Uncertainty-aware Diffusion-based Adversarial Attack for Realistic Colonoscopy Image Synthesis

Minjae Jeong*, Hyuna Cho*, Sungyoon Jung, and Won Hwa Kim

Pohang University of Science and Technology (POSTECH), Pohang, South Korea
{minjaetidtid, hyunacho, syjung, wonhwa}@postech.ac.kr

Abstract. Automated semantic segmentation in colonoscopy is crucial for detecting colon polyps and preventing the development of colorectal cancer. However, the scarcity of annotated data presents a challenge to the segmentation task. Recent studies address this data scarcity issue with data augmentation techniques such as perturbing data with adversarial noises or using a generative model to sample unseen images from a learned data distribution. The perturbation approach controls the level of data ambiguity to expand discriminative regions but the augmented noisy images exhibit a lack of diversity. On the other hand, generative models yield diverse realistic images but they cannot directly control the data ambiguity. Therefore, we propose **D**iffusion-based **A**dversarial attack for **S**emantic segmentation considering **P**ixel-level uncertainty (DASP), which incorporates both the controllability of ambiguity in adversarial attack and the data diversity of generative models. Using a hierarchical mask-to-image generation scheme, our method generates both expansive labels and their corresponding images that exhibit diversity and realism. Also, our method controls the magnitude of adversarial attack per pixel considering its uncertainty such that a network prioritizes learning on challenging pixels. The effectivity of our method is extensively validated on two public polyp segmentation benchmarks with four backbone networks, demonstrating its superiority over eleven baselines.

Keywords: Adversarial Attack · Data Augmentation · Semantic Segmentation.

1 Introduction

Semantic segmentation plays a pivotal role in medical imaging by precisely delineating anatomical structures or pathological regions such as cells [22] and tumors [1, 14]. In a colonoscopy, polyps are often found which are potential precursors to colorectal cancer if left untreated [18, 26]. Therefore, accurate and early detection of polyps is vital to prevent the development of life-threatening cancer. Recent deep neural networks (DNNs) have shown their effectiveness in polyp detection [31] and segmentation [10, 34]; however, training such DNNs requires costly large-scale data with per-pixel labels meticulously annotated by medical professionals.

* These authors contributed equally to this work.

To address data scarcity, image synthesis methods have been developed in two-fold: augmentation and generative DNNs. Augmentation methods transform or perturb existing data in various ways. For example, *cropping-based methods* [7, 12, 38] remove parts of images or feature maps, and *geometric transformations* [28] deform the shape of objects within an image, which may drop early-stage polyps which are tiny and often indistinguishable from the background intestinal wall. *Noise injection* [3,6] such as adversarial attacks preserve the morphological characteristics of objects. The attack provides challenging samples by adding small yet tricky noises to images that maximize a training objective to deceive a DNN, and the network improves in an effort to defend against the attacks [21,35]. While the level of ambiguity in attacked data (i.e., augmented data) can be controlled, their diversity is limited as the perturbed images are semantically similar to the originals. Conversely, *generative models* such as GANs [13], VAEs [20], and diffusion models [11, 15, 16, 23] produce diverse and realistic samples from a learned data distribution. However, unlike the adversarial attack, these generative models cannot sensitively control the ambiguity of the sampled data.

In this regard, we propose a novel image synthesis method for semantic segmentation that inherits advantages of the controllability of adversarial attacks and the diversity of generative models. Our method first generates pixel-level annotations and their corresponding images sequentially, using two separate generative models. As these samples are novel yet not crafted to be challenging, they are perturbed to deceive a network by the adversarial attack. During the attack, our approach adjusts the strength of pixel-wise attacks considering their uncertainty to enable the network to better learn ambiguous regions. Specifically, challenging pixels such as pixels of object edges are strongly perturbed compared to those of inner polyps (i.e., easily predictable pixels) such that a network prioritizes learning the challenging regions by minimizing supervised loss. These attacked samples with small noises are further denoised by a downstream diffusion scheme, thereby improving their realism and belonging to the pixel manifold. Consequently, our method ultimately generates novel and realistic samples that make a network prioritize learning from challenging pixels.

Our **main contributions** are summarized as follows: **1)** Our method handles the data scarcity issue for medical images by generating pairs of expensive pixel-level labels and their corresponding images. **2)** The generated images are challenging and diverse as our method effectively combines the adversarial attack to control data ambiguity and the generative networks to secure data diversity. **3)** The strength of the adversarial attack is adjusted based on the uncertainty of the pixels, allowing a network to better learn ambiguous regions such as object edges. Extensive validations on two public benchmarks with four different backbones demonstrate the effectiveness of our method on polyp segmentation.

2 Preliminary: Diffusion-based Adversarial Attack

Fig. 1a shows the effect of an adversarial attack using Projected Gradient Descent (PGD) [21]. Given a pixel p_x of an image x and a maximum noise strength

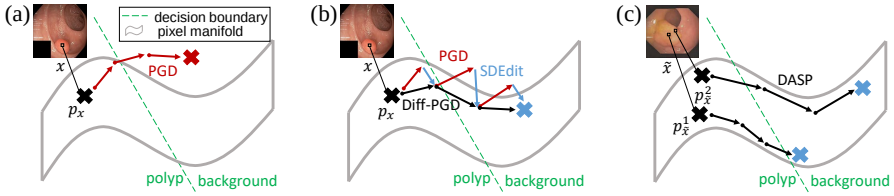

Fig. 1: (a) Adversarial attack using PGD. (b) PGD with SDEdit. (c) PGD with SDEdit considering per-pixel uncertainty.

γ, PGD iteratively adds noise to the data that maximize a training objective L as $(p_x)^k = \prod_{k=1}^{K}((p_x)^{k-1} + \gamma\,\texttt{sign}(\nabla_{(p_x)^{k-1}}L))$ for K perturbation steps, where $(p_x)^0 = p_x$. This iterative attack yields adversarial data lying beyond the network's decision boundary. Also, because of the added noises, the perturbed data contain small artifacts and thus deviate from the natural pixel manifold [37].

To remove the artifacts, authors in [35] proposed Diffusion-based PGD (Diff-PGD) that uses Stochastic Differential Editing (SDEdit). SDEdit diffuses an adversarial pixel $(p_x)^k$ for T^s steps using a forward diffusion and yields $(p_x)^k_{T^s}$. Subsequently, reverse denoising process R_ϕ parameterized by ϕ produces an edited adversarial data $(p_x)^k_0$ lying on the natural pixel manifold as in Fig. 1b as follows

$$(p_x)^k_0 = \text{SDEdit}((p_x)^k, T^s) = R_\phi(...R_\phi(R_\phi((p_x)^k_{T^s}, T^s), T^s - 1)..., 0). \tag{1}$$

Unlike PGD and Diff-PGD which add noises to the given image x, our method first samples an unseen image $\tilde{x}$ from a learned data distribution and perturbs it to secure data diversity. Also, our method allows a segmentation model to intensively learn ambiguous pixels by assigning different perturbation weights to each pixel based on its ambiguity. Fig 1c illustrates that higher weights are assigned to uncertain pixels (e.g., $p^2_{\tilde{x}}$ at object edges) compared to easier cases (e.g., $p^1_{\tilde{x}}$ in inner object regions). As a result, a segmentation network highly prioritizes learning from these ambiguous regions by minimizing a supervised loss such that the ultimate segmentation quality is improved.

3 Method

Our method aims to generate image and mask pairs for medical image segmentation. It follows these steps: **1)** First, we train a Bernoulli diffusion model [5, 33] for mask generation and a mask-conditioned Gaussian diffusion model [8] for image generation. **2)** To yield challenging data that improve the robustness of a segmentation model, diffusion-based adversarial attack is applied on the generated images considering per-pixel uncertainty. **3)** Finally, a segmentation model is trained on both the original and augmented data using a supervised loss.

3.1 Mask Generation with Bernoulli Diffusion Process

Let $\boldsymbol{y} \in \mathbb{R}^{H \times W}$ be a ground truth mask with binary values $\{0, 1\}$, where each pixel is a polyp (1) or background (0). To generate masks, we use a Bernoulli diffusion process [33], which is tailored for binary data generation using discrete Bernoulli noise as a diffusion kernel. For T diffusion timesteps, a forward diffusion process $q^M(\cdot)$ gradually adds Bernoulli noise to $\boldsymbol{y}$ using a noise scale β_t as follows

$$q^M\left(\boldsymbol{y}_{1:T} \mid \boldsymbol{y}_0\right) := \prod_{t=1}^{T} q^M\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}\right) := \prod_{t=1}^{T} \mathcal{B}\left(\boldsymbol{y}_t; (1-\beta_t)\,\boldsymbol{y}_{t-1} + \beta_t/2\right), \tag{2}$$

where $\mathcal{B}$ is a Bernoulli distribution and $\boldsymbol{y}_0 = \boldsymbol{y}$. The $q^M(\cdot)$ yields a complete Bernoulli noise $\boldsymbol{y}_T \sim \mathcal{B}(\boldsymbol{y}_T; \frac{1}{2} \cdot \mathbf{1})$, where $\mathbf{1}$ is an all-one matrix with the same size of $\boldsymbol{y}$. Based on a reparameterization trick [20], $\boldsymbol{y}_t$ is directly sampled from $\boldsymbol{y}_0$ as

$$q^M\left(\boldsymbol{y}_t \mid \boldsymbol{y}_0\right) = \mathcal{B}\left(\boldsymbol{y}_t; \bar{\beta}_t \boldsymbol{y}_0 + (1-\bar{\beta}_t)/2\right), \tag{3}$$

where $\bar{\beta}_t := \prod_{s=1}^{t}(1-\beta_s)$. From Eq. (3), Bernoulli noise ϵ_t applied to $\boldsymbol{y}_t$ is defined as $\epsilon_t = \boldsymbol{y}_t \oplus \boldsymbol{y}_0 \sim \mathcal{B}(\epsilon_t; \frac{1-\bar{\beta}_t}{2} \cdot \mathbf{1})$, where $\oplus$ is an 'exclusive or' operator.

Given $\boldsymbol{y}_T$, a reverse diffusion process $p_\theta^M(\cdot)$ is defined as a Markov process as

$$p_\theta^M\left(\boldsymbol{y}_{0:T}\right) := p\left(\boldsymbol{y}_T\right) \prod_{t=1}^{T} p_\theta^M\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t\right) := p\left(\boldsymbol{y}_T\right) \prod_{t=1}^{T} \mathcal{B}(\boldsymbol{y}_{t-1}; \mu_\theta(\boldsymbol{y}_t, t)), \tag{4}$$

where the $\mu_\theta(\boldsymbol{y}_t, t)$ is determined by a neural network $\epsilon_\theta(\boldsymbol{y}_t, t)$ [33]. The $\epsilon_\theta(\boldsymbol{y}_t, t)$ is trained to estimate the noise ϵ_t using a binary cross entropy loss ℓ_{bce} as

$$\mathcal{L}^M := \mathbb{E}_{t, \boldsymbol{y}_0, \boldsymbol{y}_t \sim q^M(\boldsymbol{y}_t \mid \boldsymbol{y}_0)}\left[\ell_{\mathrm{bce}}\left(\boldsymbol{y}_t \oplus \boldsymbol{y}_0, \epsilon_\theta\left(\boldsymbol{y}_t, t\right)\right)\right]. \tag{5}$$

In the sampling process, unseen masks $\tilde{\boldsymbol{y}}$ are sampled from the learned distribution in Eq. (4). The generated mask $\tilde{\boldsymbol{y}}$ is further used as a condition to sample unseen polyp images in the following conditional generative model.

3.2 Image Generation with Mask-conditioned Gaussian Diffusion

To generate polyp images, a conditional diffusion process is performed on the given image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$. A forward diffusion $q^I(\cdot)$ is similar to Eq. (2) but uses Gaussian noises instead of Bernoulli noise for T' diffusion steps as

$$q^I\left(\boldsymbol{x}_{1:T'} \mid \boldsymbol{x}_0\right) := \prod_{t=1}^{T'} q^I\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) := \prod_{t=1}^{T'} \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t'}\boldsymbol{x}_{t-1}, \beta_t' \mathbf{I}\right), \tag{6}$$

where $\boldsymbol{x}_0 = \boldsymbol{x}$ and β_t' is the variance of Gaussian noise distribution $\mathcal{N}(\cdot)$.

During training, the reverse process $p_\phi^I(\cdot)$ is conditioned on the original ground truth mask $\boldsymbol{y}$ to generate paired polyp images as follows:

$$p_\phi^I(\boldsymbol{x}_{0:T'} | \boldsymbol{y}) := p(\boldsymbol{x}_{T'}) \prod_{t=1}^{T'} p_\phi^I(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{y}) := p(\boldsymbol{x}_{T'}) \prod_{t=1}^{T'} \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\phi(\boldsymbol{x}_t, \boldsymbol{y}, t), \sigma_t^2 \mathbf{I}), \tag{7}$$

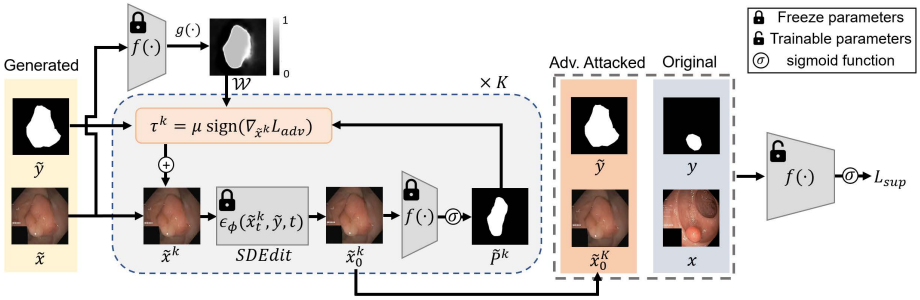

Fig. 2: Overview of DASP. Given a sampled data pair $\{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}\}$, PGD attack with SDEdit is iteratively performed on the image for K steps. The adversarial attack loss L_{adv} is weighted by the uncertainty map $\mathcal{W}$ and thus pixel-wise attack magnitude is controlled to derive the adversarial noise τ^k. A segmentation network $f(\cdot)$ is trained on both the perturbed and the given data to enhance robustness in classifying ambiguous regions.

where $\mu_\phi(\boldsymbol{x}_t, \boldsymbol{y}, t)$ is determined by a neural network $\epsilon_\phi(\boldsymbol{x}_t, \boldsymbol{y}, t)$ [15] and σ_t is determined by β'_t. As in Eq. (5), the $\epsilon_\phi(\boldsymbol{x}_t, \boldsymbol{y}, t)$ is trained to estimate Gaussian noise ϵ'_t applied to the $\boldsymbol{x}_t$ using a mean squared error loss as follows

$$\mathcal{L}^I := \mathbb{E}_{t, \boldsymbol{x}_0, \epsilon'_t \sim \mathcal{N}(0, \mathbf{I})} \left[\left\| \epsilon'_t - \epsilon_\phi \left(\boldsymbol{x}_t, \boldsymbol{y}, t \right) \right\|^2 \right]. \tag{8}$$

In the sampling process, we use the generated mask $\tilde{\boldsymbol{y}}$ as a condition such that a realistic and novel image $\tilde{\boldsymbol{x}}$ paired with $\tilde{\boldsymbol{y}}$ is generated by Eq. (7).

3.3 Diffusion-based Attack considering Pixel-wise Uncertainty

As shown in Fig. 2, our method performs an iterative adversarial attack using PGD [21] on the generated image $\tilde{\boldsymbol{x}}^0 = \tilde{\boldsymbol{x}}$ for K perturbation steps. For $k = 0, ..., K-1$ steps, the image at $(k+1)$-th perturbation $\tilde{\boldsymbol{x}}^{k+1}$ is defined as

$$\tilde{\boldsymbol{x}}^{k+1} = \tilde{\boldsymbol{x}}^k + \tau^k = \tilde{\boldsymbol{x}}^k + \gamma \mathtt{sign}(\nabla_{\tilde{\boldsymbol{x}}^k} \mathcal{L}_{adv}), \tag{9}$$

where $\tau^k \in \mathbb{R}^{H \times W \times 3}$ is an adversarial noise whose element is smaller than the threshold γ (i.e., $|\tau_i^k| \leq \gamma$, s.t. $i = 1, ..., H \times W \times 3$). By adding a noise τ^k to the image, the perturbed data $\tilde{\boldsymbol{x}}^k$ is pushed beyond the network's decision boundary. Note that, as the noise does not necessarily push the image onto the natural pixel manifold, the $\tilde{\boldsymbol{x}}^k$ contains noisy artifacts that make the image unnatural.

To obtain a seamless image, the $\tilde{\boldsymbol{x}}^k$ is further refined to remove artifacts while still remaining beyond the decision boundary. This is realized by applying the reverse denoising process R_ϕ of SDEdit [35] as in Eq. (1). With a pretrained conditional diffusion model $\epsilon_\phi(\tilde{\boldsymbol{x}}_t^k, \tilde{\boldsymbol{y}}, t)$ used for the image generation (in Section 3.2), SDEdit denoises $\tilde{\boldsymbol{x}}_{T^s}^k$ for T^s diffusion steps such that the resultant denoised image $\tilde{\boldsymbol{x}}_0^k \sim p(\tilde{\boldsymbol{x}})$ is yielded within the realistic data distribution $p(\tilde{\boldsymbol{x}})$.

To derive the noise $\tau^k = \mathrm{argmax}_\tau L_{adv}$ that deceives a segmentation network $f(\cdot)$, the adversarial loss L_{adv} is defined to reduce the difference between the

network prediction $\tilde{P}^k = \text{sigmoid}(f(\tilde{\boldsymbol{x}}_0^k))$ and a mask $\tilde{\boldsymbol{y}}$ as follows

$$\mathcal{L}_{adv}(\tilde{P}^k; \tilde{\boldsymbol{y}}, \mathcal{W}) = \frac{1}{H \times W} \sum_{i=1}^{H \times W} \left[\tilde{\boldsymbol{y}}_i \times \mathcal{W}_i \times \ell_{\text{bce}} \left((\tilde{P}^k)_i, \tilde{\boldsymbol{y}}_i \right) \right], \tag{10}$$

where $\mathcal{W}$ is an uncertainty map that contains pixel-wise weights emphasizing per-pixel uncertainty. As we aim to train the $f(\cdot)$ that robustly discriminates uncertain regions (e.g., object edges), the uncertainty map is calculated as $\mathcal{W} = g(f(\tilde{\boldsymbol{x}}))$ with a zero-mean Gaussian function $g(\cdot)$. With the $g(\cdot)$, ambiguous pixels where the network output (before sigmoid) is close to 0 are significantly attacked via maximizing the L_{adv}. Fig. 2 shows that $\mathcal{W}$ assigns high weights to uncertain edge pixels, while weights are relatively low for the inner object regions.

3.4 Training a Segmentation Network

Finally, N number of the original data pairs $\{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^N$ and N' augmented data pairs $\{\tilde{\boldsymbol{x}}_{0,n'}^K, \tilde{\boldsymbol{y}}_{n'}\}_{n'=1}^{N'}$ are combined into a unified training set. Given a Dice loss [30] $l(\cdot)$ and a sigmoid function $\sigma(\cdot)$, the network $f(\cdot)$ is trained on this dataset by minimizing a supervised loss L_{sup} which is defined as follows

$$L_{sup} = \frac{1}{N} \sum_{n=1}^{N} l(\sigma(f(\boldsymbol{x}_n)), \boldsymbol{y}_n) + \frac{1}{N'} \sum_{n'=1}^{N'} l(\sigma(f(\tilde{\boldsymbol{x}}_{0,n'}^K)), \tilde{\boldsymbol{y}}_{n'}). \tag{11}$$

4 Experiment

4.1 Experimental Setup

Dataset. We conducted experiments on two public polyp segmentation datasets: Kvasir-SEG [19] and ETIS-Larib Polyp DB (ETIS) [27]. These datasets contain 1000/196 images with corresponding pixel-wise labels. As in [6,10,29,32], we split train/validation/test sets into 80%/10%/10%. For all augmentation methods, we augmented data by doubling the number of training data (i.e., $N' = 2N$).
Implementation. DASP was trained using Adam optimizer with learning rates of 4e-3/1e-4 (Kavsir-SEG/ETIS) for 200 epochs and a batch size of 16. To generate $\{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}\}$, we followed settings in [8,33]. PGD steps K, noise strength γ, SDEdit timesteps T^s, and the variance of $g(\cdot)$ were set to 10, 1, 3, and 4, respectively. To assess the generalizability of DASP, we used 4 backbone networks: U-Net [24], U-Net++ [40], LinkNet [2], and DeepLabv3+ [4]. Codes will be released online.
Baselines & Evaluation. Along with a typical augmentation method (e.g., random horizontal and vertical flipping denoted as 'Basic' in Table 1), recent methods such as CutMix [38], CutOut [7], Elastic Transform [28], Random Erase [39], DropBlock [12], Gaussian Noise Training (GNT) [25], Logit Uncertainty (LU) [17], Tumor Copy-Paste (TumorCP) [36] and Anti-Adversarial Consistency regularization (AAC) [6] were used as baselines. We also compared DASP with a diffusion-based augmentation method, ArSDM [9]. As evaluation metrics, mean Intersection over Union (mIoU) and mean Dice coefficient (mDice) were used.

Table 1: Segmentation performance of DASP and baseline methods. The best result is marked in bold and the second-best result is indicated by an underline.

Method	U-Net	U-Net++	LinkNet	DeepLabv3+	U-Net	U-Net++	LinkNet	DeepLabv3+
	mIoU				mDice			
Kvasir-SEG								
No Aug.	81.76	63.13	73.92	85.75	86.29	74.75	85.00	89.65
Basic	87.95	88.59	88.97	89.92	93.59	93.95	94.16	94.69
CutMix [38]	86.73	89.35	88.31	91.84	92.90	94.37	93.79	95.75
CutOut [7]	89.05	89.33	89.11	91.64	94.21	94.36	94.24	95.64
Elastic Trans. [28]	87.09	88.64	88.06	91.91	93.10	93.98	93.65	95.79
Random Erase [39]	90.28	90.92	89.92	91.57	94.89	94.08	94.69	95.60
DropBlock [12]	91.15	90.57	91.08	91.36	95.37	95.10	95.48	95.48
GNT [25]	88.04	87.09	89.80	89.32	93.64	93.10	94.63	94.36
LU [17]	86.16	89.47	89.14	91.20	92.57	94.44	94.26	95.40
TumorCP [36]	90.99	91.40	89.14	91.24	95.24	95.51	94.26	95.42
AAC [6]	91.57	89.21	89.87	90.83	95.60	94.30	94.67	95.20
ArSDM [9]	92.22	92.10	91.96	92.02	95.59	95.89	95.81	95.85
DASP (Ours)	**93.10** (+0.88)	**93.06** (+0.96)	**93.05** (+1.09)	**92.98** (+0.96)	**96.43** (+0.83)	**96.41** (+0.52)	**96.40** (+0.59)	**96.36** (+0.51)
ETIS								
No Aug.	83.80	83.96	82.18	82.52	91.18	91.28	90.22	90.43
Basic	87.63	87.21	85.70	86.41	93.41	93.17	92.30	92.71
CutMix [38]	87.67	88.65	85.69	86.06	93.54	93.98	92.30	92.51
CutOut [7]	87.67	88.08	85.69	86.61	93.43	93.66	92.29	92.83
Elastic Trans. [28]	87.27	86.79	85.70	86.05	93.20	92.93	91.30	92.50
Random Erase [39]	85.69	86.55	85.55	85.64	92.30	92.79	92.21	92.27
DropBlock [12]	85.70	85.69	85.69	85.76	92.30	92.30	92.30	92.33
GNT [25]	85.69	88.10	85.62	88.41	92.30	93.67	92.25	93.85
LU [17]	85.69	85.69	85.69	86.88	92.30	92.30	92.30	92.98
TumorCP [36]	86.41	85.95	85.48	86.23	92.71	92.45	92.17	92.60
AAC [6]	89.05	88.92	85.69	86.28	94.21	94.13	92.30	92.64
ArSDM [9]	87.69	87.49	85.83	86.38	93.44	93.33	92.38	92.69
DASP (Ours)	**90.17** (+1.12)	**89.54** (+0.62)	**86.08** (+0.25)	**89.28** (+0.87)	**94.83** (+0.62)	**94.48** (+0.35)	**92.52** (+0.14)	**94.34** (+0.49)

Table 2: Ablation study on the mask type, attack method, and uncertainty map $\mathcal{W}$. The results were obtained from the Kvasir-SEG experiment.

Method			U-Net		U-Net++	
Mask type	Attack method	$\mathcal{W}$	mIoU	mDice	mIoU	mDice
$\boldsymbol{y}$	×	×	92.22	95.59	92.10	95.89
$\tilde{\boldsymbol{y}}$	×	×	92.66	96.19	92.39	96.05
$\tilde{\boldsymbol{y}}$	PGD	✓	92.88	96.31	92.79	96.26
$\tilde{\boldsymbol{y}}$	Diff-PGD	×	92.68	96.20	92.81	96.27
$\tilde{\boldsymbol{y}}$	Diff-PGD	✓	**93.10**	**96.43**	**93.06**	**96.41**

4.2 Quantitative Results

Comparison with baselines. As shown in Table 1, DASP surpasses all baseline approaches in all configurations with a maximum improvement of 1.09%p and 1.12%p in mIoU over the second-best results on the Kvasir-SEG and ETIS datasets, respectively. Notably, our method augments data from the real-world pixel manifold that resemble the given data, while geometric transformations [28, 36], cropping-based methods [7, 12, 36, 38, 39], and noise injection methods [6, 25] yield data out of the natural pixel manifold. Also, by generating both masks and images, our method synthesizes more diverse data compared to a diffusion-based method [9] which only generates images for the original masks.
Ablation study. Ablation study results on the mask generation, attack method, and uncertainty map $\mathcal{W}$ are reported in Table 2. Using a generated mask $\tilde{\boldsymbol{y}}$ improved mIoU over using a given mask $\boldsymbol{y}$ by 0.44%p and 0.29%p for U-Net and

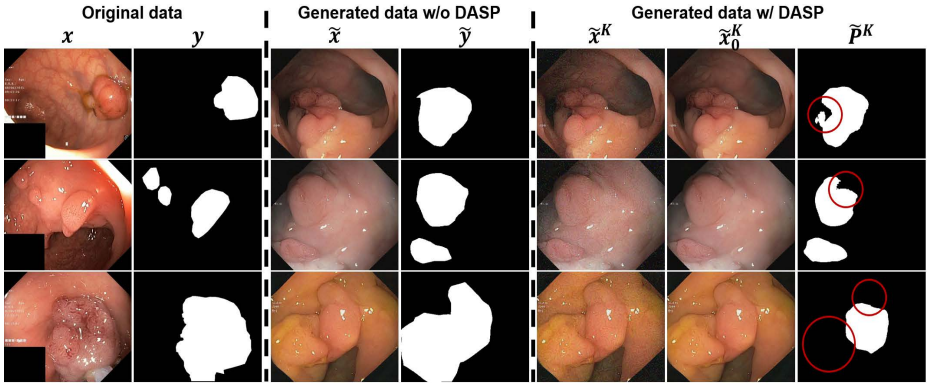

Fig. 3: Visualization of generated samples from DASP on the Kvasir-SEG.

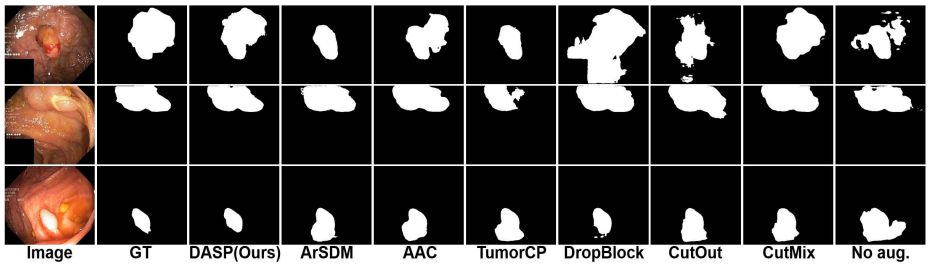

Fig. 4: Segmentation results of different models on the Kvasir-SEG test dataset.

U-Net++ backbones, respectively. For the adversarial attack, both PGD and Diff-PGD outperformed the experiments without attacks; however, Diff-PGD (in the 5th row) showed better results than PGD (in the 3rd row) as Diff-PGD allows a segmentation model to learn augmented data resembling real-world samples. Additionally, using the pixel-wise uncertainty map $\mathcal{W}$ (in the 5th row) consistently outperformed experiments without $\mathcal{W}$ (in the 4th row), indicating the effectivity of adjusting the attack magnitude based on per-pixel uncertainty.

4.3 Qualitative Results

The generated data $\{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}\}$ in Fig. 3 show the effectivity of our hierarchical mask-to-image generation scheme in producing diverse and realistic data. While PGD-attacked data $\tilde{\boldsymbol{x}}^K$ exhibit noisy artifacts, our method with SDEdit $\tilde{\boldsymbol{x}}_0^K$ generates seamless images. The difference between $\tilde{\boldsymbol{x}}^K$ and $\tilde{\boldsymbol{x}}_0^K$ is easily observed in high-resolution, as reported in the supplementary material. Given the input $\tilde{\boldsymbol{x}}_0^K$, the network struggles to accurately predict ambiguous edge pixels, as indicated by the red circles in network prediction $\tilde{P}^k$. By minimizing L_{sup}, the network focuses its learning on accurately classifying these challenging regions, leading to superior performance over various baselines as demonstrated in Fig. 4.

5 Conclusion

We propose a novel image augmentation method for semantic segmentation in medical images by perturbing data with diffusion-based adversarial attack considering per-pixel uncertainty. By controlling the strength of the adversarial attack based on per-pixel uncertainty, our method enables a segmentation network to focus on discriminating ambiguous pixels such as edge regions. Furthermore, our mask-to-image generative scheme generates both expansive pixel-wise annotations and corresponding images. Extensive experiments across multiple datasets and backbone networks validate the effectiveness of our method.

Acknowledgments. This research was supported by NRF-2022R1A2C2092336 (60%), RS-2022-II2202290 (30%), RS-2019-II191906 (AI Graduate Program at POSTECH, 10%).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Biratu, E.S., et al.: A survey of brain tumor segmentation and classification algorithms. Journal of Imaging **7**(9), 179 (2021)
2. Chaurasia, A., et al.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: IEEE Visual Communications and Image Processing. pp. 1–4. IEEE (2017)
3. Chen, C., et al.: Realistic adversarial data augmentation for mR image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 667–677. Springer (2020)
4. Chen, L.C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision. pp. 801–818 (2018)
5. Chen, T., et al.: Berdiff: Conditional bernoulli diffusion model for medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 491–501. Springer (2023)
6. Cho, H., et al.: Anti-adversarial consistency regularization for data augmentation: Applications to robust medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 555–566. Springer (2023)
7. DeVries, T., et al.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
8. Dhariwal, P., et al.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
9. Du, Y., et al.: ArSDM: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 339–349. Springer (2023)
10. Fan, D.P., et al.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 263–273. Springer (2020)

11. Frisch, Y., et al.: Synthesising rare cataract surgery samples with guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 354–364. Springer (2023)
12. Ghiasi, G., et al.: Dropblock: A regularization method for convolutional networks. Advances in Neural Information Processing Systems **31** (2018)
13. Goodfellow, I., et al.: Generative adversarial nets. Advances in Neural Information Processing Systems **27** (2014)
14. Havaei, M., et al.: Brain tumor segmentation with deep neural networks. Medical Image Analysis **35**, 18–31 (2017)
15. Ho, J., et al.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
16. Hu, X., et al.: Conditional diffusion models for weakly supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 756–765. Springer (2023)
17. Hu, Y., et al.: Data augmentation in logit space for medical image classification with limited training data. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 469–479. Springer (2021)
18. Huck, M.B., et al.: Colonic polyps: diagnosis and surveillance. Clinics in Colon and Rectal Surgery **29**(04), 296–305 (2016)
19. Jha, D., et al.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling. pp. 451–462. Springer (2020)
20. Kingma, D.P., et al.: Auto-encoding variational bayes. International Conference on Learning Representations (2013)
21. Madry, A., et al.: Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (2018)
22. Meijering, E.: Cell segmentation: 50 years down the road [life sciences]. IEEE Signal Processing Magazine **29**(5), 140–145 (2012)
23. Peng, W., et al.: Generating realistic brain mris via a conditional diffusion probabilistic model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2023)
24. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 234–241. Springer (2015)
25. Rusak, E., et al.: A simple way to make neural networks robust against diverse image corruptions. In: European Conference on Computer Vision. pp. 53–69. Springer (2020)
26. Shussman, N., et al.: Colorectal polyps and polyposis syndromes. Gastroenterology Report **2**(1), 1–15 (2014)
27. Silva, J., et al.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery **9**, 283–293 (2014)
28. Simard, P.Y., et al.: Best practices for convolutional neural networks applied to visual document analysis. In: International Conference on Document Analysis and Recognition. vol. 3 (2003)
29. Srivastava, A., et al.: MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(5), 2252–2263 (2021)
30. Sudre, C.H., et al.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 240–248. Springer (2017)

31. Tajbakhsh, N., et al.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Transactions on Medical Imaging **35**(2), 630–644 (2015)
32. Wang, J., et al.: Stepwise feature fusion: Local guides global. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 110–120. Springer (2022)
33. Wang, Z., et al.: Binary latent diffusion. In: Conference on Computer Vision and Pattern Recognition. pp. 22576–22585 (2023)
34. Wei, J., et al.: Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 699–708. Springer (2021)
35. Xue, H., et al.: Diffusion-based adversarial sample generation for improved stealthiness and controllability. Advances in Neural Information Processing Systems **36** (2024)
36. Yang, J., et al.: TumorCP: A simple but effective object-level data augmentation for tumor segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 579–588. Springer (2021)
37. Yoon, J., et al.: Adversarial purification with score-based generative models. In: International Conference on Machine Learning. pp. 12062–12072. PMLR (2021)
38. Yun, S., et al.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: International Conference on Computer Vision. pp. 6023–6032 (2019)
39. Zhong, Z., et al.: Random erasing data augmentation. In: AAAI Conference on Artificial Intelligence. vol. 34, pp. 13001–13008 (2020)
40. Zhou, Z., et al.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 3–11. Springer (2018)