Anti-adversarial Consistency Regularization for Data Augmentation: Applications to Robust Medical Image Segmentation

Hyuna Cho, Yubin Han, and Won Hwa Kim(✉)

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

Abstract. Modern deep learning methods for semantic segmentation require labor-intensive labeling for large-scale datasets with dense pixel-level annotations. Recent data augmentation methods such as dropping, mixing image patches, and adding random noises suggest effective ways to address the labeling issues for natural images. However, they can only be restrictively applied to medical image segmentation as they carry risks of distorting or ignoring the underlying clinical information of local regions of interest in an image. In this paper, we propose a novel data augmentation method for medical image segmentation without losing the semantics of the key objects (e.g., polyps). This is achieved by perturbing the objects with quasi-imperceptible adversarial noises and training a network to expand discriminative regions with a guide of anti-adversarial noises. Such guidance can be realized by a consistency regularization between the two contrasting data, and the strength of regularization is automatically and adaptively controlled considering their prediction uncertainty. Our proposed method significantly outperforms various existing methods with high sensitivity and Dice scores and extensive experiment results with multiple backbones on two datasets validate its effectiveness.

Keywords: Adversarial attack and defense · Data augmentation · Semantic segmentation

1 Introduction

Semantic segmentation aims to segment objects in an image by classifying each pixel into an object class. Training a deep neural network (DNN) for such a task is known to be data-hungry, as labeling dense pixel-level annotations requires laborious and expensive human efforts in practice [23,32]. Furthermore, semantic segmentation in medical imaging suffers from privacy and data sharing issues [13,35] and a lack of experts to secure accurate and clinically meaningful regions of interest (ROIs). This data shortage problem causes overfitting for training DNNs, resulting in the networks being biased by outliers and ignorant of unseen data.

To alleviate the sample size and overfitting issues, diverse data augmentations have been recently developed. For example, CutMix [31] and CutOut [4] augment images by dropping random-sized image patches or replacing the removed

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 555–566, 2023.
https://doi.org/10.1007/978-3-031-43901-8_53

regions with a patch from another image. Random Erase [33] extracts noise from a uniform distribution and injects it into patches. Geometric transformations such as Elastic Transformation [26] warp images and deform the original shape of objects. Alternatively, feature perturbation methods augment data by perturbing data in feature space [7,22] and logit space [9].

Although these augmentation approaches have been successful for natural images, their usage for medical image semantic segmentation is quite restricted as objects in medical images contain non-rigid morphological characteristics that should be sensitively preserved. For example, basalioma (e.g., pigmented basal cell carcinoma) may look similar to malignant melanoma or mole in terms of color and texture [6,20], and early-stage colon polyps are mostly small and indistinguishable from background entrail surfaces [14]. In these cases, the underlying clinical features of target ROIs (e.g., polyp, tumor and cancer) can be distorted if regional colors and textures are modified with blur-based augmentations or geometric transformations. Also, cut-and-paste and crop-based methods carry risks of dropping or distorting key objects such that expensive pixel-level annotations could not be properly used. Considering the ROIs are usually small and underrepresented compared to the backgrounds, the loss of information may cause a fatal class imbalance problem in semantic segmentation tasks.

In these regards, we tackle these issues with a novel augmentation method without distorting the semantics of objects in image space. This can be achieved by slightly but effectively perturbing target objects with adversarial noises at the object level. We first augment hard samples with adversarial attacks [18] that deceive a network and defend against such attacks with anti-adversaries. Specifically, multi-step adversarial noises are injected into ROIs to *maximize* loss and induce false predictions. Conversely, anti-adversaries are obtained with anti-adversarial perturbations that *minimize* a loss which eventually become easier samples to predict. We impose consistency regularization between these contrasting samples by evaluating their prediction ambiguities via supervised losses with true labels. With this regularization, the easier samples provide adaptive guidance to the misclassified data such that the difficult (but object-relevant) pixels can be gradually integrated into the correct prediction. From active learning perspective [12,19], as vague samples near the decision boundary are augmented and trained, improvement on a downstream prediction task is highly expected.

We summarize our main contributions as follows: **1)** We propose a novel online data augmentation method for semantic segmentation by imposing object-specific consistency regularization between anti-adversarial and adversarial data. **2)** Our method provides a flexible regularization between differently perturbed data such that a vulnerable network is effectively trained on challenging samples considering their ambiguities. **3)** Our method preserves underlying morphological characteristics of medical images by augmenting data with quasi-imperceptible perturbation. As a result, our method significantly improves sensitivity and Dice scores over existing augmentation methods on Kvasir-Seg [11] and ETIS-Larib Polyp DB [25] benchmarks for medical image segmentation.

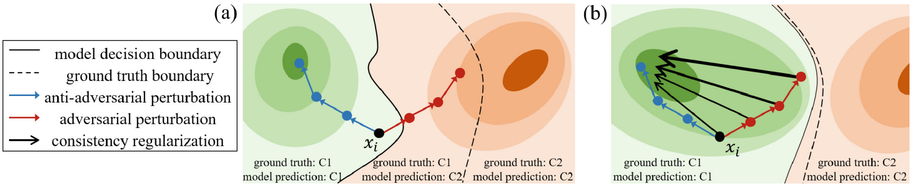

Fig. 1. (a) Conceptual illustration of the adversarial attack (red) and anti-adversarial perturbation (blue) in the latent feature space. Given a predicted sample embedding x_i, let its true label be a class 1 (C1). The adversarial attack sends the data point toward class 2 (C2) whereas the anti-adversarial perturbation increases its classification score. (b) Adaptive anti-adversarial consistency regularization (AAC) between the adversarially attacked data and the anti-adversary. (Color figure online)

2 Preliminary: Adversarial Attack and Anti-adversary

Adversarial attack is an input perturbation method that adds quasi-imperceptible noises into images to deceive a DNN. Given an image x, let μ be a noise bounded by l_∞-norm. While the difference between x and the perturbed sample $x' = x + \mu$ is hardly noticeable to human perception, a network $f_\theta(\cdot)$ can be easily fooled (i.e., $f_\theta(x) \neq f_\theta(x + \mu)$) as the μ pushes x' across the decision boundary.

To fool a DNN, Fast Gradient Sign Method (FGSM) [8] perturbs x toward *maximizing* a loss function L by defining a noise μ as the sign of loss derivative with respect to x as follows: $x' = x + \epsilon\,\mathtt{sign}(\nabla_x L)$, where ϵ controls the magnitude of perturbation. The authors in [18] proposed an extension of FGSM, i.e., Projected Gradient Descent (PGD), which is an iterative adversarial attack that also finds x' with a higher loss. Given an iteratively perturbed sample x'_t at t-th perturbation where $x'_0 = x$, the x'_t of PGD is defined as $x'_t = \prod(x'_{t-1} + \epsilon\,\mathtt{sign}(\nabla_x L))$ for T perturbation steps.

Recently, anti-adversarial methods were proposed for the benign purpose to defend against such attacks. The work in [15] used an anti-adversarial class activation map to identify objects and the authors in [1] proposed an anti-adversary layer to handle adversaries. In contrast to adversarial attacks, these works find μ that *minimizes* a loss to make easier samples to predict. Figure 1a shows multi-step adversarial and anti-adversarial perturbations in the latent space. To increase a classification score, the anti-adversarial noises move data away from the decision boundary, which is the opposite direction of the adversarial perturbations.

3 Method

Let $\{X_i\}_{i=1}^N$ be an image set with N samples each paired with corresponding ground truth pixel-level annotations Y_i. Our proposed method aims to 1) generate realistic images with adversarial attacks and 2) train a segmentation model $f_\theta(X_i) = Y_i$ for robust semantic segmentation with anti-adversarial consistency regularization (AAC). Figure 2 shows the overall training scheme with three phases: **1)** online data augmentation, **2)** computing adaptive AAC between

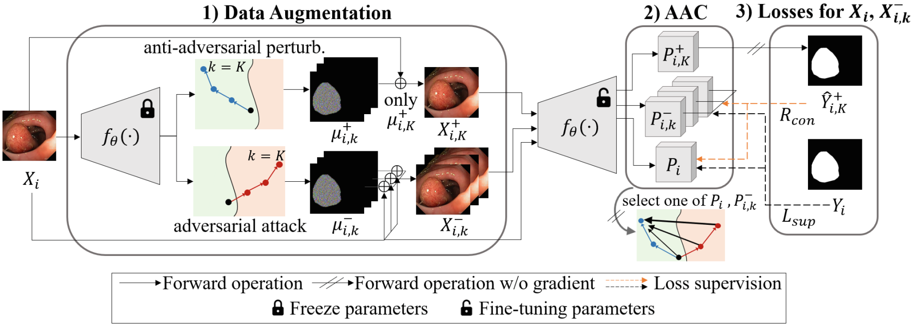

Fig. 2. An overview of our training scheme. Adversarial and anti-adversarial perturbations are iteratively performed for the objects of a given image X_i. Adversarial noise $\mu^-_{i,k}$ moves X_i across the decision boundary, whereas anti-adversarial noise $\mu^+_{i,k}$ pushes X_i away from the boundary. Downstream consistency regularization loss R_{con} minimizes the gap between adversaries $\{X^-_{i,k}\}^K_{k=1}$ and anti-adversary $X^+_{i,K}$.

differently perturbed samples, and **3)** updating the segmentation model using the loss from the augmented and original data. First, we generate plausible images with iterative adversarial and anti-adversarial perturbations. We separate the roles of perturbed data: adversaries are used as training samples and anti-adversaries are used to provide guidance (i.e., pseudo-labels) to learn the adversaries. Specifically, consistency regularization is imposed between these contrasting data by adaptively controlling the regularization magnitude in the next phase. Lastly, considering each sample's ambiguity, the network parameters θ are updated for learning the adversaries along with the given data so that discriminative regions are robustly expanded for challenging samples.

Data Augmentation with Object-Targeted Adversarial Attack. In many medical applications, false negatives (i.e., failing to diagnose a critical disease) are much more fatal than false positives. To deal with these false negatives, we mainly focus on training a network to learn diverse features at target ROIs (e.g., polyps) where disease-specific variations exist. To do so, we first exclude the background and perturb only the objects in the given image. Given o as the target object class and (p, q) as a pixel coordinate, a masked object is defined as $\hat{X}_i = \{(p, q) | X_{i_{(p,q)}} = o\}$.

As in PGD [18], we perform iterative perturbations on the $\hat{X}_i$ for K steps. Given $\hat{X}_{i,k}$ as a perturbed sample at k-th step ($k = 1, ..., K$), the adversarial and anti-adversarial perturbations use the same initial image as $X^-_{i,0} = \hat{X}_i$ and $X^+_{i,0} = \hat{X}_i$, respectively. With this input, the iterative adversarial attack is defined as

$$X^-_{i,k+1} = X^-_{i,k} + \mu^-_{i,k} = X^-_{i,k} + \epsilon\,\mathtt{sign}(\nabla_{X^-_{i,k}} L(f_\theta(X^-_{i,k}), Y_i)) \qquad (1)$$

where $\mu^-_{i,k} = argmax_\mu L(f_\theta(X^-_{i,k+1}), Y_i)$ is a quasi-imperceptible adversarial noise that fools $f_\theta(\cdot)$ and ϵ is a perturbation magnitude that limits the noise (i.e., $|\mu_{(p,q)}| \leq \epsilon$, s.t. $(p, q) \in \hat{X}_i$). Similarly, iterative anti-adversarial perturbation is defined as

$$X^{+}_{i,k+1} = X^{+}_{i,k} + \mu^{+}_{i,k} = X^{+}_{i,k} + \epsilon\, \mathtt{sign}(\nabla_{X^{+}_{i,k}} L(f_\theta(X^{+}_{i,k}), Y_i)). \tag{2}$$

In contrast to the adversarial attack in Eq. 1, the anti-adversarial noise $\mu^{+}_{i,k} = argmin_{\mu} L(f_\theta(X^{+}_{i,k+1}), Y_i)$ manipulates samples to increase the classification score.

Note that, generating noises and images are online and training-free as the loss derivatives are calculated with freezed network parameters. The adversaries $X^{-}_{i,1}, ..., X^{-}_{i,K}$ are used as additional training samples so that the network includes the non-discriminative yet object-relevant features for the prediction. On the other hand, as the anti-adversaries are sufficiently discriminative, we do not use them as training samples. Instead, only the K-th anti-adversary $X^{+}_{i,K}$ (i.e., the most perturbed sample with the lowest loss) is used for downstream consistency regularization to provide informative guidance to the adversaries.

Computing Adaptive Consistency Toward Anti-adversary. Let X'_i be either X_i or $X^{-}_{i,k}$. As shown in Fig. 1b, consistency regularization is imposed between the anti-adversary $X^{+}_{i,K}$ and X'_i to reduce the gap between samples with different prediction uncertainties. The weight of regularization between X'_i and $X^{+}_{i,K}$ is automatically determined by evaluating the gap in their prediction quality via supervised losses with ground truth Y_i as

$$w(X'_i, X^{+}_{i,K}) = max(\frac{1}{2}, \frac{l(f_\theta(X'_i), Y_i)}{l(f_\theta(X'_i), Y_i) + l(f_\theta(X^{+}_{i,K}), Y_i)}) = max(\frac{1}{2}, \frac{l(P'_i, Y_i)}{l(P'_i, Y_i) + l(P^{+}_{i,K}, Y_i)}), \tag{3}$$

where $l(\cdot)$ is Dice loss [28] and P_i is the output of $f_\theta(\cdot)$ for X_i. Specifically, if X'_i is a harder sample to predict than $X^{+}_{i,K}$, i.e., $l(P'_i, Y_i) > l(P^{+}_{i,K}, Y_i)$, the weight gets larger, and thus consistency regularization is intensified between the images.

Training a Segmentation Network. Let $\hat{Y}^{+}_{i,K}$ be a segmentation outcome, i.e., one-hot encoded pseudo-label from the network output $P^{+}_{i,K}$ of anti-adversary $X^{+}_{i,K}$. Given X_i and $\{X^{-}_{i,k}\}^{K}_{k=1}$ as training data, the supervised segmentation loss L_{sup} and the consistency regularization R_{con} are defined as

$$L_{sup} = \frac{1}{N}\sum_{i=1}^{N} l(P_i, Y_i) + \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} l(P^{-}_{i,k}, Y_i) \quad \text{and} \tag{4}$$

$$R_{con} = \frac{1}{N}\sum_{i=1}^{N} w(X_i, X^{+}_{i,K}) l(P_i, \hat{Y}^{+}_{i,K}) + \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} w(X^{-}_{i,k}, X^{+}_{i,K}) l(P^{-}_{i,k}, \hat{Y}^{+}_{i,K}). \tag{5}$$

Using the pseudo-label from anti-adversary as *a perturbation of the ground truth*, the network is supervised by diverse and realistic labels that contain auxiliary information that the originally given labels do not provide. With a hyperparameter α, the whole training loss $\mathcal{L} = L_{sup} + \alpha R_{con}$ is minimized via backpropagation to optimize the network parameters for semantic segmentation.

4 Experiments

4.1 Experimental Setup

Dataset. We conducted experiments on two representative public polyp segmentation datasets: Kvasir-SEG [11] and ETIS-Larib Polyp DB [25] (ETIS). Both are comprised of two classes: polyp and background. They provide 1000/196 (Kvasir-SEG/ETIS) input-label pairs in total and we split train/validation/test sets into 80%/10%/10% as in [5,10,24,27,29]. The images of Kvasir-SEG were resized to 512×608 ($H \times W$) and that of ETIS was set with 966×1255 resolution.

Implementation. We implemented our method on Pytorch framework with 4 NVIDIA RTX A6000 GPUs. Adam optimizer with learning rates of $4e$-3/$1e$-4 (Kavsir-SEG/ETIS) were used for 200 epochs with a batch size of 16. We set the number of perturbation steps K as 10 and the magnitude of perturbation ϵ as 0.001. The weight α for R_{con} was set to 0.01.

Baselines. Along with conventional augmentation methods (i.e., random horizontal and vertical flipping denoted as 'Basic' in Table 1), recent methods such as CutMix [31], CutOut [4], Elastic Transform [26], Random Erase [33], DropBlock [7], Gaussian Noise Training (GNT) [22], Logit Uncertainty (LU) [9] and Tumor Copy-Paste (TumorCP) [30] were used as baselines. Their hyperparameters were adopted from the original papers. The Basic augmentation was used in all methods including ours by default. For the training, we used K augmented images with the given images for all baselines as in ours for a fair comparison.

Evaluation. To verify the effectiveness of our method, evaluations are conducted using various popular backbone architectures such as U-Net [21], U-Net++ [34], LinkNet [2], and DeepLabv3+ [3]. As the evaluation metric, mean Intersection over Union (mIoU) and mean Dice coefficient (mDice) are used for all experiments on test sets. Additionally, we provide recall and precision scores to offer a detailed analysis of class-specific misclassification performance.

4.2 Comparison with Existing Methods

As shown in Table 1, our method outperforms all baselines for all settings by at most 10.06%p and 5.98%p mIoU margin on Kvasir-SEG and ETIS, respectively. Moreover, in Fig. 3, our method with U-Net on Kvasir-SEG surpasses the baselines by $\sim$8.2%p and $\sim$7.2%p in precision and recall, respectively. Note that, all baselines showed improvements in most cases. However, our method

Table 1. Performance comparison with existing data augmentation methods.

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	89.73	89.94	90.52	90.43	94.63	94.70	95.02	94.97
CutMix [31]	89.84	90.80	90.56	90.60	94.65	95.18	95.05	95.08
CutOut [4]	88.63	88.63	89.52	90.70	93.29	93.97	94.47	95.12
Elastic Trans. [26]	89.71	88.34	89.89	91.44	94.57	93.81	94.63	95.40
Random Erase [33]	88.73	89.45	90.72	90.94	94.03	94.43	95.14	95.25
DropBlock [7]	86.88	88.40	90.75	90.22	92.98	93.84	95.15	94.86
GNT [22]	82.37	88.71	90.32	90.88	90.36	94.02	94.91	95.22
LU [9]	89.51	90.84	87.71	90.52	94.46	95.20	93.45	95.02
TumorCP [30]	90.92	91.18	90.87	91.65	95.24	95.39	95.22	95.64
Ours	**92.43**	**91.43**	**91.51**	**92.42**	**96.07**	**95.52**	**95.57**	**96.06**
	ETIS							
No Aug.	83.80	83.96	82.18	82.52	91.18	91.28	90.22	90.43
Basic	86.08	87.02	84.75	82.69	92.52	93.06	91.75	90.52
CutMix [31]	86.03	86.78	82.52	82.20	92.49	92.92	90.42	90.83
CutOut [4]	84.37	84.90	86.55	84.50	91.52	91.83	92.79	91.60
Elastic Trans. [26]	85.12	85.10	86.55	84.13	91.96	91.95	92.79	91.38
Random Erase [33]	85.20	84.12	82.52	83.63	92.01	91.37	90.43	90.83
DropBlock [7]	82.52	85.33	82.49	84.27	90.42	92.08	90.41	91.46
GNT [22]	85.36	84.94	84.19	84.55	92.10	91.86	91.42	91.63
LU [9]	82.52	82.52	82.43	84.33	90.43	90.42	90.37	91.50
TumorCP [30]	82.59	84.99	85.69	85.32	90.46	91.89	92.30	92.08
Ours	**88.35**	**87.62**	**88.41**	**85.58**	**93.81**	**93.40**	**93.85**	**92.23**

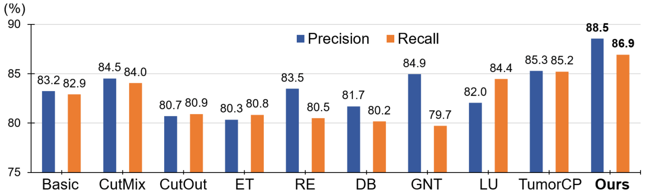

Fig. 3. Comparisons of precision and recall on the test set of Kvasir-SEG with U-Net.

performed better even compared with the TumorCP which uses seven different augmentations methods together for tumor segmentation. This is because our method *preserves the semantics of the key ROIs* with small but effective noises unlike geometric transformations [26,30], drop and cut-and-paste-based methods [4,7,30,31,33]. Also, as we augment uncertain samples that deliberately deceive a network as in Active Learning [12,16], our method is able to sensitively include the challenging (but ROI-relevant) features into prediction, unlike existing noise-based methods that extract noises from known distributions [9,22,30].

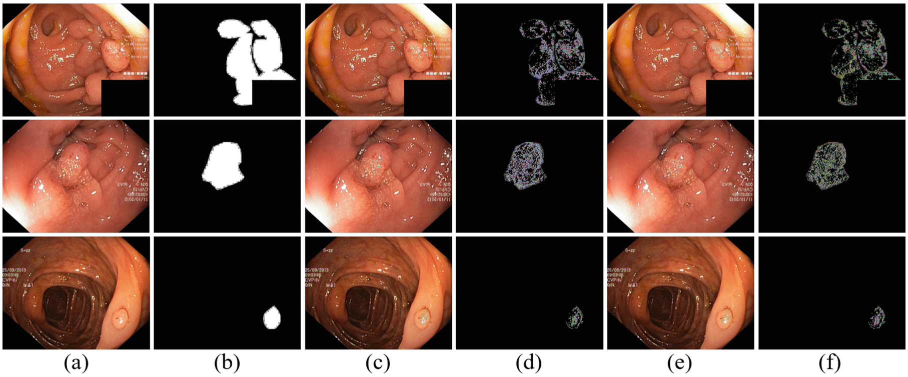

Fig. 4. (a) Input data (b) Ground truth label (c) Adversarially perturbed data (d) Adversarial noise (e) Anti-adversarially perturbed data (f) Anti-adversarial noise.

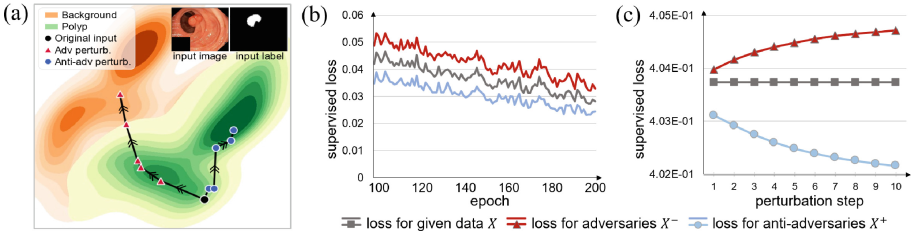

Fig. 5. (a) Density of pixel embeddings (orange and green), a sample (black) and its perturbations (red and blue) in 2D feature space via t-SNE. (b) and (c) show loss flow comparisons between a sample X and its perturbations X^- and X^+. Supervised losses are compared w.r.t. epochs and perturbation steps, and the anti-adversaries (blue) always demonstrate the lowest loss (i.e., closer to the ground truth). (Color figure online)

4.3 Analysis on Anti-adversaries and Adversaries

In Fig. 4, we visualize data augmentation results with (anti-) adversarial perturbations on Kvasir-SEG dataset. The perturbed data (c and e) are the addition of noise (d and f) to the given data (a), respectively. Interestingly, while adversaries (c) and anti-adversaries (e) are visually indistinguishable, they induce totally opposite model decisions towards different classes. In Fig. 5, we qualitatively and quantitatively compared their effects via visualizing perturbation trajectories in the feature space projected with t-SNE [17] and comparing their supervision losses. In Fig. 5a, the adversarial attacks send a pixel embedding of a polyp class to the background class, anti-adversarial perturbations push it towards the true class with a higher classification score. Also, loss comparisons in Fig. 5b and 5c demonstrate that the anti-adversaries (blue) are consistently easier to predict than the given data (grey) and adversaries (red) during the training and their differences get larger as the perturbations are iterated.

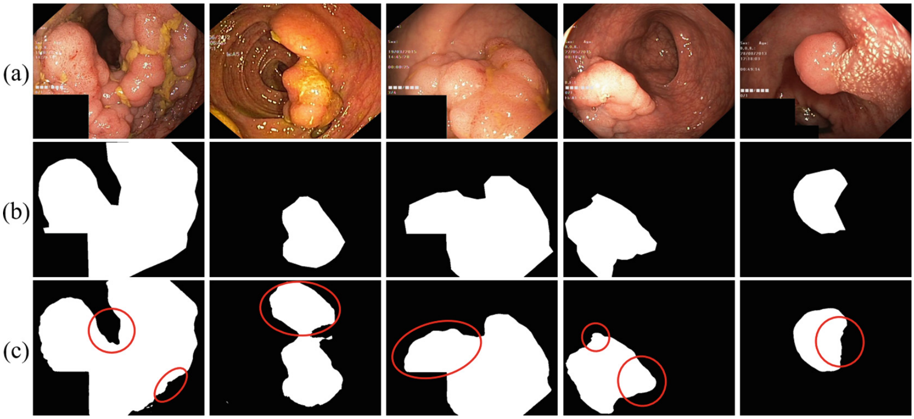

Fig. 6. Comparison of ground truths and pseudo labels from anti-adversaries. (a) Input data (b) Ground truth label (c) Pseudo-label $\hat{Y}_K^+$. (Color figure online)

These results confirm that the anti-adversaries send their pseudo label $\hat{Y}_K^+$ closer to the ground truth with a slight change. Therefore, they can be regarded as a perturbation of the ground truth that contain a potential to provide additional information to train a network on the adversaries. We empirically show that $\hat{Y}_K^+$ is able to provide such auxiliary information that the true labels do not provide, as our method performs better *with* R_{con} (i.e., $\mathcal{L} = L_{sup} + \alpha R_{con}$, 92.43% mIoU) than the case *without* R_{con} (i.e., $\mathcal{L} = L_{sup}$, 92.15% mIoU) using U-Net on Kvasir-SEG. Training samples in Fig. 6 show that the pseudo-labels $\hat{Y}_K^+$ can capture detailed abnormalities (marked in red circles) which are not included in the ground truths. Moreover, as the AAC considers sample-level ambiguity, the effect from $\hat{Y}_K^+$ is sensitively controlled and a network can selectively learn the under-trained yet object-relevant features from adversarial samples.

5 Conclusion

We present a novel data augmentation method for semantic segmentation using a flexible anti-adversarial consistency regularization. In particular, our method is tailored for medical images that contain small and underrepresented key objects such as a polyp and tumor. With object-level perturbations, our method effectively expands discriminative regions on challenging samples while preserving the morphological characteristics of key objects. Extensive experiments with various backbones and datasets confirm the effectiveness of our method.

Acknowledgement. This research was supported by IITP-2022-0-00290 (50%), IITP-2019-0-01906 (AI Graduate Program at POSTECH, 10%), IITP-2022-2020-0-01461 (ITRC, 10%) and NRF-2022R1A2C2092336 (30%) funded by the Korean government (MSIT).

References

1. Alfarra, M., Pérez, J.C., et al.: Combating adversaries with anti-adversaries. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 5992–6000 (2022)
2. Chaurasia, A., Culurciello, E.: Linknet: exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing, pp. 1–4. IEEE (2017)
3. Chen, L.C., Zhu, Y., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 801–818 (2018)
4. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
5. Fan, D.-P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26
6. Felder, S., Rabinovitz, H., et al.: Dermoscopic pattern of pigmented basal cell carcinoma, blue-white variant. Dermatol. Surg. **32**(4), 569–570 (2006)
7. Ghiasi, G., Lin, T.Y., et al.: Dropblock: a regularization method for convolutional networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
8. Goodfellow, I.J., Shlens, J., et al.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
9. Hu, Y., Zhong, Z., Wang, R., Liu, H., Tan, Z., Zheng, W.-S.: Data augmentation in logit space for medical image classification with limited training data. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 469–479. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_45
10. Jha, D., Riegler, M.A., et al.: Doubleu-net: a deep convolutional neural network for medical image segmentation. In: IEEE International Symposium on Computer-Based Medical Systems, pp. 558–564. IEEE (2020)
11. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
12. Juszczak, P., Duin, R.P.: Uncertainty sampling methods for one-class classifiers. In: Proceedings of ICML-03, Workshop on Learning with Imbalanced Data Sets II, pp. 81–88 (2003)
13. Kaissis, G.A., Makowski, M.R., et al.: Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. **2**(6), 305–311 (2020)
14. Lee, J.Y., Jeong, J., et al.: Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. Sci. Rep. **10**(1), 8379 (2020)
15. Lee, J., Kim, E., et al.: Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4071–4080 (2021)
16. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine Learning Proceedings 1994, pp. 148–156. Elsevier (1994)
17. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
18. Madry, A., Makelov, A., et al.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)

19. Nguyen, V.L., Shaker, M.H., et al.: How to measure uncertainty in uncertainty sampling for active learning. Mach. Learn. **111**(1), 89–122 (2022)
20. Parmar, B., Talati, B.: Automated melanoma types and stages classification for dermoscopy images. In: 2019 Innovations in Power and Advanced Computing Technologies, vol. 1, pp. 1–7. IEEE (2019)
21. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Rusak, E., et al.: A simple way to make neural networks robust against diverse image corruptions. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12348, pp. 53–69. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58580-8_4
23. Saleh, F.S., Aliakbarian, M.S., et al.: Effective use of synthetic data for urban scene semantic segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 84–100 (2018)
24. Sanderson, E., Matuszewski, B.J.: FCN-transformer feature fusion for polyp segmentation. In: Yang, G., Aviles-Rivero, A., Roberts, M., Schönlieb, C.B. (eds.) MIUA 2022. LNCS, vol. 13413, pp. 892–907. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-12053-4_65
25. Silva, J., Histace, A., et al.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**, 283–293 (2014)
26. Simard, P.Y., Steinkraus, D., et al.: Best practices for convolutional neural networks applied to visual document analysis. In: International Conference on Document Analysis and Recognition, vol. 3 (2003)
27. Srivastava, A., Jha, D., et al.: MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation. IEEE J. Biomed. Health Inform. **26**(5), 2252–2263 (2021)
28. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
29. Wang, J., Huang, Q., et al.: Stepwise feature fusion: Local guides global. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 110–120. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_11
30. Yang, J., Zhang, Y., Liang, Y., Zhang, Y., He, L., He, Z.: TumorCP: a simple but effective object-level data augmentation for tumor segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 579–588. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_55
31. Yun, S., Han, D., et al.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
32. Zhao, X., Vemulapalli, R., et al.: Contrastive learning for label efficient semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10623–10633 (2021)
33. Zhong, Z., Zheng, L., et al.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13001–13008 (2020)

34. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
35. Ziller, A., Usynin, D., et al.: Medical imaging deep learning with differential privacy. Sci. Rep. **11**(1), 1–8 (2021)