements, these methods rely on fully supervised training strategies and have not been thoroughly investigated for their ability to generalize across diverse datasets or adapt to multi-center unseen environments. Thus, while these techniques show promise in specific contexts, their robustness and generalization capabilities to unseen environments require further exploration and validation.

2.2. Style Transfer

Style transfer, a process of translating the style of one image into another without altering its content, has garnered significant attention. Notably, Gatys et al. [30] achieved impressive results by matching neural activation Gram matrices from different convolution layers. Li et al. [31] proposed a novel style loss, aligning the feature statistics (mean and standard deviation) of feature maps between generated and stylized images. Additionally, Huang et al. [13] introduced AdaIN, enabling real-time style transfer by replacing content image feature statistics with those of the style image. In our study, we focus on learning

domain-invariant features using feature-based style randomization for both seen source domains and unseen target domains.

Previous studies in domain adaptation and generalization [32,33] have employed methods to augment styles using Adaptive Instance Normalization (AdaIN). Luo et al. [32] utilized a pre-trained Random Adaptive Instance Normalization module with adversarial style mining to iteratively generate diverse style images. Similarly, Wang et al. [33] replaced the scaling and shifting statistics of AdaIN with learnable parameters to produce novel style images. However, these methods have limitations in terms of the diversity of style features they can generate. AdaIN primarily focuses on adjusting the mean and standard deviation of feature maps to match the style of a reference image, which may result in limited variations in style. Moreover, the reliance on pre-trained or learnable parameters within AdaIN modules might restrict the range of styles that can be effectively synthesized. Thus, while these approaches have shown promise in enhancing style diversity, further advancements are necessary to overcome these inherent limitations and enable the generation of a broader spectrum of style features. In Luo et al. [32], an anchor style is used to guide the generated domain distribution since the target domain is known in advance. Conversely, Wang et al. [33] mixed learnable parameters to generate style statistics, potentially leading to overfitting of the source domain due to the lack of additional information.

3. Methodology

We train the model on a single source domain D_s and generalize it to an unseen multiple domain D_T. All of the datasets may have different data distributions but share same label space. To solve the domain shift problem, we propose the style conversion and generation module, which has three parts: the (1) Style Conversion Module, (2) Style Generation Module, and (3) Novel Style Synthesis Module. We utilize the AdaIN method to deal with style transfer of the target style images during the training. An overview of the proposed method is shown in Figure 3.

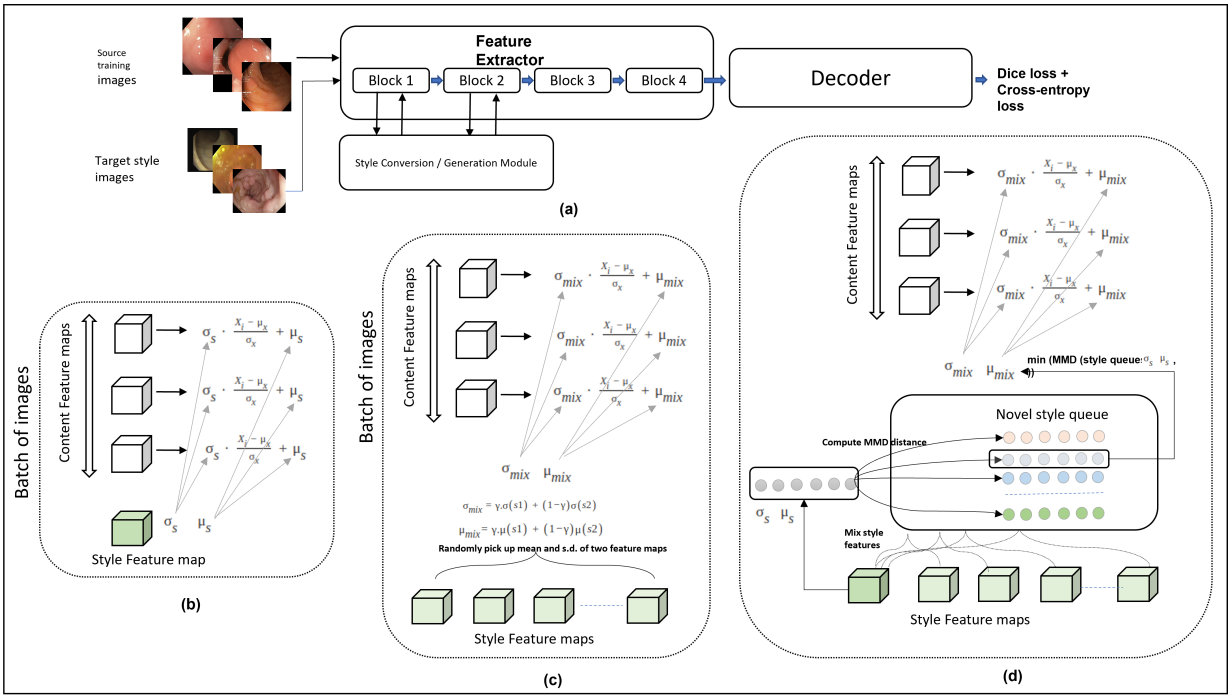

Figure 3. (**a**) Overall framework of the proposed method. We employ EfficientNet, which has 5 blocks depending upon its feature map size, as an encoder backbone. We apply the Style Conversion and Style Generation modules at earlier blocks.; (**b**) Style Conversion Module (SCM); (**c**) Style Generation Module (SGM); and (**d**) Novel Style Synthesis Module (NSSM).

3.1. Background

In style transfer, computing the instance-specific feature statistics, such as mean and standard deviation, during normalization of the feature tensors is a widely accepted technique [34] known as instance normalization (IN) [34]. Let us consider that $x \in R^{B\times C\times H\times W}$ denotes the batch size, number of channels, and height and width of the tensor, respectively. Instance normalization (IN) is formulated as

$$IN(x) = \gamma\frac{x-\mu(x)}{\sigma(x)} + \beta \tag{1}$$

where γ, $\beta \in R^C$ are learnable parameters and $\mu(x)$, $\sigma(x)$ are the mean and standard deviation of each tensor computed across the spatial dimension with each channel. $\mu(x)$, $\sigma(x)$ can be computed as:

$$\mu(x)_{b,c} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{b,c,h,w,} \tag{2}$$

and

$$\sigma(x)_{b,c} = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(x_{b,c,h,w-\mu(x)_{b,c}})^2} \tag{3}$$

Adaptive Instance Normalization (AdaIN) was designed to achieve arbitrary style transfer by changing the scale and shift parameters in Equation (1) with the mean and the standard deviation of the style image (y) as follows:

$$AdaIN(x) = \gamma(y)\frac{x-\mu(x)}{\sigma(x)} + \beta(y) \tag{4}$$

In this manuscript, we will use the aforementioned feature statistics, such as channel-wise mean and standard deviation, to generate the unique style feature in the feature space. We employ AdaIN in order to replace the existing style features with generated unique style features.

3.2. Style Conversion and Generation Module (SCGM)

In this section, we present the architecture of SCGM, as illustrated in Figure 3a. As the style conversion and generation module is executed in the feature space, more diverse transformations of the input images are expected, which ultimately cover more style distribution compared to image-level augmentation.

The general polyp generalization framework consists of a pre-trained encoder, P_f, and a decoder. We employ the Efficient UNet model, which comprises two main components: (1) a UNet encoder leveraging EfficientNet [35] as its backbone, which facilitates the extraction of diverse semantic details across multiple stages; and (2) a decoder module that amalgamates spatial information from various stages to produce a highly accurate segmentation mask. To accomplish the style mixing task, we draw inspiration from Mixstyle [14] and apply it to the preceding two layers. Our main aim is to train the encoder and decoder to focus on source-invariant semantic features across the polyp datasets through style feature conversion and generation.

3.2.1. Style Conversion Module (SCM)

More specifically, the Style Conversion Module (SCM) was inspired by Adaptive Instance Normalization (AdaIN) which replaces the learnable parameters in Equation (1) with the feature styles of target images. It transfers the feature statistics of the target style image to the source training images. It can easily be integrated into the batch while training (as shown in Figure 3a). Given a batch of images $\{x_1, x_2, x_3, x_4....x_s\}$, the SCM first integrates the style image x_s into the source training images. After concatenation on the same mini-

batch training, the SCM computes the feature statistics of each image and transfers all of them from the style to source images.

$$SCN(x) = \gamma(s)\frac{x_i - \mu(x_i)}{\sigma(x_i)} + \beta(s) \tag{5}$$

where $\mu(x_i)$, $\sigma(x_i)$ are computed across the spatial dimension for source images and $\gamma(s)$ and $\beta(s)$ are computed similarly for target images. An overview of the SCM is shown in Figure 3b.

3.2.2. Style Generation Module (SGM)

Our method, the SGM, drew inspiration from MixStyle, which was designed with the aim of regularizing the CNN by mixing the style information of the source domain during the training. However, in our case, we collected the unique style images from different datasets. Given a batch of images $\{x_1, x_2, x_3, x_4....x_{s1}, x_{s2}\}$, we sampled two random images from the collection of style images into the mini-batch settings and performed a novel style generation step. We computed the mixed style features as follows:

$$\gamma_{mix} = \lambda\sigma(x_{s1}) + (1-\lambda)\sigma(x_{s2}) \tag{6}$$

Here, we set λ to 0.5 throughout the experiments. This formulation ensures an equal contribution from each input style, resulting in a balanced proportion in the style mixing process.

$$\beta_{mix} = \lambda\mu(x_{s1}) + (1-\lambda)\mu(x_{s2}) \tag{7}$$

Finally, the mixed style feature space is calculated by the following equation:

$$SGM(x) = \gamma(mix)\frac{x_i - \mu(x_i)}{\sigma(x)} + \beta(mix) \tag{8}$$

An overview of the SGM is shown in Figure 3c.

3.2.3. Novel Style Synthesis Module (NSSM)

An overview of the NSSM is shown in Figure 3d. The problem with the SCM and SGM is that they cannot produce diverse yet plausible style features in the iterations during the training, as they only utilize a single statistic or a mere combination of two style feature statistics from the batch. Therefore, we propose seeking a novel style that should not deviate too much from the source styles and looks realistic. To achieve this, we first make a queue of style features, which is generated by combining the style statistics of two images within the batch size applying the best possible combinations (see Figure 4). We employ a technique similar to the SGM but within each pair of batches. We compute the mean and the standard deviations and mix the feature statistics following Equations (6) and (7). Then, a subset of the images that are distinct in the queue are chosen, and the maximum mean discrepancy (MMD) between the chosen styles and the other remaining target style features is computed.

Given a batch of images $\{x_1, x_2, x_3, x_4....x_{s1}, x_{s2}...x_{sn}\}$ from both the training and target style images, we randomly select one image as a base image and perform style mixing individually with each target style image by applying Equations (6) and (7). Next, we concatenate the mean and the standard deviation values of each by mixing statistics and store them in queue S. We then compare the discrepancy of each computed distribution with the base x_1 feature statistics. Let us assume S1 represents the style feature distribution of x_1 and S2 represents the style queue. We adopt the square maximum mean discrepancy (MMD) between the two distributions (S1 and S2) using a radial basis function (RBF) kernel k as follows:

$$MMD_k^2(S1, S2) = \frac{1}{|S|^2} \sum_{s_i, s_j \in S1} k(s_i, s_j) - \frac{2}{|S1||S2|} \sum_{s_i \in S1, s2_i \in S1} k(S1_i, S2_j) + \frac{1}{|S2|^2} \sum_{P_i, P_j \in S1} k(S2_i, S2_j) \quad (9)$$

Note that we apply Equation (9) for all of the combinations and compare each original style feature with the mixed one, taking those that have the minimum discrepancy.

We then normalize the feature maps following Equation (1) and inject the novel dynamic styles, which were chosen by applying the MMD for the mixed features. The plausible style injection can be formulated by:

$$NSSM(x) = \gamma(Ns)\frac{x - \mu(x)}{\sigma(x)} + \beta(Ns) \quad (10)$$

We adopted a combination of the Dice loss L_{dice} and the cross-entropy loss function to train the network parameters. The Dice loss was proposed by [36] and defined as follows:

$$L_{dice} = 1 - Dice(Y, Y') \quad (11)$$

where Dice is indicated by Dice coefficient score, which represents the spatial overlap regions between the ground truth (Y) and the predicted mask (Y′). It can be calculated as follows:

$$Dice = Mean(\frac{\sum Y' * Y + e}{\sum Y' + \sum Y + e}) \quad (12)$$

In the above Equation (12), $*$ indicates element-wise multiplication and e is a very small parameter in case of unfavorable conditions. The combination of the binary cross-entropy loss and Dice loss have been proven efficient in handling the gradient problem [37]. We can formulate the binary cross-entropy loss as follows:

$$L_{ce} = -\sum(Y * ln(Y') + (1 - Y)ln(1 - Y')) \quad (13)$$

Finally, we combine the two loss functions as follows:

$$TotalLoss = w_1 L_{dice} + w_2 L_{ce} \quad (14)$$

where w_1 and w_2 are the weights for the Dice loss and the binary-cross entropy loss, respectively.

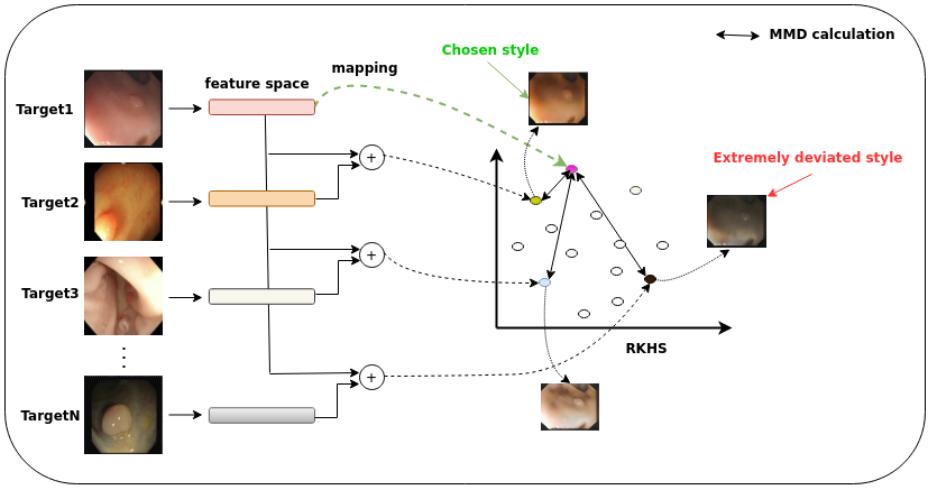

Figure 4. Overall framework of the NSSM. The network processes all target images to extract style features, which are symbolically represented by different colors. These features are then transformed into the RKHS, where a style with minimal deviation is selected.

At last, we apply both the original features and the augmented stylized features to train the network using cross-entropy and the Dice loss function aforementioned in Equation (14). The cumulated loss function is stated as follows:

$$FinalLoss = TotalLoss_{Original} + TotalLoss_{NSSM} \tag{15}$$

Similar settings were applied for the SCM and the SGM.

3.3. Datasets

We conducted experiments using five different datasets to demonstrate the effectiveness of our proposed method for polyp generalization. Our experimental settings closely followed those outlined in PraNet. Unlike some previous studies that utilized Kvasir-SEG and CVC-ClinicDB as training sets and other datasets as testing sets, we exclusively utilized Kvasir-SEG for training. We divided Kvasir-SEG and CVC-ClinicDB into training, validation and testing subsets with ratios of 80%, 10% and 10%, respectively. Additionally, we utilized other test datasets, such as CVC-ColonDB, ETIS and Hyper-Kvasir, which contain 380, 196 and 55 images, respectively, for testing purposes. In our manuscript, we present the testing results obtained from models trained on Kvasir-SEG and evaluated on other datasets. However, we also include results from models trained on CVC-ClinicDB and evaluated on other datasets, including Kvasir-SEG.

For the experiment, we resized the images to 384 $\times$ 384 pixels, consistent with the size used in many prior works. During training, we performed augmentation on the fly while loading the data into the model. This augmentation included rotation, scaling, flipping and shearing.

3.4. Implementation Details

The proposed method was implemented using the PyTorch framework [38] v1.10.2. We utilized a V100 GPU with two 32 GB integrated GPUs for training. As a baseline, we employed the EfficientUNet network, which utilizes EfficientNet as a pre-trained network. The hyperparameters chosen were consistent with those used in prior work [23]. Specifically, we trained the model using stochastic gradient descent (SGD) with a batch size of 16, a momentum of 0.9 and a weight decay of 5×10^{-4}. The total number of epochs was set to 200.

3.5. Evaluation Metrics

We employed various metrics to evaluate and compare our proposed method with state-of-the-art (SOTA) methods. These metrics include mean Intersection over Union (mIoU), Dice coefficient score (Dice), weighted F-measure (FM), structure measure (SM), mean absolute error (MAE) and max enhanced-alignment measure (EM). Among these metrics, Dice and mIoU are similar, as both assess the degree of similarity at the region level and measure consistency within. To compute the Dice, FM and IoU, we utilized 256 pairs of recall and precision values between the predicted mask and the ground truth. Specifically, we transformed the predicted output into a total of 256 binary masks by varying the threshold from 0 to 255.

$$Dice \quad Coefficient = \frac{2 * TP}{2 * TP + FP + FN} \tag{16}$$

$$Jaccard \quad Index = \frac{TP}{TP + FP + FN} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} Recall = \frac{TP}{TP + FN} \tag{18}$$

$$Weighted \quad F\text{-}measure = \frac{Precision * Recall}{Precision + Recall} \tag{19}$$

$$MAE = \frac{1}{N} \sum_{(x,y)} |P(x,y) - G(x,y)| \quad (20)$$

The F-measure is employed to evaluate a segmentation model's performance by considering both precision and recall simultaneously, providing a single metric that reflects the model's overall effectiveness. The structure measure (SM) quantifies the structural consistency between the predicted mask and the ground truth [39]. Additionally, the E-measure (maxE) evaluates the output at both the region and pixel level [40]. Similarly, the MAE serves as a pixel-level similarity comparison metric, computing the average absolute per-pixel difference between the predicted mask and the ground truth. In Equation (20), P(x,y) denotes the pixel value of the ground truth, and G(x,y) represents the pixel value location of the predicted polyp mask.

4. Results

In this section, a comparison of the proposed method with the state-of-the-art methods is presented.

4.1. Comparison with Image-Level Data Augmentation

To compare our proposed approach with the traditional method of style-based data augmentation, we created a dataset of 30,000 images for a generalization task. Here, we transfer the style features from polyp images in other datasets onto the Kvasir-SEG dataset while keeping its original content features intact, inspired by the work of Geirhos et al. [6].

We applied a commonly used style transfer technique, AdaIN [13], following the methodology outlined by Yamashita et al. [15], to generate stylized versions of the Kvasir-SEG dataset as shown in Figure 5. Each dataset in Kvasir-SEG was stylized using style features extracted from randomly selected images from CVC-ClinicDB, CVC-ColonDB, ETIS and KvasirCapsule. Additionally, we manually selected 50 target images from unseen datasets for training purposes. The selection of 50 target images was a deliberate choice aimed at achieving a balance between computational efficiency and maintaining style diversity within the dataset. A higher number of target images would impose a greater computational burden during the style transfer process, potentially hindering the practical feasibility of the approach. Conversely, a lower number of target images might compromise the diversity of styles represented in the dataset, limiting the model's ability to learn and generalize effectively. Therefore, the selection of 50 target images was aimed at striking a balance between these considerations, ensuring both computational feasibility and stylistic diversity. The rationale behind this selection was to transfer unique style characteristics from the target set and optimize the model accordingly. By curating a subset of images showcasing diverse style variations, we aimed to capture a representative sample of the style space present in the non-training set. This selective approach allows us to prioritize the most relevant style information for adaptation while mitigating potential computational burdens associated with incorporating the entire non-training dataset. No additional image transformation techniques, such as rotation, scaling or flipping, were applied. Furthermore, the image size remained consistent at 384 $\times$ 384 pixels across all experimental settings.

We present the quantitative results of the baseline model, traditional augmentation method and style augmentation in Table 1. The baseline Efficient UNet model trained on the stylized dataset achieved superior segmentation results across all evaluation metrics compared to the baseline model. This indicates that the use of style transfer as a data augmentation strategy significantly impacts the model's generalization performance. Furthermore, retraining the model on the original dataset led to marginal improvements in terms of Dice score and mIoU. Both the SCM and SGM achieved similar scores to those of style augmentation, with the distinction being that the SCM and SGM were applied in the feature space while style augmentation was performed at the image level. Lastly, our proposed method (the NSSM) demonstrated better performance compared to other methods, indicating its ability to learn more domain-irrelevant feature representations.

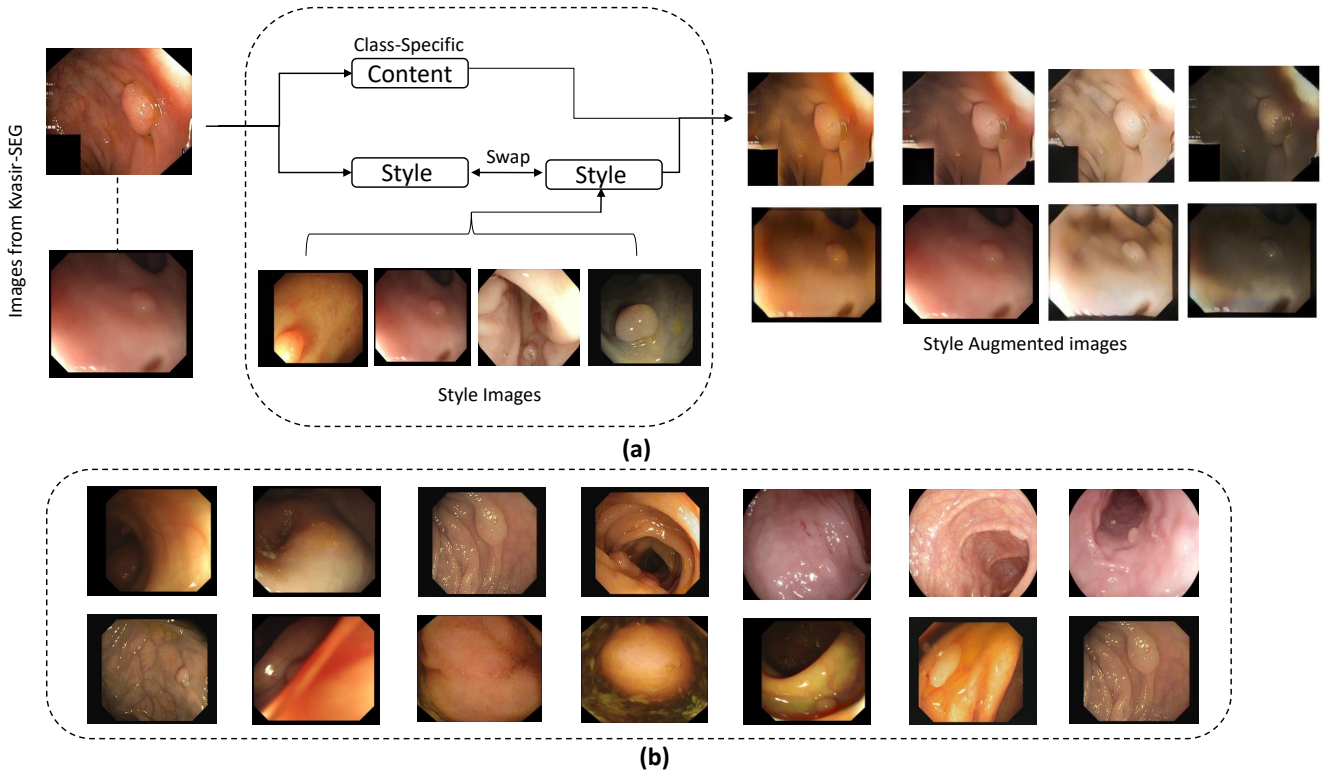

Figure 5. (**a**) Overview of stylized dataset. By applying Adaptive Instance Normalization, we generate 30,000 stylized images using target style images. In this process, we transfer the style characteristics of the target images onto the original images while preserving their original content features. (**b**) A few samples of target style images, which were taken manually from five datasets.

Table 1. Comparision of baseline model, traditional augmentation method (stylized dataset) and the proposed method in different settings on the five datasets.

Method	Kvasir-SEG		CVC-ClinicDB		CVC-ColonDB		ETIS Dataset		KvasirCapsule	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
Efficient UNet	0.892	0.805	0.840	0.73	0.728	0.580	0.643	0.492	0.612	0.436
Efficient UNet + Stylized Dataset	0.902	0.834	0.853	0.749	0.778	0.689	0.746	0.610	0.804	0.719
Efficient UNet + Stylized Dataset + Re-train	0.909	0.846	0.858	0.751	0.785	0.693	0.766	0.636	0.824	0.732
SCM	0.903	0.841	0.852	0.759	0.781	0.697	0.773	0.683	0.836	0.739
SGM	0.897	0.835	0.868	0.776	0.788	0.704	0.781	0.696	0.848	0.759
NSSM	**0.920**	**0.867**	**0.881**	**0.798**	**0.808**	**0.730**	**0.779**	**0.694**	**0.854**	**0.760**

4.2. Experimental Results of Kvasir-SEG

Quantitative results are presented in Table 2, while qualitative results are illustrated in Figures 6 and 7. Observing Table 1, we find that our proposed method achieves comparable performances across all metrics compared to other state-of-the-art (SOTA) methods. Notably, it demonstrates approximately a 3% improvement in the Dice score and IoU compared to PraNet and HardNet-MSEG. Additionally, the proposed method exhibits enhancements in other metrics such as FM, SM and MAE. In contrast, while FRCNet and HardNet-MSEG achieve MAE scores of 0.024 and 0.028, respectively, these values are slightly higher than those achieved by our proposed method. The model's performances on challenging images can be observed in the accompanying figure, showcasing its effectiveness in polyp segmentation and generalization.

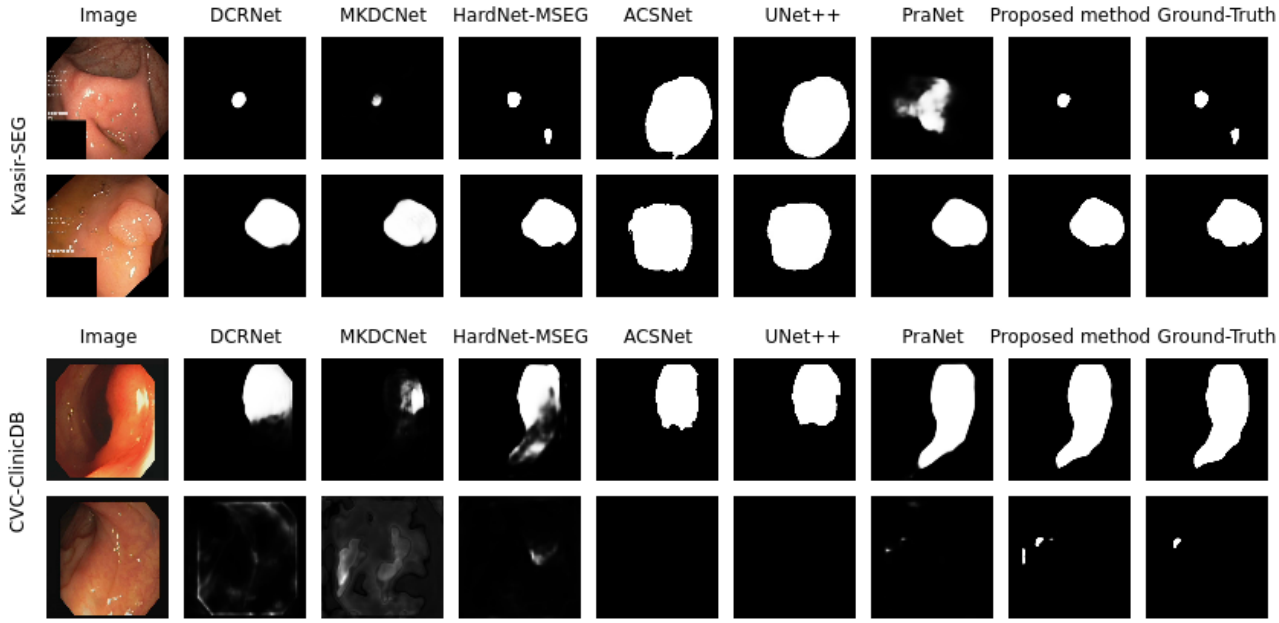

Figure 6. Qualitative comparison of the different methods on the challenging images from Kvasir-SEG testing subset and CVC-CliniCDB when trained on Kvasir-SEG.

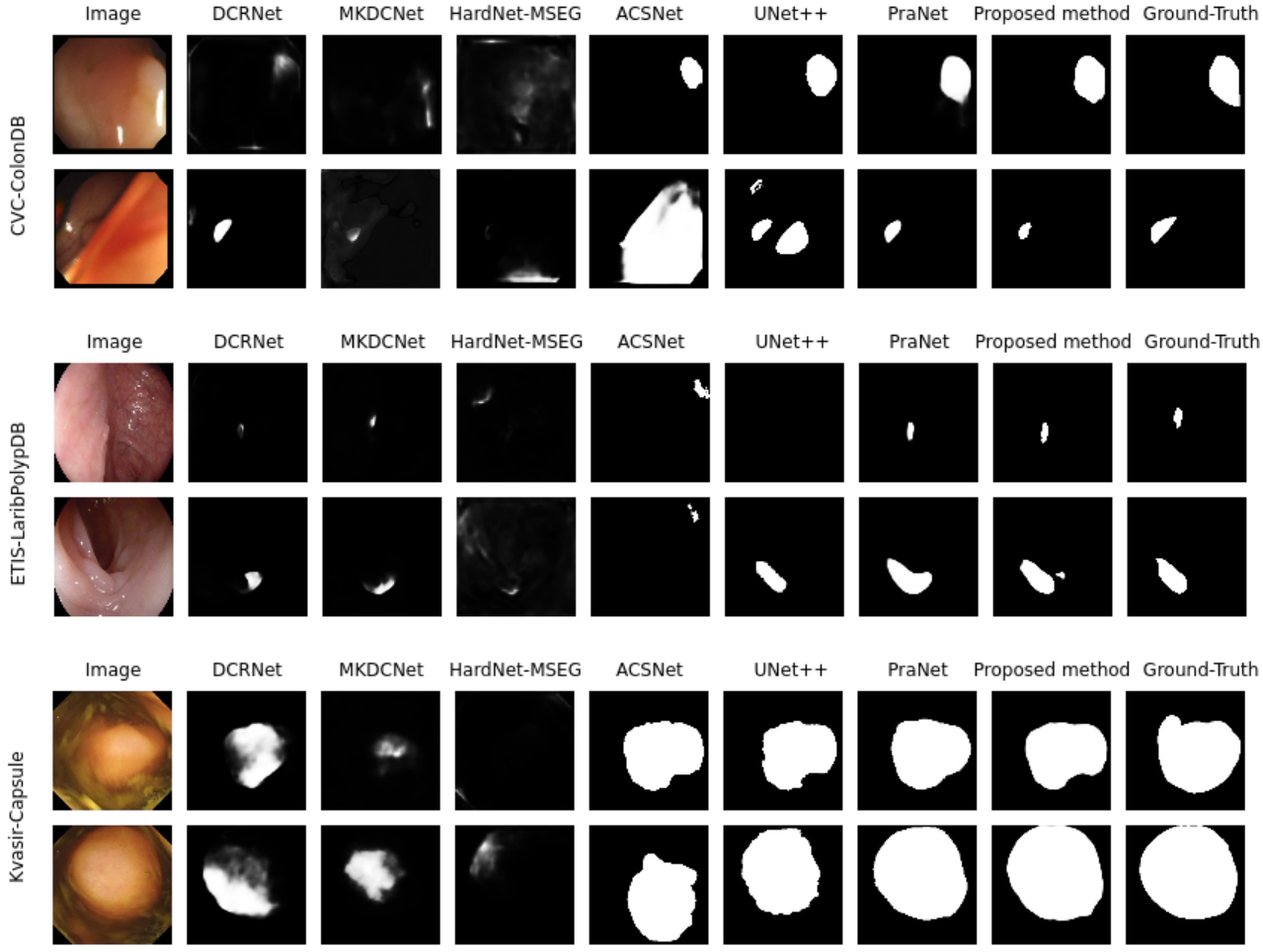

Figure 7. Qualitative comparison of the different methods on the images from challenging CVC-ColonDb (**top**), ETIS-LaribPolypDB (**middle**) and KvasirCapsule-SEG (**bottom**) when trained on Kvasir-SEG.

Table 2. Result comparison of the proposed method on the Kvasir-SEG. Note that mDice, mIoU, FM, SM, maxE and MAE represent mean dice coefficient, mean intersection over union, mean weighted F-measure, structure measure [39], max enhanced-alignment measure and mean absolute error, respectively. Evaluation scores of FRCNet [41], TransFuse [42] and SwinE-Net [43] are referred from their own research whereas evaluation scores for U-Net [21], UNet++ [22], DCRNet [28], ACSNet [26], PraNet [23], HardNet-MSEG [24] and MKDCNet [29] are computed.

Method	U-Net	UNet++	DCRNet	ACSNet	PraNet	HardNet	FRCNet	MKDCNet	TransFuse	SwinE-Net	NSSM
mDice	0.818	0.821	0.886	0.898	0.898	0.897	0.915	0.888	0.918	0.920	**0.920**
mIoU	0.746	0.743	0.825	0.838	0.840	0.839	0.849	0.826	0.868	**0.891**	0.867
FM	0.794	0.808	0.865	0.882	0.885	0.885	0.911	0.75	N/A	**0.913**	0.910
SM	0.858	0.862	0.911	0.92	0.915	0.912	0.919	0.830	N/A	**0.926**	0.924
maxE	0.893	0.91	0.941	0.952	0.948	0.948	0.959	0.903	N/A	N/A	**0.959**
MAE	0.055	0.048	0.035	0.032	0.03	0.028	0.024	0.052	N/A	0.024	**0.022**
Params	31.38	9.16	28.99	29.45	32.50	33.34	**0.78**	19.84	26.30	-	30.60
FPS	41.04	30.67	**35**	34.82	48.25	85.3	-	47.54	98.7	-	-

4.2.1. Generalizability on CVC-ClinicDB

The experiment utilized CVC-ClinicDB as the second dataset for training. Table 3 presents the quantitative results of our proposed method alongside various state-of-the-art (SOTA) methods. Our method achieves a Dice coefficient score of 0.881 and a mean IoU of 0.798, slightly surpassing DCRNet, ACSNet and HardNet-MSEG. Notably, there is a substantial performance gap between U-Net, UNet++ and our proposed method. Additionally, SwinE-Net, FRCNet and TransFuse exhibit commendable scores using all metrics, including FM, SM, maxE and MAE, surpassing our proposed method with marginal improvements. However, it is worth noting that these networks are trained on multi-source datasets. Despite being trained on a single-source dataset, our proposed method achieves comparable performance, approaching the performance levels of FRCNet, TransFuse and SwinE-Net.

Table 3. Result comparison of the proposed method trained on Kvasir-SEG and tested on the CVC-ClinicDB dataset. Evaluation scores of FRCNet [41], TransFuse [42] and SwinE-Net [43] are referred from their own research whereas evaluation scores for U-Net [21], UNet++ [22], DCRNet [28], ACSNet [26], PraNet [23], HardNet-MSEG [24] and MKDCNet [29] are computed.

Method	U-Net	UNet++	DCRNet	ACSNet	PraNet	HardNet	FRCNet	MKDCNet	TransFuse	SwinE-Net	NSSM
mDice	0.633	0.635	0.787	0.837	0.901	0.763	0.933	0.824	0.918	**0.938**	0.881
mIoU	0.543	0.547	0.721	0.826	0.857	0.693	0.886	0.746	0.868	**0.892**	0.798
FM	0.711	0.725	0.774	0.873	0.896	0.935	0.915	0.529	N/A	**0.936**	0.935
SM	0.789	0.793	0.860	0.927	0.935	0.849	0.942	0.715	N/A	**0.950**	0.947
maxE	0.854	0.831	0.896	0.959	0.957	0.878	0.981	0.828	N/A	**0.989**	0.985
MAE	0.039	0.038	0.029	0.011	0.009	0.030	0.007	0.060	N/A	**0.006**	0.010

4.2.2. Generalizability on CVC-ColonDB

For the experimental evaluation of the generalization task, CVC-ColonDB serves as the third dataset, with only Kvasir-SEG utilized for training. The quantitative results presented in Table 4 demonstrate that our proposed method outperforms previous approaches across all evaluation metrics. Notably, the NSSM achieves outstanding scores with a Dice coefficient of 0.808, mean IoU of 0.73 and MAE of 0.026. FRCNet, HardNet and ACSNet also attain competitive scores, with MAE values of 0.036, 0.038 and 0.039, respectively.

Table 4. Comparative result comparison of the proposed method on the CVC-ColonDB dataset. Evaluation scores of FRCNet [41], TransFuse [42] and SwinE-Net [43] are referred from their own research whereas evaluation scores for U-Net [21], UNet++ [22], DCRNet [28], ACSNet [26], PraNet [23], HardNet-MSEG [24] and MKDCNet [29] are computed.

Method	U-Net	UNet++	DCRNet	ACSNet	PraNet	HardNet	FRCNet	MKDCNet	TransFuse	SwinE-Net	NSSM
mDice	0.512	0.483	0.704	0.716	0.712	0.735	0.741	0.367	0.773	0.804	**0.808**
mIoU	0.444	0.41	0.631	0.649	0.64	0.666	0.67	0.296	0.696	0.725	**0.730**
FM	0.498	0.467	0.684	0.697	0.699	0.724	0.728	0.351	N/A	0.787	**0.791**
SM	0.712	0.691	0.821	0.829	0.82	0.834	0.831	0.627	N/A	0.869	**0.871**
maxE	0.776	0.76	0.848	0.851	0.072	0.875	0.878	0.766	N/A	0.910	**0.915**
MAE	0.061	0.064	0.052	0.039	0.043	0.038	0.036	0.103	N/A	0.028	**0.026**

4.2.3. Generalizability on ETIS Dataset

The results of our proposed method, trained on the Kvasir-SEG dataset and tested on the ETIS dataset, are presented in Table 5. Our method achieves a Dice score of 0.779 and a mean IoU of 0.694, along with FM, SM, maxE and MAE values of 0.739, 0.853, 0.904 and 0.011, respectively. Among the models evaluated, SwinE-Net and TransUnet rank second and third with Dice scores of 0.758 and 0.733, respectively. Notably, our proposed method significantly outperforms U-Net and UNet++, demonstrating its robustness and generalizability on unseen datasets.

Table 5. Comparative result comparison of the proposed method on the ETIS dataset. Evaluation scores of FRCNet [41], TransFuse [42] and SwinE-Net [43] are referred from their own research whereas evaluation scores for U-Net [21], UNet++ [22], DCRNet [28], ACSNet [26], PraNet [23], HardNet-MSEG [24] and MKDCNet [29] are computed.

Method	U-Net	UNet++	DCRNet	ACSNet	PraNet	HardNet	FRCNet	MKDCNet	TransFuse	SwinE-Net	NSSM
mDice	0.398	0.401	0.556	0.578	0.628	0.70	0.712	0.432	0.733	0.758	**0.779**
mIoU	0.335	0.344	0.496	0.509	0.567	0.63	0.647	0.371	0.659	0.687	**0.694**
FM	0.366	0.390	0.506	0.560	0.60	0.671	0.682	0.409	N/A	0.726	**0.739**
SM	0.684	0.683	0.736	0.754	0.794	0.828	0.837	0.679	N/A	0.864	**0.853**
maxE	0.74	0.776	0.773	0.764	0.808	0.89	0.89	0.745	N/A	0.902	**0.904**
MAE	0.036	0.035	0.096	0.059	0.031	0.015	0.015	0.061	N/A	0.012	**0.011**

4.2.4. Generalizability on Hyper Kvasir Capsule

Table 6 provides quantitative results, indicating that our proposed method demonstrates superior generalization capabilities compared to other approaches. It achieves an exceptional Dice score of 0.854 and a mean IoU of 0.76. In contrast, U-Net, UNet++, DCRNet, MKDCNet and ACSNet attain Dice scores of 0.384, 0.421, 0.213, 0.269 and 0.578, respectively. The proposed method outperforms these methods not only in mean Dice but also across all major metrics. During the training phase, our method effectively utilizes style statistics from a diverse collection of target images from different datasets. Consequently, despite the substantial domain gap, our method accurately predicts this domain compared to prior methods.

4.3. Experimental Results of CVC-ClinicDB

To assess the efficacy of our proposed method, we trained it on the CVC-ClinicDB dataset and evaluated its performance on the remaining datasets, including its own test set. It is important to note that we utilized the same number of style images as those used for training on Kvasir-SEG. The results are presented in Table 7. Our proposed method demonstrates superior accuracy on unseen datasets compared to prior methods. It is noteworthy that all methods perform well when the training and testing samples are from the same source. However, we observed performance drops, particularly on the CVC-ColonDB, ETIS and KvasirCapsule datasets. For instance, DCRNet, ACSNet, PraNet and

HardNet-MSEG exhibit impressive performance when tested on their own testing sets but struggle on other unseen datasets. In contrast, our proposed method consistently improves performance across all datasets except for Kvasir-SEG. Although PraNet achieves a Dice score of 0.876 and a mean IoU of 0.832, outperforming our proposed method in this specific dataset, this is not the case for other datasets.

Table 6. Comparative result comparison of the proposed method on the Hyper KvasirCapsule dataset. Evaluation scores of U-Net [21], UNet++ [22], DCRNet [28], ACSNet [26], PraNet [23], HardNet-MSEG [24] and MKDCNet [29] are computed.

Method	U-Net	UNet++	DCRNet	MKDCNet	ACSNet	PraNet	HardNet	Proposed Method (NSSM)
mDice	0.384	0.421	0.213	0.269	0.578	**0.937**	0.393	0.854
mIoU	0.343	0.345	0.136	0.142	0.509	**0.890**	0.292	0.760
FM	0.386	0.397	0.215	0.272	0.560	**0.926**	0.395	0.906
SM	0.673	0.694	0.318	0.335	0.754	**0.873**	0.436	0.759
maxE	0.68	0.746	0.323	0.349	0.764	**0.951**	0.426	0.771
MAE	0.034	0.032	0.523	0.510	0.059	**0.074**	0.436	0.164

Table 7. Comparison of the proposed method trained on CVC-ClinicDB and tested on other datasets.

Method	CVC-ClinicDB		Kvasir-SEG		CVC-ColonDB		ETIS		KvasirCapsule	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net	0.874	0.806	0.748	0.682	0.586	0.527	0.557	0.419	0.358	0.301
UNet++	0.886	0.835	0.752	0.689	0.604	0.548	0.603	0.517	0.416	0.349
DCRNet	0.930	0.876	0.793	0.715	0.694	0.625	0.397	0.338	0.701	0.630
MKDCNet	0.887	0.832	0.815	0.772	0.723	0.676	0.613	0.506	0.597	0.516
ACSNet	0.914	0.895	0.845	0.724	0.738	0.684	0.617	0.536	0.646	0.572
PraNet	0.915	0.867	**0.876**	**0.832**	0.716	0.644	0.630	0.575	0.737	0.664
HardNet-MSEG	0.924	0.891	0.840	0.73	0.712	0.634	0.652	0.574	0.541	0.437
Proposed Method (NSSM)	**0.943**	**0.898**	0.842	0.764	**0.756**	**0.692**	**0.681**	**0.606**	**0.763**	**0.676**

5. Discussion

We have introduced augmentation strategies operating in the feature space, which have exhibited superior performance across various polyp datasets and demonstrated the ability to generalize in challenging and unseen environments compared to traditional data augmentation techniques and previous methodologies. We hypothesize that our proposed method facilitates the learning of deep learning models to acquire domain-agnostic and content-specific visual representations by substituting or exchanging original style components with new ones, primarily focusing on domain-irrelevant and class-specific aspects. Indeed, simply transferring style statistics at both the image level and feature space level has led to significant performance improvements. Training on stylized versions of the polyp dataset has resulted in notably enhanced performance compared to traditional augmentation methods. Similarly, experiments involving the transfer of style features at the feature space level have yielded comparable performance gains. Furthermore, the experiments on generating a more diverse style transfer technique (the NSSM) demonstrated that the proposed method can deal with arbitrary styles, in contrast to traditional augmentation approaches, the SGM and AGM, which rely on a fixed set of data style transformations.

Allegedly, prior research has not employed style-based data augmentation, either at the image level or the feature space level, for the task of polyp segmentation and generalization using deep learning models. Some earlier studies have applied style transfer techniques in different domains, such as skin lesion classification [44] and histology datasets [15]. However, these studies utilized transformations relevant to medical contexts to address image scarcity and imbalance issues in datasets. One variant of our proposed method (the SCM) bears some resemblance to this approach. Moreover, the study by [15] emphasized

the use of medically irrelevant transformations with natural images and demonstrated their superiority over previous approaches. This could be attributed to the potential of diverse transformations enabled by employing a wider range of style features, unlike medically relevant transformations, which are inherently limited. Our proposed methods, particularly the SGM and NSSM, align with this concept. We achieved style diversification by leveraging target images solely from different polyp datasets and generating plausible styles similar to the source image. Learning features that are specific to classes and independent of domains is crucial for deep learning models, akin to human cognition. Utilizing style transfer techniques at the feature space level with extensive diversification can enhance the model's representation significantly.

While data augmentation at the image level is recognized as an effective technique for enhancing the performance and generalization of deep learning models, its application and potential in medical imaging remain largely unexplored, thus warranting further investigation. Additionally, determining an optimal configuration for data augmentation methods can vary depending on the datasets being utilized. As suggested by our proposed method, employing style data augmentation at the feature space level holds promise for learning domain-agnostic and class-specific representations. The findings presented in Table 1 underscore the need for future research to explore optimal settings for data augmentation and style transformation in a diverse and plausible manner, akin to the approach proposed in the NSSM, alongside existing methods.

In clinical scenarios, the performance of deep learning models often diminishes due to variations in domain shift. Models that can generalize across multiple datasets are highly advantageous. While it is commonly believed that training deep learning models on diverse multi-institutional datasets can facilitate generalization to unseen datasets, our proposed method demonstrates that a well-curated dataset or thoughtfully designed architecture can compel the model to learn features that are both class-specific and invariant to domain shifts. Specifically, the NSSM achieved comparable performance within its dataset but exhibits superior performance on other unseen datasets. For instance, when trained on Kvasir-SEG and tested on CVC-ClinicDB, the NSSM achieved the highest Dice score of 0.933 and a mean IoU of 0.893. Similar performance gains are observed on the CVC-ColonDB and ETIS datasets. Notably, the proposed method outperforms prior methods significantly, as evidenced in Table 6, where it attains a Dice score of 0.844 and a mean IoU of 0.750, surpassing other major metrics, while prior methods struggle with generalization. Similarly, when trained on CVC-ClinicDB and tested on other datasets, the proposed method achieves the highest scores on all metrics. Through extensive experimentation, the results indicate that the proposed method exhibits superior generalizability, attributable to its style augmentation approach at the feature space level, which consistently generates diverse style features while preserving key content features.

We conducted ablation studies to determine the optimal settings. We experimented with different mixing ratios between the source training images and target style images at the feature space. Our findings indicate that an equal mixing ratio resulted in greater diversity in the features, leading to improved performance on unseen datasets. The quantitative results of these experiments are presented in Table 8.

One significant limitation of our study is that we tested our approach solely on a simple UNet architecture. This decision was made to avoid potential interpretability issues that could arise from using more complex models. However, future research endeavors are necessary to investigate whether our approach can effectively address these limitations and demonstrate its robustness and efficiency in various scenarios. Specifically, it would be valuable to explore the applicability of our method within the frameworks of prior methods' backbones. Additionally, extending the evaluation to other domains such as classification and detection tasks, as well as different medical imaging domains including histopathology, dermatology and radiology, would provide further insights into the versatility of our approach. Another limitation worth mentioning is that our proposed method (the NSSM)

requires longer training times compared to previous works, primarily due to the style augmentation being performed at the feature space level on the fly.

Table 8. Ablation studies of the proposed method in different settings on the five datasets when trained on Kvasir-SEG.

Method	Kvasir-SEG		CVC-ClinicDB		CVC-ColonDB		ETIS		KvasirCapsule	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
Mixing ratios [0.9, 0.1]	0.892	0.805	0.840	0.73	0.768	0.680	0.733	0.662	0.812	0.736
Mixing ratios [0.8, 0.2]	0.902	0.834	0.853	0.749	0.788	0.693	0.746	0.610	0.814	0.738
Mixing ratios [0.7, 0.3]	0.909	0.846	0.858	0.751	0.784	0.691	0.766	0.636	0.849	0.742
Mixing ratios [0.6, 0.4]	0.903	0.841	0.872	0.783	0.791	0.714	0.773	0.683	0.829	0.735
Mixing ratios [0.5, 0.5]	**0.920**	**0.867**	**0.881**	**0.798**	**0.808**	**0.730**	**0.779**	**0.694**	**0.854**	**0.760**

In summary, we have introduced the NSSM, a novel style-based data augmentation method designed to learn diverse style features from a comprehensive set of medically relevant polyp images originating from various sources. Our approach aims to facilitate the acquisition of domain-agnostic and class-specific feature representations within the polyp domain. Through our experiments, we have demonstrated notable enhancements in the performance of segmentation tasks across different unseen datasets, particularly when confronted with domain shift challenges. Our investigation underscores two key findings. Firstly, CNNs exhibit a bias towards style features and may rely on low-level attributes such as color and texture, rendering them susceptible to domain shifts within polyp domains. Secondly, we posit that the incorporation of a medically relevant NSSM can serve as a practical strategy to alleviate this reliance, thereby offering a potential avenue for acquiring domain-agnostic representations.

Author Contributions: Data curation, S.P. and S.-W.L.; Funding acquisition, S.-W.L.; Methodology, S.P.; Supervision, S.-W.L.; Writing—original draft, S.P.; Writing—review and editing, S.-W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Gachon University research fund of 2019 (GCU-2019-0771) and the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. RS-2023-00250978).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request.

Acknowledgments: We would like to acknowledge and thank Astha Adhikari for aiding in data collection for the training of the model.

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

References

1. Patino-Barrientos, S.; Sierra-Sosa, D.; Garcia-Zapirain, B.; Castillo-Olea, C.; Elmaghraby, A. Kudo's classification for colon polyps assessment using a deep learning approach. *Appl. Sci.* **2020**, *10*, 501. [CrossRef]
2. Sornapudi, S.; Meng, F.; Yi, S. Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps. *Appl. Sci.* **2019**, *9*, 2404. [CrossRef]
3. Shin, W.; Lee, M.S.; Han, S.W. COMMA: Propagating complementary multi-level aggregation network for polyp segmentation. *Appl. Sci.* **2022**, *12*, 2114. [CrossRef]
4. Stacke, K.; Eilertsen, G.; Unger, J.; Lundström, C. Measuring domain shift for deep learning in histopathology. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 325–336. [CrossRef] [PubMed]
5. Li, J.; Feng, C.; Lin, X.; Qian, X. Utilizing GCN and Meta-Learning Strategy in Unsupervised Domain Adaptation for Pancreatic Cancer Segmentation. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 79–89. [CrossRef] [PubMed]
6. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv* **2018**, arXiv:1811.12231.

7. Blanchard, G.; Lee, G.; Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; p. 24
8. Zhao, Y.; Zhong, Z.; Yang, F.; Luo, Z.; Lin, Y.; Li, S.; Sebe, N. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6277–6286.
9. Kiyasseh, D.; Tadesse, G.A.; Nhan, L.N.T.; Tan, L.V.; Thwaites, L.; Zhu, T.; Clifton, D. PlethAugment: GAN-based PPG augmentation for medical diagnosis in low-resource settings. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3226–3235. [CrossRef]
10. Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; Gong, B. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 2100–2110.
11. Zhou, K.; Yang, Y.; Hospedales, T.; Xiang, T. Deep domain-adversarial image generation for domain generalisation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13025–13032.
12. Jackson, P.T.; Abarghouei, A.A.; Bonner, S.; Breckon, T.P.; Obara, B. Style augmentation: Data augmentation via style randomization. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 15–20 June 2019; Volume 6, pp. 10–11.
13. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
14. Zhou, K.; Yang, Y.; Qiao, Y.; Xiang, T. Domain generalization with mixstyle. *arXiv* **2021**, arXiv:2104.02008.
15. Yamashita, R.; Long, J.; Banda, S.; Shen, J.; Rubin, D.L. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Trans. Med Imaging* **2021**, *40*, 3945–3954. [CrossRef]
16. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; Lange, T.d.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. In Proceedings of the International Conference on Multimedia Modeling, Daejeon, Republic of Korea, 5–8 January 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 451–462.
17. Bernal, J.; Sánchez, F.J.; 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]
18. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med Imaging* **2015**, *35*, 630–644. [CrossRef] [PubMed]
19. Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* **2014**, *9*, 283–293. [CrossRef] [PubMed]
20. Jha, D.; Tomar, N.K.; Ali, S.; Riegler, M.A.; Johansen, H.D.; Johansen, D.; de Lange, T.; Halvorsen, P. Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy. In Proceedings of the 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Aveiro, Portugal, 7–9 June 2021; pp. 37–43.
21. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
22. Zhou, Z.; Siddiquee, M.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: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
23. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. *arXiv* **2020**, arXiv:2006.11392.
24. 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. *arXiv* **2021**, arXiv:2101.07172.
25. Poudel, S.; Lee, S.W. Deep multi-scale attentional features for medical image segmentation. *Appl. Soft Comput.* **2021**, *109*, 107445. [CrossRef]
26. Zhang, R.; Li, G.; Li, Z.; Cui, S.; Qian, D.; Yu, Y. Adaptive context selection for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 253–262.
27. Chao, P.; Kao, C.Y.; Ruan, Y.S.; Huang, C.H.; Lin, Y.L. Hardnet: A low memory traffic network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 3552–3561.
28. Yin, Z.; Liang, K.; Ma, Z.; Guo, J. Duplex contextual relation network for polyp segmentation. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, 28–31 March 2022; pp. 1–5.
29. Tomar, N.K.; Srivastava, A.; Bagci, U.; Jha, D. Automatic Polyp Segmentation with Multiple Kernel Dilated Convolution Network. In Proceedings of the 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS), Shenzhen, China, 21–22 July 2022; pp. 317–322.
30. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
31. Li, Y.; Wang, N.; Liu, J.; Hou, X. Demystifying neural style transfer. *arXiv* **2017**, arXiv:1701.01036.
32. Luo, Y.; Liu, P.; Guan, T.; Yu, J.; Yang, Y. Adversarial style mining for one-shot unsupervised domain adaptation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 20612–20623.

33. Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; Baktashmotlagh, M. Learning to diversify for single domain generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 834–843.
34. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv* **2016**, arXiv:1607.08022.
35. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
36. Soomro, T.A.; Afifi, A.J.; Gao, J.; Hellwich, O.; Paul, M.; Zheng, L. Strided U-Net model: Retinal vessels segmentation using dice loss. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, ACT, Australia, 10–13 December 2018; pp. 1–8.
37. Mehrtash, A.; Wells, W.M.; Tempany, C.M.; Abolmaesumi, P.; Kapur, T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med Imaging* **2020**, *39*, 3868–3878. [CrossRef]
38. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035. Available online: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (accessed on 22 March 2024).
39. Fan, D.P.; Cheng, M.M.; Liu, Y.; Li, T.; Borji, A. Structure-measure: A new way to evaluate foreground maps. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4548–4557.
40. Fan, D.P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.M.; Borji, A. Enhanced-alignment measure for binary foreground map evaluation. *arXiv* **2018**, arXiv:1805.10421.
41. Shi, L.; Wang, Y.; Li, Z. FRCNet: Feature Refining and Context-Guided Network for Efficient Polyp Segmentation. *Front. Bioeng. Biotechnol.* **2022**, *10*, 799541 [CrossRef]
42. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Pasadena, CA, USA, 22–25 February 2024; Springer: Berlin/Heidelberg, Germany, 2021; pp. 14–24.
43. Park, K.B.; Lee, J.Y. SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer. *J. Comput. Des. Eng.* **2022**, *9*, 616–632. [CrossRef]
44. Mikołajczyk, A.; Grochowski, M. Style transfer-based image synthesis as an efficient regularization technique in deep learning. In Proceedings of the 2019 24th International Conference on Methods and Models in Automation and Robotics (MMAR), Miedzyzdroje, Poland, 26–29 August 2019; pp. 42–47.