S2S2: Semantic Stacking for Robust Semantic Segmentation in Medical Imaging

Yimu Pan[1], **Sitao Zhang**[1], **Alison D. Gernand**[1], **Jeffery A. Goldstein**[2], **James Z. Wang**[1]

[1]The Pennsylvania State University, University Park
[2]Northwestern University, Chicago
{ymp5078, sitao.zhang, adg14}@psu.edu, ja.goldstein@northwestern.edu, jwang@psu.edu

Abstract

Robustness and generalizability in medical image segmentation are often hindered by scarcity and limited diversity of training data, which stands in contrast to the variability encountered during inference. While conventional strategies—such as domain-specific augmentation, specialized architectures, and tailored training procedures—can alleviate these issues, they depend on the availability and reliability of domain knowledge. When such knowledge is unavailable, misleading, or improperly applied, performance may deteriorate. In response, we introduce a novel, domain-agnostic, add-on, and data-driven strategy inspired by image stacking in image denoising. Termed "semantic stacking," our method estimates a denoised semantic representation that complements the conventional segmentation loss during training. This method does not depend on domain-specific assumptions, making it broadly applicable across diverse image modalities, model architectures, and augmentation techniques. Through extensive experiments, we validate the superiority of our approach in improving segmentation performance under diverse conditions.

Code — https://github.com/ymp5078/Semantic-Stacking
Extended version — https://arxiv.org/abs/2412.13156

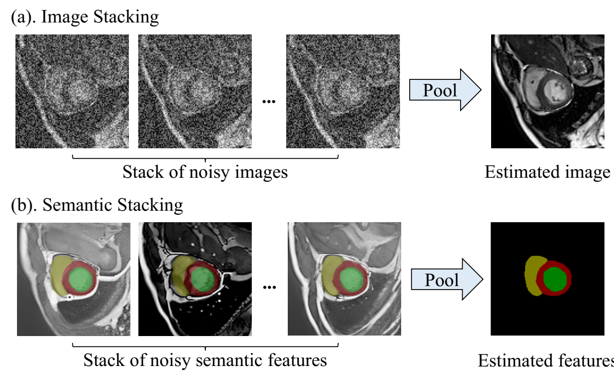

Figure 1: An illustration of the proposed semantic stacking approach compared to traditional image stacking for noise reduction. (a) Image stacking for noise reduction in imagery. (b) Our semantic stacking technique, aimed at reducing feature noise. Here, we illustrate semantic features through semantic segmentation maps for clarity, though our method operates on encoded features.

1 Introduction

In the rapidly evolving field of computer vision, significant progress in image recognition has been driven by not only groundbreaking developments in model architectures (He et al. 2016; Dosovitskiy et al. 2021; Ronneberger, Fischer, and Brox 2015) but also deliberated training recipes (Wightman, Touvron, and Jégou 2021; Liu et al. 2022; Woo et al. 2023) and innovative augmentation techniques (Cubuk et al. 2020, 2019; Hendrycks et al. 2020; Yun et al. 2019). These advancements largely stem from the abundance and diversity of natural image datasets (Russakovsky et al. 2015; Lin et al. 2014; Krishna et al. 2017), which enable models to learn robust, generalizable features.

In contrast, medical image analysis faces distinct challenges. Data are often scarce and originate from a limited number of sites, captured through specific imaging devices, or within certain modalities (Litjens et al. 2017). High annotation costs further exacerbates these challenges, making the pursuit of training robust models in medical image analysis a paramount yet elusive goal (Aggarwal et al. 2021; Nguyen et al. 2023). The need for model robustness in medical imaging is critical: errors can have severe clinical consequences (Esteva et al. 2019). While augmentation techniques can mitigate data limitation, they can be inadequate, or even detrimental, if misapplied to medical contexts (Perez et al. 2018; Ozbulak, Van Messem, and De Neve 2019). The heterogeneous nature of medical images—ranging from Computerized Tomography (CT) and Magnetic Resonance Imaging (MRI) scans to standard RGB photographs—further complicates the development of universally applicable augmentation strategies. Together, these issues underscore the urgent need for approaches that enhance model robustness without succumbing to the pitfalls of domain-specified bias (Seyyed-Kalantari et al. 2021; Roberts et al. 2021).

In this work, we introduce an add-on training strategy, Semantic Stacking for Semantic Segmentation (S2S2), that can be seamlessly integrated into existing pipelines. Unlike previous approaches that focus narrowly on either in-domain performance (Chen et al. 2021) or out-of-domain robustness (Su et al. 2023), our method enhances both. Drawing inspiration from image stacking in image denoising, where multiple noisy images are stacked and pooled to estimate a denoised image, we propose *semantic stacking*: we first estimate a denoised semantic representation from a stack of synthetic images and then encourage the network to learn from this representation. We argue that this estimated denoised semantic representation more closely reflects the underlying ground truth, thus reducing both bias and variance. This method directs models toward a denoised semantic representation, distinguishing itself through a data-driven design that avoids domain-specific assumptions. This versatility makes our approach particularly advantageous across diverse image modalities, serving as an invaluable asset in scenarios where broad generalizability is critical and specific domain knowledge remains elusive.

Additionally, directly estimating the semantic stacking requires obtaining the semantic representation from all images in the stack. Running the network through all images in the stack at each iteration is resource- and time-intensive, or even impractical, as the stack grows. Through theoretical analysis, we derived a practical upper bound for semantic variations, transforming the semantic stacking into an operation involving only two images per iteration. This transformation makes learning from the semantic stacking feasible.

As general-purpose interactive segmentation tools gain traction (Kirillov et al. 2023; Ma et al. 2024; Pan et al. 2023), the need for training methodologies compatible with mixed image modalities is becoming increasingly critical. In such contexts, the data-driven design of S2S2 offers significant benefits, as integrating knowledge from different domains into a single training strategy is challenging. We validate our proposed strategy across popular network architectures and demonstrate its effectiveness in improving both in-domain performance and single-source domain generalization across various CT, MRI, and RGB images.

Our main contributions are summarized as follows:

- We propose a versatile add-on training strategy, semantic stacking, that enhances robustness without requiring specialized domain knowledge.
- We provide theoretical analysis enabling a practical, efficient method for learning the semantic stacking that scales to large datasets.
- We demonstrate our method's ability to improve both in-domain performance and out-of-domain robustness.

2 Related Work

Data Augmentation in Medical Image Analysis

Early approaches adapted data augmentation strategies from natural images to medical images (Ronneberger, Fischer, and Brox 2015; Milletari, Navab, and Ahmadi 2016). For example, nnU-net (Isensee et al. 2021) employed a predefined pipeline with operations like rotation, scaling, Gaussian noise, and blur. Inspired by AutoAug (Cubuk et al. 2019), later approaches explored automated data augmentation strategies, but these relied on traditional spatial and color transformations (Xu, Li, and Zhu 2020; Qin et al. 2020; Lyu et al. 2022; Yang et al. 2019).

While traditional methods are simple and effective, they fail to fully exploit the distinctive characteristics of medical images. Recently, generative models, such as GANs (Goodfellow et al. 2020; Denton et al. 2015; Beers et al. 2018; Yi, Walia, and Babyn 2019) and diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Kazerouni et al. 2023), have been used for synthesizing medical images. Most generative approaches have focused on classification tasks (Pinaya et al. 2022; Khader et al. 2023; Tang et al. 2023; Ye et al. 2023; Peng et al. 2023; Deo et al. 2023), with segmentation tasks receiving less attention. Notable exceptions include brainSPADE (Fernandez et al. 2022), which trained segmentation models solely with synthetic data, and DPGAN (Chai et al. 2022), which used synthetic augmentation to address class imbalance. However, these methods exhibit limitations in performance and applicability. Our work addresses this gap by proposing a versatile add-on training strategy that enhances both in-domain performance and out-of-domain robustness.

Single-Source Domain Generalization

Domain generalization aims to train models that perform reliably on previously unseen data distributions (Wang et al. 2022; Zhou et al. 2022a). Our goal aligns with single-source domain generalization (SDG), where models are trained without access to target or additional source domain information. Recent SDG approaches use specialized augmentations, adapt model architectures, or propose unique training methods (Zhou et al. 2022a; Su et al. 2023; Xu et al. 2021; Zhou et al. 2022b; Huang et al. 2020; Hu, Liao, and Xia 2023; Guo, Liu, and Yuan 2024; Liao et al. 2024). Different from these methods that only focus of out-of-domain robustness, our approach provides a versatile add-on strategy that enhances model robustness and semantic representation without changing existing augmentations, architectures, or training paradigms, ensuring strong in-domain performance while preparing models for deployment across varied medical imaging domains.

3 Method

From Image Stacking to Semantic Stacking

In semantic segmentation, our objective is to recover the ground truth segmentation map y from an input image x. This entails classifying each pixel in the semantic feature map t to yield the segmentation map $y = \mathcal{H}(t)$, where $\mathcal{H}$ denotes a classifier. Ideally, the goal is to minimize the discrepancy between the estimated semantic feature map $\hat{t}$ and the truth but unknown t. Since t itself is not directly observable, our practical objective shifts to reducing the difference between the estimated segmentation map $\hat{y}$ and the true segmentation map y. However, because the classifier $\mathcal{H}$ may map different inputs to the same output label, suggesting that the feature map derived from the training data guided by pixel-level supervision may inherently carry bias. To address

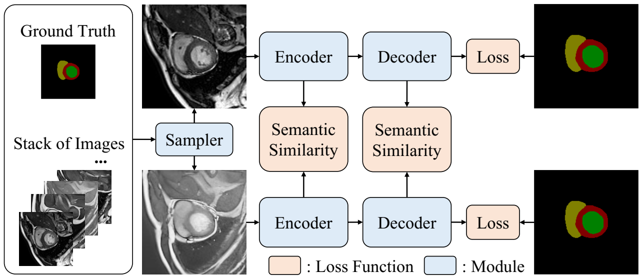

Figure 2: Illustration of the proposed S2S2 framework. A stack of images given is generated from the ground truth semantic segmentation map. Two samples from the stack are then fed into the network, where the training process is guided by the consistency between features alongside the segmentation loss.

this, we leverage the concept of image stacking, a technique traditionally utilized in image denoising, to obtain a more accurate approximation of t.

In image denoising, as depicted in Fig. 1 (a), the primary objective is to estimate the unknown ground truth image x. Image stacking employs multiple noisy images to approximate the ground truth image. Let $\{x_1, \cdots, x_n\}$ represent a collection of noisy images sampled from $\mathcal{N}(x, \sigma_x)$, where x denotes the ground truth image and σ_x the noise variance. Let $\hat{x} = \mathcal{P}(x_1, \cdots, x_n)$ denotes the pooled result of the image stack using mean or median pooling method $\mathcal{P}$, then

$$\hat{x} \sim \mathcal{N}\left(x, \frac{\sigma_x}{\sqrt{n}}\right) \ . \tag{1}$$

As n grows, the precision of the estimated image $\hat{x}$ relative to the ground truth image x enhances.

Adapting this principle for semantic feature estimation, as illustrated in Fig. 1 (b), allows us to approach semantic feature mapping with a novel perspective. Specifically, for a given network $\mathcal{F}$ without regularization, we can acquire a semantic feature map $t_i = \mathcal{F}(x_i) \sim \mathcal{N}(t, \sigma)$, where t represents the ground truth semantic feature map. Following the same principle as in image stacking, if we possess a collection of semantic features $\{t_1, \cdots, t_n\}$ corresponding to the identical semantic feature map, pooling these features as $\hat{t} = \mathcal{P}(t_1, \cdots, t_n)$ yields an estimated feature map with diminished variance, expressed as:

$$\hat{t} \sim \mathcal{N}\left(t, \frac{\sigma}{\sqrt{n}}\right) \ . \tag{2}$$

Let $\mathcal{D}$ denote a distance metric. Utilizing $\hat{t}$ as an approximation of t allows for the optimization of $\mathcal{D}(t_i, \hat{t})$ to enhance the training of $\mathcal{F}$, aiming for $\mathcal{F}$ to generate an accurate $\hat{t}$.

Practical Objective for Semantic Stacking

Direct approximation of $\hat{t}$ from t necessitates constructing a stack of n feature maps, which becomes impractical for large n due to the need for multiple activation copies. To overcome this, we use Bayesian updating. Given a sequence of feature maps $\{t_1, \cdots, t_n\}$, the estimated posterior distribution of $\hat{t}$ is defined as:

$$\mathbb{E}[\hat{t}] = \frac{\sigma^2 t_0 + \sigma_0^2 \sum_{i=1}^n t_i}{\sigma^2 + n\sigma_0^2} \ , \quad \mathrm{Var}[\hat{t}] = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2} \ , \tag{3}$$

where σ_0 and t_0 are the prior distribution's hyperparameters. Assuming $\mathcal{D}$ satisfies the triangle inequality and adopting the L_1 distance for simplicity, minimizing $\mathcal{D}(t_i, \mathbb{E}[\hat{t}])$ is achieved through the following optimization:

$$\begin{aligned}
\mathcal{D}\left(t_i, \mathbb{E}[\hat{t}]\right) &= \left| t_i - \frac{\sigma^2 t_0 + \sigma_0^2 \sum_{j=1}^n t_j}{\sigma^2 + n\sigma_0^2} \right| \\
&= \frac{1}{\sigma^2 + n\sigma_0^2} \left| \sigma^2 (t_i - t_0) + \sigma_0^2 \sum_{j \neq i}^n (t_i - t_j) \right| \\
&\leq \frac{\sigma^2}{\sigma^2 + n\sigma_0^2} |t_i - t_0| + \frac{\sigma_0^2}{\sigma^2 + n\sigma_0^2} \sum_{j \neq i}^n |t_i - t_j| \\
&\leq \frac{\sigma^2}{\sigma^2 + n\sigma_0^2} \mathcal{D}(t_i, t_0) + \frac{\sigma_0^2}{\sigma^2 + n\sigma_0^2} \sum_{j \neq i}^n \mathcal{D}(t_i, t_j)
\end{aligned} \tag{4}$$

We observe that $\mathcal{D}(t_i, \mathbb{E}[\hat{t}])$ is upper-bounded by a weighted sum of all $\mathcal{D}(t_i, t_j)$. Therefore, minimizing $\mathcal{D}(t_i, \mathbb{E}[\hat{t}])$ effectively requires minimizing $\mathcal{D}(t_i, t_j)$ between any pair of feature maps in the stack. This insight permits sampling just two images at a time from the stack and minimizing the distance between their corresponding feature maps. The resulting semantic consistency loss is formulated as:

$$\mathcal{L}_{\text{sc}} = \mathcal{D}\left(\mathcal{F}(x_i), \mathcal{F}(x_j)\right) \ , \tag{5}$$

where $\mathcal{D}$ is a suitable distance metric that adheres to the triangle inequality, with x_i, x_j being two distinct samples from the stack of images corresponding to the same segmentation map. This methodology, termed S2S2, is illustrated in Fig. 2.

Constructing Semantic Stack

Generating images that align with a specific semantic segmentation map poses a significant challenge, particularly in medical image analysis, where annotations are costly and scarce. Recent advances in generative models have provided new ways for synthesizing realistic medical images. In contrast to traditional photometric adjustments like intensity or scale (Cai, Fan, and Fang 2023) changes that only account for variations due to equipment differences, variations in human organs are can be learned and simulated using generative models. Utilizing a conditional image generation approach, we generate a set of images based on a given segmentation map. This generative strategy not only enhances diversity but also reduces reliance on dataset-specific knowledge, such as particular intensity variations or color shifts introduced in methods like SLAug (Su et al. 2023), thereby offering a more generalized solution. Specifically, we fine-tune a Stable Diffusion model (Rombach et al. 2022), employing ControlNet (Zhang, Rao, and Agrawala 2023) for segmentation map control. Although the synthesized images might not precisely replicate the ground truth distribution, we suggest that generating a substantial volume of high-quality images can improve model performance.

After generating a series of semantic feature maps $\{t_1, \cdots, t_n\} \sim \mathcal{N}(t^g, \sigma^g)$ from the synthesized images, where σ^g reflects the variance indicative of the generated feature maps' quality, and t^g represents the mean, we posit that $t^g \approx t$. This assumption rests on the premise that fine-tuning the generative model with accurate ground truth annotations aligns the mean of the generated feature maps with the ground truth mean, while the variance captures residual discrepancies. In line with previous formulations (Eq. 2), we have: $\hat{t}^g \sim \mathcal{N}(t, \frac{\sigma^g}{\sqrt{n}})$. Specifically, if $\frac{\sigma^g}{\sqrt{n}} \leq \sigma$, then $\hat{t}^g$ offers a more accurate estimate of the ground truth semantic feature map, which indicates the potential of enhancing model performance through the minimization of $\mathcal{D}(t_i, \hat{t}^g)$. Although empirically validating this condition may be challenging due to the unknown values of σ^g and σ, theoretical guarantees ensure its validity as n increases.

4 Dataset

To comprehensively evaluate the efficacy of our method across diverse medical image segmentation scenarios, we conducted experiments assessing both in-domain and out-of-domain performance. These evaluations covered a variety of imaging modalities, including RGB, CT, and MRI. Details on data prepossessing are in the Appendix.

For RGB images, we utilized two polyp segmentation datasets: CVC-ClinicDB (Bernal et al. 2015) and Kvasir-SEG (Jha et al. 2020). CVC-ClinicDB comprises 612 labeled images, while Kvasir-SEG includes 1,000 labeled images. These datasets, originating from distinct sites and captured using different devices, provide variability in the data. The processing of RGB datasets adhered to the methods described in previous studies (Sanderson and Matuszewski 2022). For CT images, we evaluated using the Synapse multi-organ segmentation dataset [1], which includes 30 abdominal CT scans with comprehensive annotations for multi-organ segmentation tasks. In the MRI category, our evaluation encompassed several datasets focused on abdominal and cardiac segmentation. The Combined Healthy Abdominal Organ Segmentation (CHAOS) (Kavur et al. 2021) dataset consists of 20 T2-SPIR MRI images focused on abdominal organ segmentation. For cardiac segmentation, we included a dataset (Zhuang et al. 2022) comprising 45 late gadolinium enhanced (LGE) MRI images and 45 balanced steady-state free precession (bSSFP) MRI images, alongside the Automatic Cardiac Diagnosis Challenge (ACDC) (Bernard et al. 2018) dataset, which features 100 cases of Cine MRI images.

5 Results

Only average metrics are reported in this section for clarity; the class-specific metrics are detailed in the Appendix. Since S2S2 is applicable to any method, we evaluate its performance on representative methods and include baseline methods as references. These baseline methods include MSRF-Net (Srivastava et al. 2021) and PraNet (Fan et al. 2020) for the Kvasir and CVC datasets; R50-AttnUNet (Schlemper et al. 2019), ViT-CUP (Dosovitskiy et al. 2021), and R50-ViT-CUP (Dosovitskiy et al. 2021) for the Synapse and ACDC datasets; and Cutout (DeVries and Taylor 2017), RSC (Huang et al. 2020), MixStyle (Zhou et al. 2021), AdvBias (Carlucci et al. 2019), RandConv (Xu et al. 2021), and CSDG (Ouyang et al. 2022) for abdominal and cardiac datasets.

Implementation Details

We compared S2S2 against several established approaches in medical image analysis, as well as a state-of-the-art technique in single-source domain generalization. These established techniques serve as baseline methods for our experiments. All experimental procedures adhered to the methodologies outlined by these baselines, with exceptions made solely for components that integrate our proposed approach (detailed in the Appendix). Synthetic images were generated using Stable Diffusion 2.5 fine-tuned on training images with segmentation-map-controlled ControlNet for 100 epochs. Further details are provided in the Appendix.

Contemporary models for semantic segmentation are typically comprised of an encoder for capturing high-level semantics and a decoder for pixel-level details. We hypothesize that both levels of features are useful and apply our semantic consistency loss to both components, denoted as $\mathcal{L}_{\text{sc}}^{\text{enc}}$ and $\mathcal{L}_{\text{sc}}^{\text{dec}}$, respectively. The final loss function is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \alpha^{\text{enc}}\mathcal{L}_{\text{sc}}^{\text{enc}} + \alpha^{\text{dec}}\mathcal{L}_{\text{sc}}^{\text{dec}} , \quad (6)$$

where $\mathcal{L}_{\text{seg}}$ represents the segmentation loss derived from any chosen method. The variables α^{enc} and α^{dec} are the weights for the consistency losses. For simplicity, we define the distance function as $\mathcal{D}(t_i, t_j) = 1 - \text{CosSim}(t_i, t_j)$ where CosSim is cosine similarity.

In-domain Performance

As an add-on method, our foundational premise posits that the integration of S2S2 should not detrimentally affect the performance of the baseline method within the scope of in-domain evaluation. To verify this, we rigorously evaluated S2S2 across a variety of acclaimed network architectures on datasets derived from RGB, CT, and MRI images. Furthermore, we aim to underscore the advantages of adopting a universally applicable method over approaches that are narrowly tailored to specific tasks. To this end, we incorporated SLAug (Su et al. 2023), a state-of-the-art method devised for enhancing single-domain generalization in CT/MRI imaging, into our in-domain benchmarks.

As shown in Table 1, the integration of S2S2 significantly elevates the in-domain performance for CT/MRI datasets on widely recognized models. Similarly, Table 2 demonstrates that the deployment of S2S2 concurrently amplifies the efficacy of FCBFormer on RGB datasets.

[1] https://www.synapse.org/\#!Synapse:syn3193805/wiki/217789

Method	Synapse	ACDC	Mean
R50-AttnUNet	75.57	86.75	81.16
ViT-CUP	67.86	81.45	74.66
R50-ViT-CUP	71.29	87.57	79.43
TransUNet	76.86	88.86	82.86
+S2S2	**81.19**	**90.40**	**85.80**+2.94

Table 1: In-domain performance comparison on the Synapse multi-organ CT dataset and ACDC dataset. Dice score (%) is used as the evaluation metric. The best-performing method is highlighted in bold, and the second-best is underlined. The improvement achieved by S2S2 is indicated.

Notably, the baseline methods already incorporate augmentation techniques such as color space and spatial augmentation, indicating that S2S2 operates independently of the baseline method or image modality. The semantic stack provides a superior representation of the ground truth semantic feature map than the original unconstrained semantic feature map. We observe an enhanced performance with an increase in the number of classes, potentially attributable to the generative model's refined control over image generation or the amplified complexity of maintaining semantic consistency across broader classes.

Method	Kvasir	CVC	Mean
MSRF-Net	92.17	94.20	93.19
PraNet	89.80	89.90	89.90
SLAug	84.85	85.39	85.12
SLAug+S2S2	85.33	88.76	87.05+1.93
FCBFormer	91.90	93.46	92.68
+S2S2	**93.20**	**94.88**	**94.04**+1.36

Table 2: In-domain performance comparison on RGB datasets. Dice score (%) is used as the evaluation metric.

In addition, our analysis reveals that SLAug (Su et al. 2023), despite being specifically engineered for CT/MRI imaging modalities through the exploitation of domain-specific knowledge, fails to deliver comparable benefits for RGB imaging (Table 2). However, the subsequent application of S2S2 atop SLAug results in a discernible enhancement in performance metrics, indicating that S2S2 introduces an additional layer of supervision beyond the capabilities of domain-specific augmentation techniques. More importantly, even for the CT/MRT images, which SLAug was originally tailored for, S2S2 outperforms the baseline method, as shown in Table 3. This finding suggests that methods focused on domain-specific generalization may inadvertently compromise in-domain performance while optimizing for out-of-domain applicability. In contrast, our approach avoids making assumptions about the application domain, thereby ensuring consistent improvements in in-domain performance across diverse datasets and imaging modalities.

Qualitative Evaluation. The comparison in Fig. 3 revealed several distinct advantages of our approach. First, S2S2 demonstrates superior capability in identifying the presence or absence of small objects, as evident in rows 1, 4, and 7.

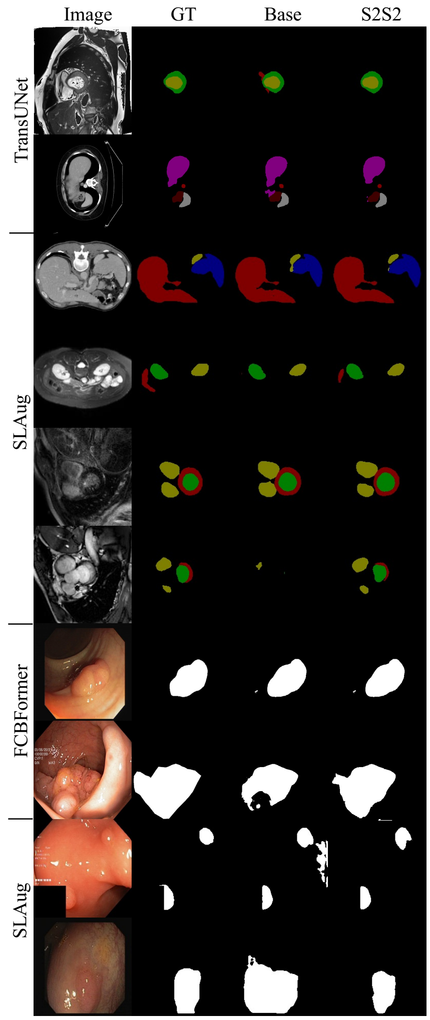

Figure 3: Visualization of the improvement achieved by applying S2S2 to the base method in the in-domain setting. 'GT' is the ground truth. 'Base' refers to the corresponding method without S2S2.

Method	Abdominal		Cardiac		Mean
	CT	MRI	bSSFP	LGE	
Supervised (CSDG)	**89.74**	90.85	88.16	**88.15**	**89.23**
SLAug	82.66	90.60	**92.27**	87.35	88.22
+S2S2	84.21	**91.28**	92.16	87.62	88.82+0.60

Table 3: In-domain performance comparison on slices of 3D medical image datasets. Dice score (%) is used as the evaluation metric.

Second, it tends to generate smoother segmentation masks, observable in rows 2, 8, and 10. Lastly, S2S2 adopts a more conservative approach in its predictions, particularly highlighted in row 9.

Out-of-domain Performance

In our out-of-domain evaluations, we benchmarked the S2S2 method against reproducible state-of-the-art, aligning with the settings of FCBFormer (Sanderson and Matuszewski 2022) for polyp segmentation tasks on RGB images and SLAug (Su et al. 2023) for abdominal organ and cardiac segmentation tasks on CT/MRI images. These comparisons validate not only the robustness of our approach in established domains but also its superior generalization capabilities in unseen domains.

Method	Abdominal		Cardiac		Mean
	CT-MRI	MRI-CT	bSSFP-LGE	LGE-bSSFP	
Cutout	80.12	70.50	78.87	85.92	78.85
RSC	74.09	66.07	77.51	85.60	75.82
MixStyle	77.80	63.95	75.21	86.34	75.83
AdvBias	80.17	64.84	79.62	86.27	77.73
RandConv	80.66	76.56	83.73	87.24	82.05
CSDG	86.31	80.40	85.01	86.99	84.68
SLAug	**88.55**	81.70	**86.42**	87.17	85.96
+S2S2	87.75	**83.15**	86.06	**87.49**	**86.11**+.15

Table 4: Out-of-domain performance comparison on slices of 3D medical image datasets. Dice score (%) is used as the evaluation metric.

Beyond demonstrating improvements in in-domain performance, our method also exhibits notable improvements in out-of-domain generalization, as shown in Table 5. Similar to what we observed in in-domain evaluation, the domain-specific method SLAug delivers suboptimal performance on RGB images. However, integrating the proposed S2S2 method fills this gap, enhancing its effectiveness. These results underscore the applicability of S2S2 in augmenting out-of-domain generalization capabilities without necessitating prior insights into the imaging modality or base models. This adaptability renders S2S2 particularly valuable in scenarios where domain-specific knowledge is unavailable. Furthermore, when such expertise is present, domain-specific strategies like SLAug exhibit superior generalization within their intended application domains, as indicated in Table 4. While domain-specific approaches are anticipated to excel, the supplementary application of S2S2 on top of SLAug still results in a marginal improvements on both the in-domain and out-of-domain performance. This result consolidates the relevance of S2S2, even in the presence of domain-specific methodologies.

Qualitative Evaluation. From Fig. 4, we observe similar ability to identify small objects and maintain boundary smoothness in the in-domain samples. Additionally, it is noteworthy that the base method is prone to misclassification issues in RGB images under conditions of significant glare (rows 6 and 8), the presence of unexpected objects (row 7), or insufficient lighting (rows 5 and 8). These conditions introduce what can be considered semantic noise. Our method, designed to mitigate semantic noise within the feature representation, remains robust and unaffected by such artifacts.

6 Ablation Study

In our ablation study, we aim to analyze the contribution of each module to performance, as well as the effect of the hyperparameters for the proposed loss. Our strategy to

Method	Kvasir-CVC	CVC-Kvasir	Mean
MSRF-Net	62.38	72.96	67.67
PraNet	79.12	79.50	79.31
SLAug	75.62	77.09	76.36
+S2S2	76.44	80.52	78.48+2.12
FCBFormer	91.16	86.46	88.81
+S2S2	**92.85**	**88.72**	**90.79**+1.98

Table 5: Out-of-domain performance comparison on Polyp segmentation (RGB medical image datasets). Dice score (%) is used as the evaluation metric.

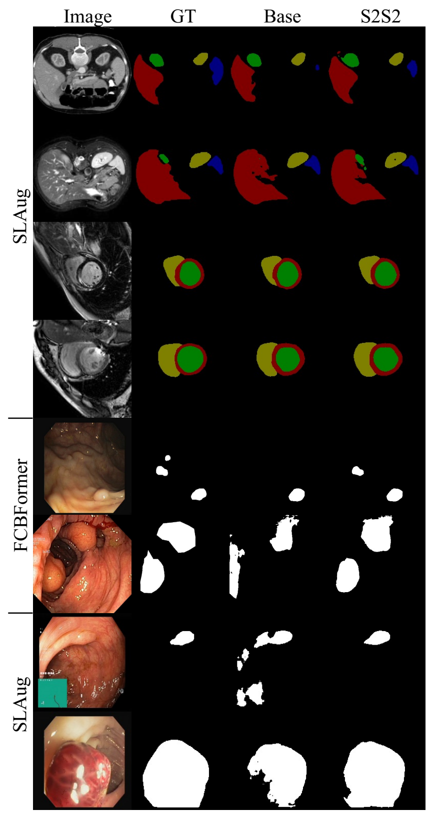

Figure 4: Visualization of the improvement achieved by applying S2S2 to the base method in the out-of-domain setting. 'GT' is the ground truth. 'Base' refers to the corresponding method without S2S2.

Synthetic	$\mathcal{L}^{enc}$	$\mathcal{L}^{dec}$	ACDC	Synapse
			88.86	76.86
✓			89.66+.80	77.61+.75
✓	✓		90.64+1.78	80.29+3.43
✓	✓	✓	90.40+1.54	81.19+4.33

Table 6: Performance of TransUNet using different proposed modules, measured in DSC (%). 'Synthetic' indicates the use of synthetic images. $\mathcal{L}^{enc}$ denotes the application of consistency loss on encoder features. $\mathcal{L}^{dec}$ denotes the application of consistency loss on decoder features.

accurately gauge the contributions of our module involves leveraging a baseline model that makes minimal assumptions and favors widespread adoption. For this purpose, we select TransUNet as the base model, adhering to its established training pipeline. The results of this investigation are detailed in Table 6. Employing solely synthetic images in the absence of semantic consistency loss yields a result comparable to the documented in prior works (Pinaya et al. 2022; Khader et al. 2023; Tang et al. 2023; Ye et al. 2023), with negligible improvements. The integration of semantic consistency loss $\mathcal{L}^{\text{enc}}$, however, marks a significant elevation in performance. Although the subsequent application of $\mathcal{L}^{\text{dec}}$, in conjunction with $\mathcal{L}^{\text{enc}}$, results in performance improvement on the Synapse dataset (with 9 classes), a marginal decline in performance is observed on the ACDC dataset (with 4 classes). This result is consistent with our earlier insight, indicating the superiority of the S2S2 method in datasets characterized by a greater number of classes. Moreover, the result suggests that the quality of generated images plays a vital role in the method's effectiveness. $\mathcal{L}^{\text{enc}}$ performs on a higher level semantic feature that is less sensitive to low-level detail of the generated images whereas $\mathcal{L}^{\text{dec}}$ operates on the pixel level that is very sensitive to the low-level detail. This dynamic is reflected in our experiments, wherein the inclusion of $\mathcal{L}^{\text{dec}}$ may potentially detract from out-domain performance. Nonetheless, the application of any form of semantic consistency loss invariably transcends the performance of the baseline model, underscoring the overall efficacy of the proposed S2S2 method.

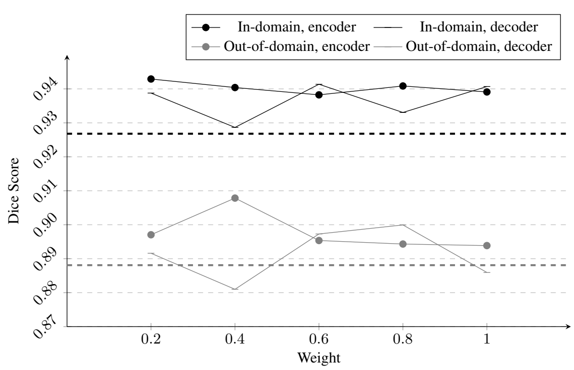

Figure 5: Ablation study results using FCBFormer with the proposed S2S2 method. Dashed lines indicate the performance of the base method.

To further investigate the impact of loss weighting on performance in both in-domain and out-of-domain contexts, we conducted an ablation study using FCBFormer on RGB images. We measured the Dice score on both in-domain and out-of-domain datasets, focusing on the effects of α^{enc} and α^{dec}. Each variable was analyzed in isolation by setting the alternative to zero for individual assessments. From the analysis presented in Fig. 5, it is observed that α^{enc} exerts a relatively consistent influence on in-domain performance, with the most notable improvement in out-domain performance is observed at $\alpha^{\text{enc}} = 0.4$. In contrast, the impact of α^{dec} appears less consistent, with the greatest fluctuations occurring within the range $\alpha^{\text{dec}} \in [0.2, 0.6]$ for both in-domain and out-of-domain datasets. This discrepancy in the behavior of losses on top of the encoder and decoder may stem from the generative model's capacity to more effectively capture higher-level semantic details as opposed to lower-level information, thereby rendering the encoder features more stable than those of the decoder, which aligns with our previous results. Moreover, the decoder features are subjected to additional layers of network weights, potentially amplifying errors inherent within the network architecture. This result suggests a preference for $\mathcal{L}^{\text{enc}}$ over $\mathcal{L}^{\text{dec}}$, attributed to its reduced sensitivity to variations in image quality. Despite the distinct behaviors observed, both semantic consistency losses contribute to the overall enhancement in model performance. Finally, if we apply the semantic consistency loss with only photometric augmentation such as Gaussian blur and color jitters, we get worse performance than the base method (detailed in the Appendix). This result further suggests the importance of the semantic stacking in addition to traditional augmentation.

7 Discussion and Conclusion

We introduce S2S2, a novel and broadly applicable add-on training strategy inspired by the image stacking technique, designed to improve both in-domain performance and out-of-domain robustness. However, the practical application of S2S2 encounters certain constraints. Primarily, the method's reliance on a fine-tuned generative model for semantic stacking, while innovative, introduces computational demands that may limit its suitability for situations with abundant data, such as natural image segmentation tasks. Additionally, the performance of S2S2 is inherently tied to the generative model's effectiveness across various datasets, which could significantly influence outcomes.

In conclusion, our findings present a compelling case for S2S2 as a powerful complement to existing domain-specific augmentation methods and architectural modifications. This strategy not only enhances model robustness but also represents a meaningful step toward the development of universally applicable solutions in image segmentation.

Acknowledgements

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIH) under award R01EB030130. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work used cluster computers at the National Center for Supercomputing Applications through an allocation from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation (NSF) grants 2138259, 2138286, 2138307, 2137603, and 2138296. The work also used the Extreme Science and Engineering Discovery Environment (XSEDE) under NSF grant 1548562.

References

Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D. S.; Karthikesalingam, A.; King, D.; Ashrafian, H.; and Darzi, A. 2021. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digital Medicine*, 4(1): 65.

Beers, A.; Brown, J.; Chang, K.; Campbell, J. P.; Ostmo, S.; Chiang, M. F.; and Kalpathy-Cramer, J. 2018. High-resolution medical image synthesis using progressively grown generative adversarial networks. *arXiv preprint arXiv:1805.03144*.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43: 99–111.

Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.-A.; Cetin, I.; Lekadir, K.; Camara, O.; Ballester, M. A. G.; et al. 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11): 2514–2525.

Cai, Y.; Fan, L.; and Fang, Y. 2023. SBSS: Stacking-based semantic segmentation framework for very high-resolution remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2229–2238.

Chai, L.; Wang, Z.; Chen, J.; Zhang, G.; Alsaadi, F. E.; Alsaadi, F. E.; and Liu, Q. 2022. Synthetic augmentation for semantic segmentation of class imbalanced biomedical images: A data pair generative adversarial network approach. *Computers in Biology and Medicine*, 150: 105985.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. AutoAugment: Learning augmentation policies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. RandAugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition Workshops*, 702–703.

Denton, E. L.; Chintala, S.; Fergus, R.; et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in Neural Information Processing Systems*, 28.

Deo, Y.; Dou, H.; Ravikumar, N.; Frangi, A. F.; and Lassila, T. 2023. Shape-guided conditional latent diffusion models for synthesising brain vasculature. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 164–173. Springer.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; and Dean, J. 2019. A guide to deep learning in healthcare. *Nature Medicine*, 25(1): 24–29.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. PraNet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 263–273. Springer.

Fernandez, V.; Pinaya, W. H. L.; Borges, P.; Tudosiu, P.-D.; Graham, M. S.; Vercauteren, T.; and Cardoso, M. J. 2022. Can segmentation models be trained with fully synthetically generated data? In *International Workshop on Simulation and Synthesis in Medical Imaging*, 79–90. Springer.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Guo, X.; Liu, J.; and Yuan, Y. 2024. Infproto-Powered Adaptive Classifier and Agnostic Feature Learning for Single Domain Generalization in Medical Images. *International Journal of Computer Vision*, 1–24.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 770–778.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hu, S.; Liao, Z.; and Xia, Y. 2023. Devil is in Channels: Contrastive Single Domain Generalization for Medical Image Segmentation. *arXiv preprint arXiv:2306.05254*.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, 124–140. Springer.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2): 203–211.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; Lange, T. d.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 451–462. Springer.

Kavur, A. E.; Gezer, N. S.; Barış, M.; Aslan, S.; Conze, P.-H.; Groza, V.; Pham, D. D.; Chatterjee, S.; Ernst, P.; Özkan, S.; et al. 2021. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69: 101950.

Kazerouni, A.; Aghdam, E. K.; Heidari, M.; Azad, R.; Fayyaz, M.; Hacihaliloglu, I.; and Merhof, D. 2023. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88: 102846.

Khader, F.; Müller-Franzes, G.; Tayebi Arasteh, S.; Han, T.; Haarburger, C.; Schulze-Hagen, M.; Schad, P.; Engelhardt, S.; Baeßler, B.; Foersch, S.; et al. 2023. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, 13(1): 7303.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123: 32–73.

Liao, S.; Peng, T.; Chen, H.; Lin, T.; Zhu, W.; Shi, F.; Chen, X.; and Xiang, D. 2024. Dual-Spatial Domain Generalization for Fundus Lesion Segmentation in Unseen Manufacturer's OCT Images. *IEEE Transactions on Biomedical Engineering*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Lyu, J.; Zhang, Y.; Huang, Y.; Lin, L.; Cheng, P.; and Tang, X. 2022. AADG: automatic augmentation for domain generalization on retinal image segmentation. *IEEE Transactions on Medical Imaging*, 41(12): 3699–3711.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, 565–571. Ieee.

Nguyen, D. M. H.; Pham, T. N.; Diep, N. T.; Phan, N. Q.; Pham, Q.; Tong, V.; Nguyen, B. T.; Le, N. H.; Ho, N.; Xie, P.; et al. 2023. On the Out of Distribution Robustness of Foundation Models in Medical Image Segmentation. *arXiv preprint arXiv:2311.11096*.

Ouyang, C.; Chen, C.; Li, S.; Li, Z.; Qin, C.; Bai, W.; and Rueckert, D. 2022. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4): 1095–1106.

Ozbulak, U.; Van Messem, A.; and De Neve, W. 2019. Impact of adversarial examples on deep learning models for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 300–308. Springer.

Pan, Y.; Zhang, S.; Gernand, A. D.; Goldstein, J. A.; and Wang, J. Z. 2023. AI-SAM: Automatic and Interactive Segment Anything Model. *arXiv preprint arXiv:2312.03119*.

Peng, W.; Adeli, E.; Bosschieter, T.; Park, S. H.; Zhao, Q.; and Pohl, K. M. 2023. Generating realistic brain mris via a conditional diffusion probabilistic model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 14–24. Springer.

Perez, F.; Vasconcelos, C.; Avila, S.; and Valle, E. 2018. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018*, 303–311. Springer.

Pinaya, W. H.; Tudosiu, P.-D.; Dafflon, J.; Da Costa, P. F.; Fernandez, V.; Nachev, P.; Ourselin, S.; and Cardoso, M. J. 2022. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, 117–126. Springer.

Qin, T.; Wang, Z.; He, K.; Shi, Y.; Gao, Y.; and Shen, D. 2020. Automatic data augmentation via deep reinforcement learning for effective kidney tumor segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1419–1423. IEEE.

Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A. I.; Etmann, C.; McCague, C.; Beer, L.; et al. 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3): 199–217.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 234–241. Springer.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252.

Sanderson, E.; and Matuszewski, B. J. 2022. FCN-transformer feature fusion for polyp segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, 892–907. Springer.

Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; and Rueckert, D. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53: 197–207.

Seyyed-Kalantari, L.; Zhang, H.; McDermott, M. B.; Chen, I. Y.; and Ghassemi, M. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12): 2176–2182.

Srivastava, A.; Jha, D.; Chanda, S.; Pal, U.; Johansen, H. D.; Johansen, D.; Riegler, M. A.; Ali, S.; and Halvorsen, P. 2021. MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(5): 2252–2263.

Su, Z.; Yao, K.; Yang, X.; Huang, K.; Wang, Q.; and Sun, J. 2023. Rethinking data augmentation for single-source domain generalization in medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(2), 2366–2374.

Tang, F.; Ding, J.; Wang, L.; Xian, M.; and Ning, C. 2023. Multi-Level Global Context Cross Consistency Model for Semi-Supervised Ultrasound Image Segmentation with Diffusion Model. *arXiv preprint arXiv:2305.09447*.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Wightman, R.; Touvron, H.; and Jégou, H. 2021. ResNet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16133–16142.

Xu, J.; Li, M.; and Zhu, Z. 2020. Automatic data augmentation for 3D medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 378–387. Springer.

Xu, Z.; Liu, D.; Yang, J.; Raffel, C.; and Niethammer, M. 2021. Robust and Generalizable Visual Representation Learning via Random Convolutions. In *International Conference on Learning Representations*.

Yang, D.; Roth, H.; Xu, Z.; Milletari, F.; Zhang, L.; and Xu, D. 2019. Searching learning strategy with reinforcement learning for 3D medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 3–11. Springer.

Ye, J.; Ni, H.; Jin, P.; Huang, S. X.; and Xue, Y. 2023. Synthetic augmentation with large-scale unconditional pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 754–764. Springer.

Yi, X.; Walia, E.; and Babyn, P. 2019. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58: 101552.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6023–6032.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022a. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain Generalization with MixStyle. In *International Conference on Learning Representations*.

Zhou, Z.; Qi, L.; Yang, X.; Ni, D.; and Shi, Y. 2022b. Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20856–20865.

Zhuang, X.; Xu, J.; Luo, X.; Chen, C.; Ouyang, C.; Rueckert, D.; Campello, V. M.; Lekadir, K.; Vesal, S.; RaviKumar, N.; et al. 2022. Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge. *Medical Image Analysis*, 81: 102528.