Effective Disjoint Representational Learning for Anatomical Segmentation

Priya Tomar[*1,2] PRIYA.PRIYA@IAIS.FRAUNHOFER.DE
Aditya Parikh[*1] ADITYA.PARIKH@IAIS.FRAUNHOFER.DE
Philipp Feodorovici[3] PHILIPP.FEODOROVICI@UKBONN.DE
Jan Arensmeyer[3] JAN.ARENSMEYER@UKBONN.DE
Hanno Matthaei[3] HANNO.MATTHAEI@UKBONN.DE
Christian Bauckhage[1,2] CHRISTIAN.BAUCKHAGE@IAIS.FRAUNHOFER.DE
Helen Schneider[†1] HELEN.SCHNEIDER@IAIS.FRAUNHOFER.DE
Rafet Sifa[†1,2] RAFET.SIFA@IAIS.FRAUNHOFER.DE

[1] *Fraunhofer IAIS, Germany*
[2] *University of Bonn, Germany*
[3] *University Hospital Bonn, Germany*

Editors: Under Review for MIDL 2025

Abstract

In the wake of the limited availability of pertinent datasets, the application of computer vision methods for semantic segmentation of abdominal structures is mainly constrained to surgical instruments or organ-specific segmentations. Multi-organ segmentation has the potential to furnish supplementary assistance in multifarious domains in healthcare, for instance, robot-assisted laparoscopic surgery. However, in addition to the complexity involved in discriminating anatomical structures due to their visual attributes and operative conditions, the representation bias pertaining to organ size results in poor segmentation performance on organs with smaller pixel proportions. In this work, we focus on alleviating the influence of representation bias by involving different encoder-decoder frameworks for learning organ-specific features. In particular, we investigate the effect of organ-specific decoders on binary segmentation of anatomical structures in abdominal surgery. Additionally, we analyze the effect of organ-specific pretraining on the multi-label segmentation in two model training settings including knowledge sharing and disjoint learning, in relation to the contextual feature sharing between organ-specific decoders. Our results illustrate the significant gain in segmentation performance by incorporating organ-specific decoders, especially for less represented organs.

Keywords: Semantic segmentation, robot-assisted laparoscopic surgery, organ segmentation, Dresden surgical anatomy dataset.

1. Introduction

The applications of deep learning (DL) approaches have exhibited unforeseen advancements in medical image segmentation (Qureshi et al., 2023)(Wang et al., 2022). Nonetheless, the

[*] Contributed equally
[†] Contributed equally

limited availability of pertient datasets due to the restrictions involved in medical data processing and requirement of domain-specific expertise in annotating the intricate data features have restricted the developments to application-specific approaches. Medical imaging segmentation datasets predominantly focus on single-organ segmentation or specific anatomical regions (Rister et al., 2020), typically in computed tomography (CT) scans, magnetic resonance imaging (MRI) or X-Ray modalities (Fu et al., 2021)(Ji et al., 2022) (Schneider et al., 2023)(Häntze et al., 2024). In the domain of minimally invasive surgery, though datasets like EndoVis (Allan et al., 2020) have contributed to the significant developments on surgical instruments or organ-specific segmentation, datasets containing multi-organ annotations remain scarcely available (Carstens et al., 2023)(Hong et al., 2020). Furthermore, the surgical environment characteristic involving variable organ appearances, dynamic viewing conditions, frequent camera movements, and organ occlusions, present challenges in the extensive developments of DL-based methods (Rueckert et al., 2024).

For multi-organ segmentation in surgical data, current methods primarily focus on organ-specific learning. Kolbinger et al. (2023) investigated structure-specific models in comparison to common encoder and structure-specific decoder approach, reporting better performance for former. Similarly, Maack et al. (2024) focused on multi-teacher knowledge distillation (MT-KD) approach involving organ-specific decoders to learn through guided features of anatomy-specific teacher networks. Jenke et al. (2024) introduced an implicit learning method which emphasizes only the annotated classes in the images and considering missing ones as unknowns. It illustrated superior performance in comparison to organ-specific ensemble model but inferior when compared with full-supervised learning for multi-organ segmentation. Notwithstanding the improvements in organ-specific performance, these approaches face challenges due to the class imbalance. For instance, the involved Dresden surgical anatomy (DSA) dataset (Carstens et al., 2023) contains organ pixel percentages ranging from 27.32% for abdominal wall to 1.25% for intestinal veins which often leads to learning bias towards the highly represented organs, thereby neglecting smaller or less frequent structures.

To improve the performance for underrepresented or intricate classes in laparoscopic images, Sinha and Dolz (2021) proposed a multi-scale attention mechanism combined with semi-supervised learning and effectively leveraged a small set of labeled laparoscopic images alongside unlabeled data. Similarly, Qiu et al. (2022) introduced a class-wise confidence-aware active learning framework which focused on dynamically selecting informative samples and exploiting unlabeled data, achieving significant improvements. (Zhang et al., 2024) propose a method that integrates class-wise contrastive learning with multi-scale feature extraction. By leveraging classification labels and employing a multi-scale projection head, the model effectively learns representations even with limited annotated data.

Our work builds on recent advancements in surgical image segmentation, specifically addressing key challenges in laparoscopic imaging and organ-specific learning for multi-organ segmentation. Drawing motivation from (Kolbinger et al., 2023), we focus on tackling the complexities arising from anatomical variability and inter-organ relationships in surgical imaging datasets. The significant variance in pixel proportions across organs in our dataset, where seven organs comprise less than 7% of the relative pixel area, necessitates a careful examination of organ-specific feature learning approaches. Our primary contributions include: (1) a comprehensive evaluation of four different segmentation architectures,

including vanilla U-Net variants and state-of-the-art model, for binary segmentation of abdominal organs; (2) an in-depth analysis of three different training frameworks for organ-specific binary segmentation, investigating the influence of organ-specific feature learning in relation to model capacity; (3) examination of organ-specific transfer learning capabilities for multi-class segmentation focusing on the three organs with the highest pixel proportion; and (4) investigation of effective intra-organ feature knowledge sharing between organ-specific decoders, incorporating both knowledge-sharing and disjoint learning frameworks for multi-label segmentation.

The application of our proposed methodology extends beyond surgical organ segmentation to other medical imaging domains. The organ-specific feature learning and knowledge sharing frameworks could benefit areas such as brain tumor segmentation (Chen et al., 2023), cardiac chamber analysis (Zhang et al., 2023), and musculoskeletal imaging (Wang et al., 2023) where anatomical structures exhibit similar challenges of size variation and complex spatial relationships.

2. Methodology

2.1. Dataset

In our work, we utilize the DSA dataset (Carstens et al., 2023), a comprehensive, high-resolution dataset specifically curated for computer-aided surgical applications and machine learning approaches in medical imaging. The dataset comprises of 13,195 laparoscopic images, with a minimum of 20 surgeries documented for each organ. Binary segmentation masks are provided for eleven anatomical classes including abdominal wall, colon, inferior mesenteric artery, intestinal veins, liver, pancreas, small intestine, spleen, stomach, ureter, and vesicular glands. Additionally, the dataset includes a multi-class subset focusing on six abdominal organs (abdominal wall, colon, liver, pancreas, small intestine, and spleen). Furthermore, the dataset exhibits significant class imbalance, with organ pixel percentages ranging from 27.32% for abdominal wall to 1.25% for intestinal veins.

For robust evaluation, we focused on organs with foreground-to-background percentages exceeding 1% in the multi-class subset, specifically the abdominal wall, colon, and stomach (see Appendix Table 8). To ensure fair comparison with existing approaches, we adopted the training, validation, and testing split proposed by (Kolbinger et al., 2023). We resize the input to 256×256 and use standard data augmentation approaches including color jittering, random rotation, and image resizing to enhance model robustness and prevent overfitting.

2.2. Model Architecture

In this study, we investigated four backbone architectures: U-Net (Ronneberger et al., 2015), DeepLabv3 (Chen et al., 2017), Pyramid Vision Transformer (PVT) U-Net (Zhu et al., 2023) and a variant of U-Net (Oktay et al., 2018), referred to as Attention U-Net (AU-Net). The U-Net and AU-Net architectures follow a symmetric encoder-decoder structure with channel depths [64, 128, 256, 512]. In the AU-Net variant, we enhanced the standard U-Net by incorporating attention gates within the decoder pathway. These attention gates integrate gating signals from the decoder with skip connections from the encoder, enabling the network to focus on task-relevant spatial regions at each decoding stage. The

PVT U-Net hybrid architecture combines the hierarchical structure of U-Net with modern vision transformer concepts. This design builds upon the work of (Wang et al., 2021), adapting their pyramid vision transformer concept into a U-Net framework optimized for multi-organ segmentation.

We conducted an extensive comparative analysis of two distinct architectural approaches:

- This architecture follows the standard paradigm where a single encoder-decoder pipeline processes the input, with the decoder's output channels corresponding to individual organ segmentation masks (Fig. 1a). This approach leverages shared features across all organs while maintaining computational efficiency. To investigate whether performance differences stem from architectural design or model capacity, we implemented Expanded CECD (E-CECD), a variant with a 4× increase in decoder channel dimensions. This choice was motivated by prior work showing that increasing width (channel dimensions) rather than depth (number of layers) improves feature representation and training stability while maintaining computational efficiency (Zagoruyko and Komodakis, 2016)(Tan and Le, 2019). This controlled comparison allows us to differentiate between gains achieved through architectural design choices versus those stemming from increased parameter capacity.
- Common Encoder-Multiple Decoder (CEMD): This architecture utilizes a shared encoder but implements dedicated decoders for each of the eleven organs (Fig.1b). The shared encoder learns common representations of the abdominal anatomy, while individual decoders specialize in organ-specific learning. This architectural choice enables us to investigate whether organ-specific decoders provide advantages over the conventional single-decoder approach in terms of segmentation performance and feature learning.

This comparative appraisal aims to determine the necessity and effectiveness of organ-specific decoders versus the standard multi-class segmentation approach, particularly in the context of complex abdominal organ segmentation tasks.

2.3. Parameter Sharing Strategies

In this work, we furthermore investigate two parameter sharing strategies for fine-tuning the CEMD framework for multi-class organ segmentation. The first approach focuses on cumulative knowledge sharing, enabling simultaneous updates across all decoders during the fine-tuning process. This approach is motivated by the inherent anatomical relationships in abdominal imaging, where organs share common contextual features such as tissue interfaces, vascular patterns, and neighboring structures. The simultaneous update mechanism potentially allows decoders to leverage cross-organ anatomical knowledge, particularly beneficial for structures with similar imaging characteristics or adjacent anatomical relationships.

The second strategy employs disjoint learning, where each decoder is individually updated by preventing the parameter updates for other decoders. This approach maintains the specialized anatomical features learned during initial training, critical for preserving organ-specific boundary characteristics and morphological patterns.

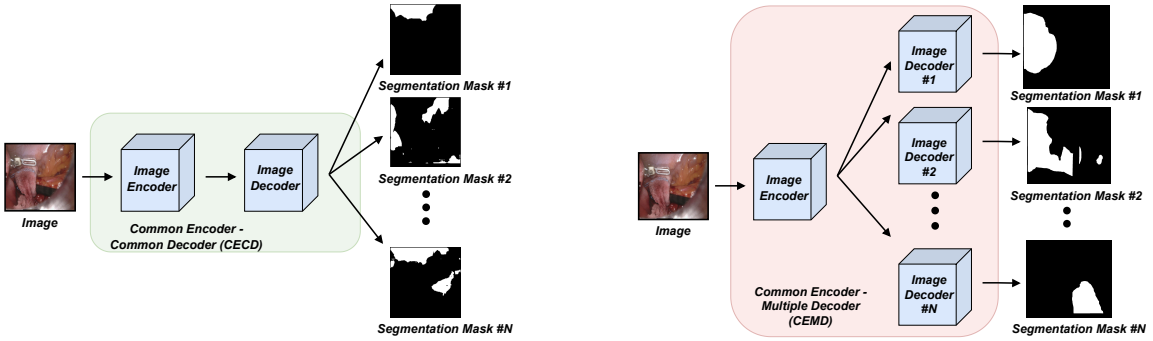

Figure 1: Comparison of two architectural approaches for organ segmentation. (a) CECD uses a single decoder to generate multi-channel outputs for multiple organ masks, while (b) CEMD employs multiple decoders, each dedicated to producing a specific organ mask. Both architectures share a common encoder for feature extraction.

These contrasting approaches address a key trade-off in medical image segmentation between preserving specialized features and leveraging shared anatomical knowledge. This distinction is particularly relevant for laparoscopic imaging, where varying viewpoints and tissue deformation benefit from both, specialized and shared feature learning.

2.4. Experiment Configurations and Evaluation

For all the experiments, we use a batch size of 4 samples, Adam optimizer with an initial learning rate of 1×10^{-3}, and a step-based learning rate scheduler, implementing a weight decay of 0.1 after every 50 epochs. Early stopping is implemented with a patience of 10 epochs, monitoring the validation DICE score to prevent overfitting. Each experiment runs for a maximum of 200 epochs, though convergence typically occurs earlier. Training is performed on an NVIDIA A100 GPU (40GB VRAM) using PyTorch framework.

To ensure reproducibility, we initialize all the models with random weights using fixed seed values. We use DICE loss for training as it exhibited superior performance in comparison to Binary Cross-Entropy (BCE) loss in our preliminary experiments (refer Appendix Table 6) and use it to select best-performing model checkpoint. For performance evaluation, we report the DICE Score (expressed in percentage for better interpretability) on held-out test set, as it provides an intuitive measure of spatial overlap while being robust to class imbalance, making it a standard metric in medical image segmentation tasks.

3. Results and Discussion

3.1. Comparison of Backbone Models

Our evaluation (Table 1) demonstrates that the AU-Net architecture consistently outperforms standard U-Net, DeepLabv3, and PVT U-Net models for most anatomical structures within the CEMD framework. AU-Net achieved superior performance in 9 out of 11 organs,

validating the effectiveness of attention mechanisms in medical image segmentation. The performance comparison between U-Net and AU-Net reveals that incorporating attention gates in the decoder pathway significantly improves segmentation accuracy. This improvement is particularly pronounced in organs with thin and clear anatomical boundaries with distinctive features (see Appendix Figure 3 for qualitative comparisons) . Additional analysis comparing IoU Scores in Appendix Table 4.

Table 1: Model Performance Comparison: DICE Scores (in %) (Mean ± Standard Deviation) across different backbone architectures. % indicates the foreground-to-background ratio for each organ in the training set. An additional table comparing the pixel % of the split in Appendix 8.

Organ	%	U-Net	DeepLabv3	PVT U-Net	AU-Net
Abdominal Wall	27.32	88.66 ±5.40	82.51 ±6.75	83.82 ±8.49	**88.73** ±5.73
Colon	11.07	73.98 ±12.36	61.12 ±15.64	60.12 ±17.23	**77.46** ±10.26
Inferior Artery	3.14	**48.54** ±11.68	36.42 ±14.46	38.87 ±14.55	47.95 ±14.20
Intestinal Veins	1.25	50.58 ±11.00	42.31 ±18.38	47.56 ±15.30	**58.83** ±15.64
Liver	24.59	**77.03** ±16.54	71.02 ±17.82	57.83 ±20.17	73.55 ±19.85
Pancreas	2.03	28.77 ±18.07	26.62 ±14.93	22.72 ±18.99	**29.80** ±18.68
Small Intestine	15.32	82.72 ±8.04	73.05 ±7.94	73.57 ±5.91	**85.38** ±7.04
Spleen	3.45	83.23 ±9.11	73.42 ±10.82	63.26 ±19.67	**87.58** ±6.98
Stomach	4.73	65.48 ±16.19	58.32 ±16.60	60.03 ±17.17	**71.38** ±16.57
Ureter	1.38	27.31 ±18.94	29.93 ±16.73	29.26 ±15.96	**50.07** ±19.24
Vesicular Glands	2.37	32.25 ±15.14	28.76 ±13.54	45.56 ±15.39	**47.27** ±15.71
Overall	-	60.78 ±23.09	53.04 ±20.77	51.29 ±18.70	**65.27** ±19.85

As anticipated, segmentation performance correlates with the organ pixel percentage (%) (foreground-to-background ratio). To investigate this, we computed Pearson (Kirch, 2008) and Spearman correlation coefficients (Spearman, 1904) between pixel % and DICE scores for each segmentation model. The results reveal a positive correlation between pixel % and DICE score for all models, with Pearson correlation coefficients of 0.64, 0.76, 0.48, and 0.69 for U-Net, DeepLabv3, PVT U-Net and, AU-Net, respectively (Appendix Fig 2). These findings highlight the influence of class imbalance on model performance, underscoring the inherent challenge of segmenting smaller or less frequently visible structures in laparoscopic imagery.

3.2. Comparison of Base Architectures

Following our comprehensive model evaluation, we focus on AU-Net for architectural variant comparisons considering its superior performance across organs. The evaluation (Table 2) reveals significant differences in both computational requirements and segmentation performance where the CEMD consistently outperformed CECD and its expanded variant E-CECD. While both E-CECD and CEMD have comparable parameter counts (152.83M and 156.83M respectively), their Giga Multiply-Accumulate Operations (GMACs) differ

substantially (482.65 vs 443.71). This disparity in computational efficiency arises from how parameters are distributed across the network. In CEMD, the dedicated decoders process only their respective organ's features, whereas E-CECD's enlarged decoder processes all features through the same computational path, leading to increased multiply-accumulate operations. For additional analysis refer Appendix Table 5.

Table 2: Architectural Comparison: DICE Scores (in %) (Mean ± Standard Deviation) with pixel % and computational requirements for CECD, E-CECD, and CEMD variants.

Organ	Pixel (%)	CECD	E-CECD	CEMD
Parameters (M)	-	31.39	152.83	156.83
GMACs	-	55.95	482.65	443.71
Abdominal Wall	27.32	85.30 ±6.79	86.39 ±6.52	**88.73** ±5.73
Colon	11.07	71.35 ±14.15	71.06 ±12.45	**77.46** ±10.26
Inferior Artery	3.14	36.42 ±18.20	42.61 ±15.68	**47.95** ±14.20
Intestinal Veins	1.25	56.56 ±19.65	**63.89** ±13.70	58.83 ±15.64
Liver	24.59	70.09 ±20.56	71.52 ±17.64	**73.55** ±19.85
Pancreas	2.03	**34.29** ±19.97	27.47 ±20.01	29.80 ±18.68
Small Intestine	15.32	82.48 ±8.81	81.63 ±7.42	**85.38** ±7.04
Spleen	3.45	84.28 ±7.89	79.06 ±13.38	**87.58** ±6.98
Stomach	5.36	64.90 ±17.35	66.93 ±13.97	**71.38** ±16.57
Ureter	1.38	46.27 ±17.81	43.17 ±17.10	**50.07** ±19.24
Vesicular Glands	2.37	**47.41** ±16.48	40.32 ±19.52	47.27 ±15.71
Overall	-	61.78 ±23.91	61.14 ±23.93	**65.27** ±19.85

Despite similar parameter counts between E-CECD and CEMD, the latter demonstrates superior performance for 8 out of 11 organs. It reflects the significant influence of architectural design in achieving superior outcomes and effectiveness in capturing organ-specific anatomical features in comparison to the sheer number of model parameters. Notably, CEMD shows particular strength in segmenting organs with lower pixel percentages, suggesting its enhanced capability in handling class imbalance through its specialized architectural design.

3.3. Evaluating Parameter Sharing Strategies

We further analyze the influence of transfer learning and two different parameter sharing approaches including Knowledge Sharing Fine-tuning (KS-FT) and Disjoint Fine-tuning (D-FT) for multi-class segmentation (introduced in Section 2.3). We consider AU-Net CEMD architecture as baseline and fine-tune it on a three-class organ subset subsuming abdominal wall, colon, and stomach.

Results (Table 3) demonstrate that D-FT consistently outperforms the baseline and KS-FT for all organs, achieving an overall improvement of 5.33% in DICE score after fine-tuning. It suggests that preserving organ-specific features during fine-tuning is crucial for

maintaining segmentation accuracy. While KS-FT showed improvements for the abdominal wall and stomach, its performance degraded significantly for colon segmentation, indicating that simultaneous parameter updates may lead to feature interference in complex anatomical structures.

Table 3: Fine-tuning Strategy Comparison: DICE and IoU scores (in %) for base model, Knowledge Sharing Fine-tuning (KS-FT), and Disjoint Fine-tuning (D-FT) on three-organ subset.

		DICE Score (in %)			IoU Score (in %)		
Organ	Pixel (%)	Base	KS-FT	D-FT	Base	KS-FT	D-FT
Abdominal Wall	7.69	56.17	62.64	**63.98**	47.19	55.76	**56.76**
Colon	5.60	33.38	21.42	**38.56**	25.76	15.16	**29.55**
Stomach	6.97	65.21	65.52	**66.22**	55.26	55.80	**56.59**
Overall	-	51.59	49.86	**56.92**	42.07	42.24	**47.63**

4. Conclusion

Our work presents several contributions to the field of multi-organ segmentation in laparoscopic surgery. Firstly, our evaluation of backbone architectures demonstrates that AU-Net consistently outperforms conventional architectures. The integration of attention gates in standard U-Net proved particularly effective for organs with distinct boundaries. Secondly, our investigation of parameter sharing strategies revealed that architectural design, rather than model capacity alone, plays a crucial role in segmentation performance. The CEMD architecture, despite having similar parameter counts to E-CECD, achieved superior results, suggesting that organ-specific decoders are better suited for capturing unique anatomical features in surgical organ images. Our analysis of parameter sharing approaches for fine-tuning indicates the importance of preserving specialized features in transfer learning, particularly for complex anatomical structures.

Future work should address the persistent challenge of class imbalance between organ representations and segmentation performance. Additionally, investigating more sophisticated attention mechanisms and developing strategies for efficient knowledge transfer between organ-specific decoders could further improve performance for less-represented anatomical structures. Finally, exploring techniques such as Parameter-Efficient Fine-Tuning (PEFT) methods, including low-rank adaptations (LoRA), or pruning strategies, could be explored to optimize the architecture. These approaches have the potential to maintain high segmentation accuracy while significantly lowering computational requirements, making the model more practical for deployment in resource-constrained settings.

Acknowledgments

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, LAMARR22B.

Additional Results

Table 4: Model Performance Comparison: IOU Scores (in %) (Mean ± Standard Deviation) across different backbone architectures. % indicates the foreground-to-background ratio for each organ in the test set.

Organ	%	U-Net	DeepLabv3	PVT U-Net	AU-Net
Abdominal Wall	27.32	80.02 ±8.21	70.73 ±8.95	72.96 ±11.30	**80.19** ±8.80
Colon	11.07	60.14 ±14.63	45.70 ±15.22	45.07 ±17.17	**64.27** ±12.82
Inferior Artery	3.14	**32.80** ±9.87	23.19 ±10.60	25.17 ±11.67	32.64 ±11.85
Intestinal Veins	1.25	44.32 ±11.15	28.49 ±14.37	32.50 ±13.09	**48.30** ±14.81
Liver	24.59	**65.14** ±18.69	57.83 ±17.82	42.56 ±11.77	61.66 ±22.51
Pancreas	2.03	18.25 ±13.86	16.25 ±10.52	14.20 ±12.96	**19.10** ±14.60
Small Intestine	15.32	71.29 ±11.18	58.14 ±9.58	58.53 ±7.38	**75.11** ±10.25
Spleen	3.45	72.26 ±12.62	59.11 ±12.97	49.12 ±20.07	**78.56** ±10.54
Stomach	4.73	50.77 ±17.53	42.94 ±15.35	44.97 ±17.16	**57.92** ±19.01
Ureter	1.38	17.32 ±13.80	18.75 ±11.73	18.25 ±11.98	**35.46** ±16.29
Vesicular Glands	2.37	20.26 ±11.54	17.55 ±9.56	30.80 ±13.18	**32.34** ±13.60
Overall	-	48.42 ±23.33	39.88 ±19.83	37.92 ±17.96	**52.78** ±21.21

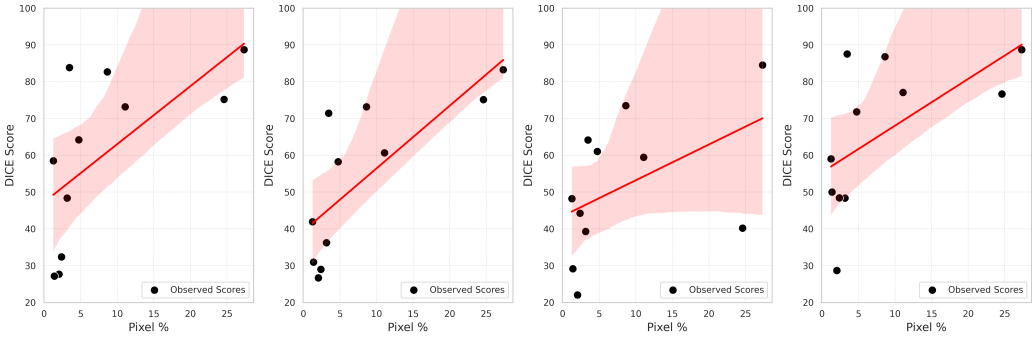

Figure 2: Scatter plots showing the correlation between pixel % and DICE scores (in %) for U-Net, DeepLabv3, PVT U-Net, and AU-Net. Each plot includes a regression trendline and corresponding Pearson correlation coefficient.

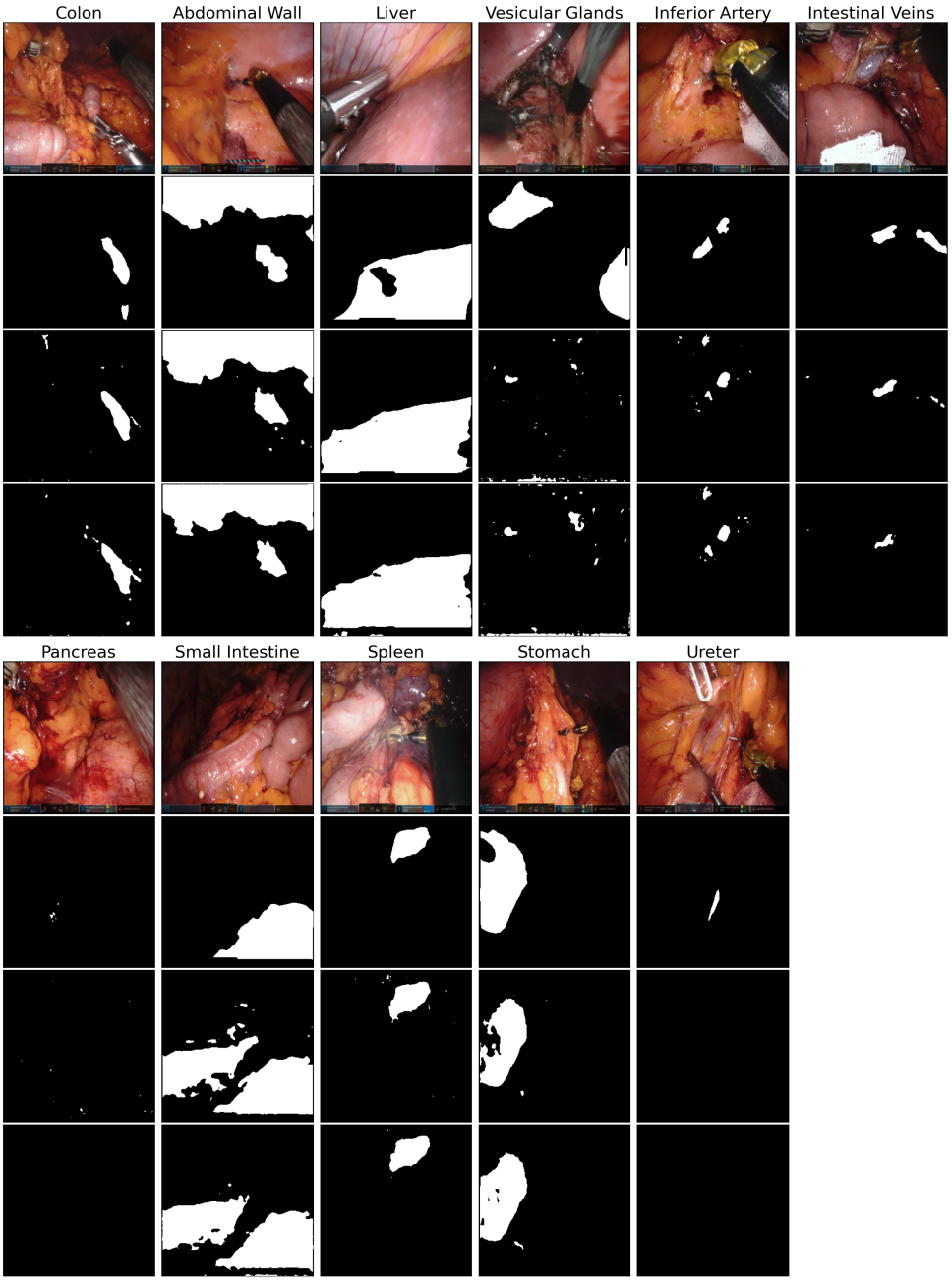

Figure 3: Qualitative comparison of segmentation results for 11 anatomical structures, comparing U-Net and AU-Net across two comprehensive visualizations. Each subplot includes (1) the original surgical image, (2) the ground truth segmentation mask, (3) U-Net predicted segmentation mask, and (4) AU-Net predicted segmentation mask. The first visualization presents results for the first six organs, while the second visualization shows results for the remaining five organs. AU-Net demonstrates superior boundary delineation and reduced false positives compared to U-Net, particularly in regions with complex anatomical interfaces.

Table 5: Architectural Comparison: IoU Scores (in %) (Mean ± Standard Deviation) and computational requirements for CECD, E-CECD, and CEMD variants.

Organ	Pixel %	CECD	E-CECD	CEMD
Parameters (M)	-	31.39	152.83	156.83
GMACs	-	55.95	482.65	443.71
Abdominal Wall	27.32	74.95 ±9.83	76.59 ±9.46	**80.19** ±8.80
Colon	11.07	57.22 ±16.04	56.45 ±13.98	**64.27** ±12.82
Inferior Artery	3.14	23.84 ±14.25	28.36 ±12.97	**32.64** ±11.85
Intestinal Veins	1.25	41.77 ±17.22	**48.33** ±13.99	48.30 ±14.81
Liver	24.59	57.41 ±22.11	58.42 ±20.01	**61.66** ±22.51
Pancreas	2.03	**22.58** ±15.78	17.73 ±15.79	19.10 ±14.60
Small Intestine	15.32	71.07 ±12.02	69.62 ±10.42	**75.11** ±10.25
Spleen	3.45	73.62 ±11.60	67.21 ±16.86	**78.56** ±10.54
Stomach	6.58	50.40 ±18.62	51.95 ±15.81	**57.92** ±19.01
Ureter	1.38	31.87 ±15.44	29.00 ±13.61	**35.46** ±16.29
Vesicular Glands	2.37	**32.56** ±13.88	27.17 ±15.76	32.34 ±13.60
Overall	-	48.81 ±24.22	48.09 ±23.93	**52.78** ±21.21

Table 6: Training Loss Function Comparison: DICE vs BCE, evaluating using DICE and IOU scores (in %) (Mean ± Standard Deviation) using the AU-Net architecture. The DICE loss function was chosen for its ability to better handle the imbalanced class distribution commonly present in medical image segmentation tasks. It emphasizes the overlap between predicted and ground truth masks, which is crucial when segmenting small, irregularly shaped organs where pixel-wise accuracy is less informative. As shown in the table, the use of DICE loss results in higher DICE and IoU scores for most organs compared to BCE loss, particularly for organs with smaller or more challenging shapes. The overall performance of DICE loss (65.27 ± 19.65 for DICE Score % and 52.78 ± 21.21 for IoU Score %) also outperforms BCE loss (57.63 ± 23.87 for DICE Score % and 45.35 ± 23.71 for IoU Score %), highlighting its superior sensitivity to challenging segmentation tasks. This is particularly important for precise delineation in medical imaging, where even small misclassifications can have significant clinical implications.

Organ	DICE Score (%)		IoU Score (%)	
	DICE Loss	BCE Loss	DICE Loss	BCE Loss
Abdominal Wall	**88.73** ±5.73	87.90 ±5.78	**80.19** ±8.80	78.87 ±8.79
Colon	**77.46** ±10.26	72.56 ±12.81	**64.27** ±12.82	58.40 ±14.65
Inferior Artery	**47.95** ±14.20	40.75 ±13.75	**32.64** ±11.85	26.51 ±10.74
Intestinal Veins	**58.83** ±15.64	49.88 ±13.52	**43.30** ±14.81	34.28 ±11.75
Liver	73.55 ±19.85	**76.76** ±15.55	61.66 ±22.51	**64.60** ±18.43
Pancreas	**29.80** ±18.68	25.77 ±17.50	**19.10** ±14.60	16.06 ±12.71
Small Intestine	**85.38** ±7.04	82.41 ±7.11	**75.11** ±10.25	70.67 ±9.79
Spleen	**87.58** ±6.98	78.13 ±11.93	**78.56** ±10.54	65.57 ±15.01
Stomach	**71.38** ±16.57	63.82 ±16.79	**57.92** ±19.01	48.97 ±17.29
Ureter	**50.07** ±19.24	29.10 ±19.48	**35.46** ±16.29	18.61 ±13.93
Vesicular Glands	**47.27** ±15.71	26.86 ±14.32	**32.34** ±13.60	16.33 ±9.98
Overall	**65.27** ±19.65	57.63 ±23.87	**52.78** ±21.21	45.35 ±23.71

Table 7: Average Organ Ratios (foreground to background pixels) as percentages for Train, Test, and Validation Sets for binary segmentation dataset. The table highlights similar trends in pixel percentages across splits for each organ, demonstrating consistency in the data distribution.

Organ	Train Set (%)	Test Set (%)	Validation Set (%)
Abdominal Wall	27.32	22.34	27.74
Colon	11.07	14.04	11.75
Inferior Artery	3.14	2.19	2.51
Intestinal Veins	1.25	1.66	1.05
Liver	24.59	11.62	14.99
Pancreas	2.03	4.06	3.18
Small Intestine	15.32	15.99	16.26
Spleen	3.45	2.19	4.00
Stomach	4.73	5.69	5.21
Ureter	1.38	0.85	1.18
Vesicular Glands	2.37	3.34	3.83

Table 8: Average Organ Ratios (foreground to background pixels) as percentages for Train, Test, and Validation Sets for multi-class segmentation subset. The table highlights inconsistencies in organ pixel distribution across splits, with only organs having pixel percentages greater than 1% in all splits (Abdominal Wall, Colon and Stomach) considered for evaluation. This ensures robustness, avoids bias from underrepresented organs, and addresses the issue of poor model training and performance on organs with extremely low pixel percentages.

Organ	Train Set (%)	Test Set (%)	Validation Set (%)
Abdominal Wall	**13.99**	**7.67**	**19.29**
Colon	**2.79**	**5.60**	**1.59**
Liver	1.74	0.49	9.53
Pancreas	0.29	0.54	0.04
Small Intestine	0.46	0.20	0.41
Spleen	0.60	0.00	0.00
Stomach	**6.19**	**6.97**	**6.55**

References

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, et al. 2018 robotic scene segmentation challenge, 2020.

Matthias Carstens, Franziska Rinner, Sebastian Bodenstedt, Alexander Jenke, Jürgen Weitz, et al. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10, 01 2023. doi: 10.1038/s41597-022-01719-2.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

R. Chen, J. Liu, and S. Wang. Brain tumor segmentation using multi-decoder networks with region-specific feature learning. *Medical Image Analysis*, 84:102694, 2023.

Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. A review of deep learning based methods for medical image multi-organ segmentation. *Physica Medica*, 85:107–122, 2021.

W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

Hartmut Häntze, Lina Xu, Christian J. Mertens, Felix J. Dorfner, Leonhard Donle, et al. Mrsegmentator: Multi-modality segmentation of 40 classes in mri and ct, 2024.

Alexander C. Jenke, Sebastian Bodenstedt, Fiona R. Kolbinger, Marius Distler, Jürgen Weitz, and Stefanie Speidel. One model to use them all: Training a segmentation model with complementary datasets, 2024.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems*, 2022.

Wilhelm Kirch, editor. *Pearson's Correlation Coefficient*, pages 1090–1091. Springer Netherlands, Dordrecht, 2008. ISBN 978-1-4020-5614-7. doi: 10.1007/978-1-4020-5614-7_2569.

Fiona R. Kolbinger, Franziska M. Rinner, Alexander C. Jenke, Matthias Carstens, Stefan Leger, et al. Anatomy segmentation in laparoscopic surgery: Comparison of machine learning and human expertise. *medRxiv*, 2023. doi: 10.1101/2022.11.11.22282215.

Lennart Maack, Finn Behrendt, Debayan Bhattacharya, Sarah Latus, and Alexander Schlaefer. Efficient anatomy segmentation in laparoscopic surgery using multi-teacher knowledge distillation. In *Medical Imaging with Deep Learning*, 2024.

Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, et al. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018.

Jie Qiu, Yuichiro Hayashi, Masahiro Oda, Takayuki Kitasaka, and Kensaku Mori. Class-wise confidence-aware active learning for laparoscopic images segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 18:473–482, 2022.

Imran Qureshi, Junhua Yan, Qaisar Abbas, Kashif Shaheed, Awais Bin Riaz, Abdul Wahid, Muhammad Waseem Jan Khan, and Piotr Szczuko. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90:316–352, 2023.

Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):1–13, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

Tobias Rueckert, Daniel Rueckert, and Christoph Palm. Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art. *Computers in Biology and Medicine*, page 107929, 2024.

Helen Schneider, David Biesner, Akash Ashokan, Maximilian Broß, Rebecca Kador, et al. Segmentation and analysis of lumbar spine mri scans for vertebral body measurements. *ESANN 2023 proceesdings*, 2023.

Ashish Sinha and Jose Dolz. Multi-scale self-guided attention for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25:121–130, 2021. doi: 10.1109/JBHI.2020.2986926.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

M. Wang, K. Yang, and X. Chen. Orthopedic image segmentation with anatomical structure-specific feature extraction. *Computerized Medical Imaging and Graphics*, 105: 102156, 2023.

Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET image processing*, 16(5): 1243–1267, 2022.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *CoRR*, abs/2102.12122, 2021.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

L. Zhang, Y. Wu, and H. Li. Cardiac multi-structure segmentation: A knowledge-sharing framework for chamber analysis. *IEEE Transactions on Medical Imaging*, 42(5):1245–1256, 2023.

Luyang Zhang, Yuichiro Hayashi, Masahiro Oda, and Kensaku Mori. Towards better laparoscopic video segmentation: A class-wise contrastive learning approach with multi-scale feature extraction. *Healthcare Technology Letters*, 11:126–136, 2024. doi: 10.1049/htl2.12069.

Jinghua Zhu, Yue Sheng, Hui Cui, Jiquan Ma, Jijian Wang, et al. Cross pyramid transformer makes u-net stronger in medical image segmentation. *Biomedical Signal Processing and Control*, 86:105361, 2023. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2023.105361.