

Annotation Is (Almost) Indispensable: The Limitations of Unpaired Domain Translation and Semi-Supervised Approaches to Ultrasound Simplification

Precious Kolawole^a and Matthew S. Holden^a

^aCarleton University, School of Computer Science, Ottawa, Canada

ABSTRACT

We systematically evaluate whether ultrasound segmentation-based simplification can succeed with minimal or no paired data, to directly address the annotation bottleneck in clinical settings. Using the Open Kidney Ultrasound Dataset, we benchmark fully supervised, unpaired CycleGAN, and semi-supervised hybrid approaches across both binary (kidney/background) and multiclass tasks, with ablation studies sampling paired examples from 5 to over 90. Our results show that unpaired methods perform poorly (Dice: 0.058 binary, 0.043 multiclass), underscoring fundamental limitations of CycleGAN for ultrasound simplification without paired data. Fully supervised models achieve high accuracy with sufficient annotation (Dice: 0.900 binary, 0.627 multiclass) but rapidly degrade with less paired data. Semi-supervised hybrids bridge this gap, with binary segmentation reaching Dice 0.464 at just 5 pairs and 0.842 with 93 pairs, nearly matching supervised results. These findings establish annotation thresholds for effective ultrasound simplification and demonstrate that even sparse expert annotation dramatically outperforms purely unpaired approaches, providing actionable guidance for implementing ultrasound simplification under real-world constraints.

Keywords: ultrasound segmentation, semi-supervised learning, CycleGAN, annotation bottleneck, deep learning in ultrasound, medical image annotation, unpaired image translation, binary vs. multiclass segmentation.

1. INTRODUCTION

Ultrasound imaging provides radiation-free visualization in real time, which is crucial for clinical diagnosis and intervention; however, the gap between what ultrasound offers and what clinicians and researchers can extract from it limits its transformative impact in automated analysis.¹ A fundamental challenge in medical imaging is to simplify ultrasound scans into interpretable representations without requiring extensive manual annotation. Segmentation - delineating anatomical structures from background - serves as one approach to ultrasound simplification, but traditional methods require substantial paired training data (images with corresponding expert-drawn masks). This annotation bottleneck is significant in resource-limited environments.² The question thus emerges: can we effectively simplify ultrasound through segmentation without paired annotation?

While fully supervised segmentation represents the gold standard in medical image analysis, with U-Net variants^{3,4} demonstrating impressive performance across anatomical targets, their reliance on extensive paired data limits practical application. Unpaired translation approaches like CycleGAN⁵ have shown promise in other domains, with applications to cardiac ultrasound⁶ and shape-constrained adaptations,⁷ but their performance on complex ultrasound structures remains questionable. Semi-supervised methods that incorporate consistency regularization⁸ and adversarial training⁹ represent a middle ground, although the minimum effective annotation remains unquantified.

In this work, we systematically evaluate whether ultrasound simplification through segmentation can be accomplished as an unpaired task or with minimal annotation. We benchmark three paradigms, fully supervised models, unpaired CycleGAN translation, and semisupervised hybrids, across both binary and multiclass segmentation tasks. Our primary objectives are to: (1) determine whether unpaired approaches can produce clinically useful segmentations; (2) quantify the minimum annotation required for viable simplification; and (3) determine how segmentation complexity fundamentally reshapes these requirements. These findings will provide evidence-based guidance for implementing ultrasound simplification systems under real-world annotation constraints, potentially transforming clinical accessibility across resource-limited environments.

Table 1: Summary of the experimental paradigms and their architectural features.

Paradigm	Data Regime	Task Type	Model/Core	Data Flow	Losses Used
Fully Supervised (Paired)	All paired and N-paired (5,10,25,50,93)	Multi-class	Unet++ (ResNet101/34)	RGB $\rightarrow$ Mask	Dice Loss
Fully Supervised (Paired)	All paired and N-paired (5,10,25,50,93)	Binary	Unet++ (ResNet101/34)	RGB $\rightarrow$ Mask	Dice Loss
Unpaired CycleGAN	Unpaired A and B Domains	Multi-class	CycleGAN w/Unet++ gen. + PatchGAN disc.	Grayscale $\rightarrow$ Mask, Mask $\rightarrow$ Grayscale	Adversarial, Cycle-consistency
Unpaired CycleGAN	Unpaired A and B Domains	Binary	CycleGAN w/Unet++ gen. + PatchGAN disc.	Grayscale $\rightarrow$ Mask, Mask $\rightarrow$ Grayscale	Adversarial, Cycle-consistency, Identity Losses
Semi-Supervised Hybrid CycleGAN	All unpaired data + N-paired (5,10,25,50,93)	Multi-class	CycleGAN w/Unet++ gen. + PatchGAN disc.	Unpaired on all, paired seg. of N	Adversarial, Cycle-consistency, Segmentation Losses
Semi-Supervised Hybrid CycleGAN	All unpaired data + N-paired (5,10,25,50,93)	Binary	CycleGAN w/Unet++ gen. + PatchGAN disc.	Unpaired on all, paired seg. of N	Adversarial, Cycle-consistency, Identity, Segmentation Losses

2. METHODS

2.1 Dataset

We used the Open Kidney Ultrasound Dataset.¹⁰ It contains 534 2D B-mode abdominal ultrasound images and corresponding segmentation masks in PNG format. Images were acquired in real-world clinical settings between 2015 and 2019, with subjects including adults with chronic kidney disease, prospective donors, and transplant recipients. Two experienced sonographers independently annotated four anatomical kidney regions for each image: cortex, capsule, medulla, and central echo complex (comprising calyces, renal sinus, and related central structures).

2.2 Simplification Pipeline

At a high level, our approach explores three major regimes: fully supervised learning, unpaired (CycleGAN-based) learning, and semi-supervised (hybrid) learning. Each regime was partitioned according to segmentation complexity (multiclass - four anatomical regions plus background, or binary - kidney versus background), and, crucially, by the amount of annotated (paired) data available. The fully supervised track benchmarks the upper bound of segmentation under ideal annotation, training with paired image-mask samples. The unpaired CycleGAN track addresses the practical challenge of unlabeled data by simulating real-world labeling constraints, where masks and images may not be paired. Finally, the hybrid track bridges the two paradigms, utilizing limited ground truth amidst abundant unlabeled data. The key characteristics and distinctions of each architecture are summarized in Table 1.

As pre-processing, each annotation was converted from a polygon to a corresponding pixel-wise mask image. To optimize information content and network focus, every image and mask was center-cropped (retaining 70% of the region) and resized to 256×256 pixels. We normalized images according to model requirements ($[0,1]$ for supervised models and $[-1,1]$ for GANs), and masks were encoded using the full set of class labels or binarized for binary tasks. To rigorously control information leakage, we split the curated dataset into two non-overlapping halves: **Set A** and **Set B**. Each set contained 267 unique image-mask pairs, divided into subsets of 70% training, 15% validation, and 15% test. This partitioning was crucial for our CycleGAN training, which used images from Set A and masks from Set B, thereby preserving a truly unpaired regime. For all evaluations, only paired images and masks (from Set A validation and test) were used. Image data augmentation was done through random flips, rotations, and blurring. Only deterministic resizing and cropping were applied during validation and testing.

2.3 Segmentation Approaches

For the fully supervised baseline, we leveraged U-NET++ architectures with ResNet-101 encoders (pretrained on ImageNet), optimized for pixel-level mask prediction. All available image-mask pairs were used to train models for both multiclass (five classes) and binary (foreground/background) segmentation. Training was conducted with a weighted Dice loss to address class imbalance. To evaluate segmentation performance with just the very limited paired data, we conducted N -paired ablation studies: only N randomly sampled pairs ($N = 5, 10, 25, 50, 93$) from the training data were used to compute performance as a function of training set size.

CycleGAN frameworks were chosen for the strictly unpaired scenario, in which all training images (from Set A) and non-corresponding masks (from Set B) were used, explicitly eliminating spatial correspondence. Our implementation included two generator–discriminator pairs, each comprising U-NET++ generators and Patch-GAN discriminators to enforce realism. The networks learned both US→mask and mask→US mappings (with softmax mask predictions for multiclass tasks), using adversarial loss (hinge or LSGAN-formulated) together with an L1 cycle-consistency loss to enforce reversibility in the mapping. For binary segmentation, we additionally implemented an identity loss to further stabilize and regularize domain translation. Identity loss could not be applied for multiclass segmentation due to fundamental domain incompatibilities.

To investigate real-world scenarios where only a limited pool of paired masks was available, we developed a hybrid semi-supervised CycleGAN setup. Here, N random paired samples (varying from 5 to 93) were made available. In each training batch, these paired samples were included alongside unpaired ones, and a cross-entropy segmentation loss was computed on them in addition to the standard adversarial and cycle-consistency terms. The cycle-consistency and adversarial losses still operated on the full unpaired batch (A images and B masks), but the model was “anchored” by the supervised signal from a small number of paired examples. We chose this hybrid approach, as it directly addresses our core research question through CycleGAN’s domain translation capabilities while incorporating varying quantities of direct supervision. This setting directly probed whether the incorporation of just a few annotations could meaningfully boost segmentation, especially when all other data remained unpaired.

2.4 Implementation and Evaluation

All models and data loaders were implemented in PyTorch using the `segmentation-models-pytorch` library, trained for 40 epochs using the Adam optimizer ($\text{lr} = 1 \times 10^{-4}$), with validation and testing using paired data for fair comparison. Augmentation and normalization schemes were the same across approaches. Early stopping and model selection were based on validation Dice score. Segmentation quality was measured using Dice and Intersection-over-Union (IoU) scores, both globally (taken as a mean over foreground classes) and per class.

3. RESULTS

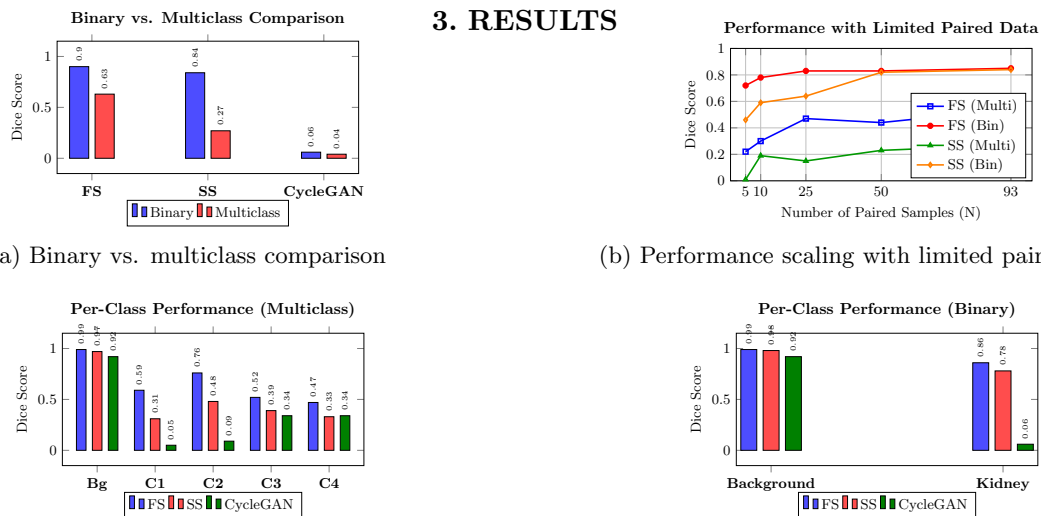


Figure 1: Performance analysis: (a) Binary vs. multiclass comparison; (b) Performance scaling with limited paired data; (c) Per-class multiclass performance (“all data”); (d) Per-class binary performance. (FS=Fully Supervised, SS=Semi-Supervised)

Table 2 summarizes Dice and IoU scores for all experimental configurations, spanning fully supervised, semi-supervised, and unpaired CycleGAN paradigms for both binary and multiclass segmentation. Performance is further detailed in Figure 1, which compares methods overall and per-class Dice scores.

The fully paired approach achieved the highest average Dice and IoU scores for both binary segmentation (Dice: 0.900, IoU: 0.820) and multiclass segmentation (Dice: 0.627 and IoU: 0.470). Segmentation Dice score for the background class was consistently highest (0.991), with individual foreground classes ranging from 0.470 (Class 4) to 0.762 (Class 2).

Table 2: Segmentation performance across approaches, complexities, and data regimes.

Method	Complexity	Dice Score / IoU Score					
		All Data	N=5	N=10	N=25	N=50	N=93
Fully Supervised	Binary	0.900 /0.820	0.725/0.573	0.777/0.638	0.832/0.720	0.825/0.716	0.854/0.762
	Multiclass	0.627/0.470	0.224/0.138	0.296/0.192	0.468/0.325	0.442/0.307	0.544/0.391
Semi-Supervised	Binary	-/-	0.464/0.319	0.586/0.420	0.643/0.481	0.818/0.698	0.842/0.732
	Multiclass	-/-	0.010/0.005	0.192/0.116	0.148/0.091	0.233/0.145	0.274/0.179
Unpaired CycleGAN	Binary	0.058/0.030	-/-	-/-	-/-	-/-	-/-
	Multiclass	0.043/0.024	-/-	-/-	-/-	-/-	-/-

The unpaired CycleGAN models performed substantially below the paired configurations (Table 2). In the multiclass regime, the average Dice was 0.043 and IoU was 0.024. The binary segmentation model attained a Dice of 0.058 and IoU of 0.030. Hybrid semi-supervised experiments, which supplemented CycleGAN training with a supervised loss on a small set of paired samples, outperformed the unpaired CycleGAN models and approached fully supervised performance as the number of paired samples increased (Table 2). For instance, in multiclass segmentation with $N = 93$ paired examples, the Dice score reached 0.274, while in binary segmentation it reached 0.842. Figure 1 illustrates the trade-off between annotation effort and segmentation performance for the fully paired and semi-supervised approaches. In both tasks, Dice scores improved with increasing paired sample counts. Binary segmentation required fewer paired examples to achieve clinically relevant Dice scores compared to multiclass segmentation. Figure 2 displays example segmentation results for fully supervised, semi-supervised and CycleGAN models in both multiclass and binary settings, in different quantities of paired samples (N).

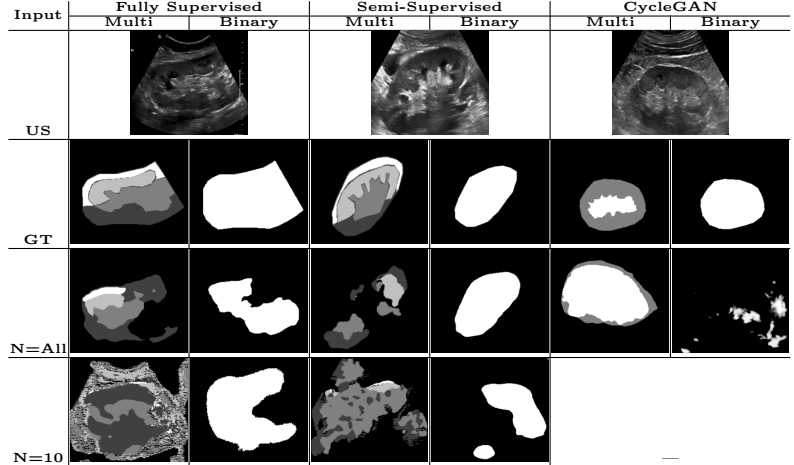


Figure 2: Qualitative segmentation results across approaches and training data sizes. For each approach, results for multi-class and binary segmentation are shown. Semi-supervised shows results at $N=93$ for the "All" variant.

4. DISCUSSIONS

Our systematic evaluation reveals clear patterns in the effectiveness of deep learning architectures across varying supervision levels and segmentation complexities. The fully supervised approach achieves clinical viability ($Dice > 0.7$) for binary tasks with a small number of samples ($N=5$), while multiclass segmentation requires substantially more paired examples to approach acceptable performance. In both the binary and multi-class tasks, poor performance is achieved with the unpaired approach ($Dice = 0.043$ for multi-class and $Dice = 0.058$ for binary), where the gap may be explained by the inability to implement identity losses in the multi-class task due to structural incompatibilities. These findings suggest that purely unpaired approaches face inherent limitations in ultrasound segmentation tasks without additional constraints or architectural innovations. While the results from the semi-supervised approach show improvement over unpaired results, they do not exceed the results of the fully supervised approach for multiclass. This suggests that using unpaired ultrasound images and masks has limited value in ultrasound segmentation.

Given the results of this work, we suggest using paired datasets for ultrasound segmentation and simplification whenever possible. We propose, in the future, this paired data can be collected using tracked ultrasound. Tracked ultrasound can be used to reconstruct a volumetric image of the relevant anatomy, and this image can be co-registered to a CT or MR image of the anatomy (alternatively, several fiducials could be localized in both coordinate frames for fiducial registration). Simplification (e.g., segmentation) can be performed once on the CT or MR image. Finally, the slice of the simplified 3D image corresponding to an arbitrary ultrasound image can be found using the transformation computed during registration. This is consistent with the approach recommended by China et al.,¹¹ but could be used for arbitrary simplification tasks.

ACKNOWLEDGMENTS

This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada grant RGPIN-2020-05582 and the Black and Indigenous Summer Research Internship (BISRI) program at Carleton University. This research was enabled in part by support provided by [Compute Ontario](#) and the [Digital Research Alliance of Canada](#).

REFERENCES

- [1] Dietrich, C., Goudie, A., Chiorean, L., Cui, X., Gilja, O., Dong, Y., Abramowicz, J., Vinayak, S., Westerway, S., Nolsøe, C., Chou, Y., and Blaivas, M., “Point of care ultrasound: A wfumb position paper,” *Ultrasound in Medicine and Biology* **43**, 49–58 (Jan. 2017). Includes bibliographical references.
- [2] Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D., and Lungren, M. P., “Preparing medical imaging data for machine learning,” *Radiology*, 192224 (2020).
- [3] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J., “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” in [*Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*], Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J. M. R. S., Bradley, A., Papa, J. P., Belagiannis, V., Nascimento, J. C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., and Madabhushi, A., eds., *Lecture Notes in Computer Science* **11045**, 3–11, Springer International Publishing, Cham (2018).
- [4] Oktay, O., Schlemper, J., Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N., Kainz, B., Glocker, B., and Rueckert, D., “Attention u-net: Learning where to look for the pancreas,” (04 2018).
- [5] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in [*2017 IEEE International Conference on Computer Vision (ICCV)*], 2242–2251 (2017).
- [6] Zhang, Z., Yang, L., and Zheng, Y., “Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network,” 9242–9251 (06 2018).
- [7] Chen, C., Dou, Q., Chen, H., Qin, J., and Heng, P., “Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation,” *IEEE Transactions on Medical Imaging* **PP**, 1–1 (02 2020).
- [8] Tarvainen, A. and Valpola, H., “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” (2018).
- [9] Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., and Srinivasan, B., “A generalized deep learning framework for whole-slide image segmentation and analysis,” *Scientific Reports* **11**, 11579 (06 2021).
- [10] Singla, R., Ringstrom, C., Hu, G., Lessoway, V., Reid, J., Ngan, C., and Rohling, R., “The open kidney ultrasound data set,” in [*International Workshop on Advances in Simplifying Medical Ultrasound (ASMUS)*], *Lecture Notes in Computer Science*, 155–164, Springer Nature Switzerland (2023).
- [11] China, D., MacLean, L. J., Wei, J., Theodore, N., Johnson, N., Crawford, N., Ding, K., and Uneri, A., “Automated labeling using tracked ultrasound imaging: Application in tracking vertebrae during spine surgery,” *Medical image analysis*, 103672 (2025).