Evaluating BEFUnet in the Small-Data Regime: A Study on Kvasir-SEG with U-Net and Edge-Gated Fusion

Rahul B (22b3976)[1]

[1]Department of Electrical Engineering, IIT Bombay

Abstract

Medical image segmentation models are increasingly moving towards hybrid CNN–Transformer architectures with explicit boundary modeling. BEFUnet is one such architecture that combines an edge encoder, a transformer-based body encoder, and specialized fusion modules to improve boundary accuracy. While prior work demonstrates strong performance on relatively large datasets, the behavior of such high-capacity models in the small-data regime is not well understood. In this work, we systematically compare a U-Net baseline, a full BEFUnet variant with LCAF+DLF fusion, and a lightweight BEFUnet variant using Edge-Gated Fusion (EGF) on the Kvasir-SEG polyp segmentation dataset under an 80/10/10 train/validation/test split. We report performance in terms of Dice coefficient, Intersection over Union (IoU), and sensitivity, and additionally analyze training-time and convergence behavior. Our results show that, under limited data, U-Net achieves higher mean Dice and IoU than both BEFUnet variants, suggesting that the strong convolutional inductive bias and lower capacity of U-Net yield better generalization in the small-data regime. At the same time, our EGF-based BEFUnet achieves similar performance to the full LCAF+DLF fusion while training approximately twice as fast, making it a more practical fusion strategy when computational budget is constrained.

1 Introduction

Deep learning has become the dominant paradigm for medical image segmentation, enabling automatic delineation of anatomical structures and lesions in modalities such as colonoscopy, MRI, CT, and ultrasound. Accurate segmentation boundaries are essential for downstream tasks, including lesion size estimation, treatment planning, and disease monitoring. However, achieving both region-level accuracy and sharp boundary localization is challenging, particularly in the presence of low contrast, noise, and irregular object shapes.

U-Net and its variants remain the de facto standard for medical image segmentation, primarily due to their encoder–decoder architecture with skip connections, strong convolutional inductive bias, and robustness to relatively small datasets. Recent work on BEFUnet-type architectures augments this paradigm with additional edge encoders, Transformer-based body encoders, and sophisticated fusion modules to enhance boundary modeling. These architectures are usually evaluated under settings with sufficiently large and diverse training data, where their capacity can be fully exploited.

In many realistic medical scenarios, however, annotated datasets are small due to the cost of expert labeling and the difficulty of data sharing. In such *small-data* regimes, high-capacity models can overfit easily, and classical architectures like U-Net may still be competitive or even superior. This raises a natural question: *How does BEFUnet behave when trained on limited data, and does it still outperform U-Net?*

In this paper, we address this question through an empirical study on the Kvasir-SEG polyp segmentation dataset. We implement:

- a U-Net baseline,
- a full BEFUnet variant with Local Context Aggregation Fusion (LCAF) followed by Dual-Level Fusion (DLF),
- and a lightweight BEFUnet variant that replaces LCAF+DLF with our proposed Edge-Gated Fusion (EGF).

We specifically focus on the small-data regime: Kvasir-SEG is split into 80% training, 10% validation, and 10% test, corresponding to only 800 training images. We then compare the three models in terms of segmentation quality and efficiency.

Contributions

Our contributions are:

- We conduct a controlled comparison of U-Net and two BEFUnet-style architectures on Kvasir-SEG under a small-data setting.
- We introduce a lightweight Edge-Gated Fusion (EGF) module that replaces LCAF+DLF, significantly reducing fusion complexity while retaining boundary awareness.
- We show that U-Net achieves better mean Dice and IoU than both BEFUnet variants when trained on limited data, highlighting the importance of matching model capacity to dataset size.
- We demonstrate that BEFUnet with EGF achieves similar performance to the full LCAF+DLF fusion at roughly half the training time per batch, making it a practical choice when computational budget is limited.

2 Related Work

U-Net and variants. U-Net introduced a symmetric encoder–decoder architecture with skip connections that has become foundational for biomedical segmentation. Numerous variants extend this design with dense skip connections (UNet++), multi-scale aggregation (U-Net 3+), or attention mechanisms (Attention U-Net), but all retain the central CNN-based inductive bias.

Boundary-aware segmentation. Improving boundary quality has been explored via edge supervision, contour-based losses, and explicit boundary refinement modules. These methods seek to augment region-level predictions with edge-aware features that better capture fine structures and object contours.

Transformer-based medical segmentation. Hybrid CNN–Transformer models such as TransUNet and Swin-UNet use self-attention to capture long-range dependencies and global context. While powerful, these models often require larger datasets and more careful regularization to avoid overfitting.

BEFUnet-style architectures. BEFUnet-like models combine an edge encoder, a Transformer-based body encoder, and specialized fusion modules to explicitly model both boundaries and regions. Existing evaluations typically focus on relatively large datasets, whereas we study their behavior in a resource-constrained setting.

3 Method

3.1 Baseline U-Net

We adopt a standard 2D U-Net as baseline. The encoder consists of repeated `DoubleConv` blocks (two consecutive 3×3 convolutions with batch normalization and ReLU), followed by max-pooling for downsampling. The decoder mirrors the encoder with transposed convolutions for upsampling and skip connections that concatenate encoder features with decoder features at each scale. A final 1×1 convolution produces a single-channel segmentation logit map.

The U-Net serves as a strong baseline due to its:

- convolutional inductive bias suitable for small datasets,
- relatively modest parameter count,
- and stable training behavior.

3.2 BEFUnet Encoders

Edge encoder. The edge encoder is a lightweight CNN that extracts boundary-focused features at multiple scales. It consists of stacked `ConvBlock`s (two 3×3 convolutions with batch normalization and ReLU) with max-pooling between stages, producing feature maps E_1, E_2, E_3, E_4 at progressively lower resolutions and higher channel dimensions. These features are intended to capture local edges and fine spatial structures.

Body encoder (Transformer). The body encoder uses a Swin-Tiny model, instantiated via the `timm` library with `features_only` output. This encoder yields multi-scale feature maps B_1, B_2, B_3, B_4 corresponding to different stages in the transformer hierarchy. These features provide global context and robust region-level semantics.

3.3 Fusion Variants

We consider two fusion strategies between edge features E_i and body features B_i:

3.3.1 BEFUnet (LCAF+DLF): Full Fusion

The full BEFUnet variant employs a two-stage fusion mechanism:

- **Local Context Aggregation Fusion (LCAF)**: aligns and combines edge and body features using 3×3 convolutions and local context mixing, producing intermediate fused features.
- **Dual-Level Fusion (DLF)**: refines these fused features by combining additive and multiplicative interactions between edge and body representations, followed by additional convolutional refinement.

While effective, this LCAF+DLF stack is computationally heavy due to multiple convolutional layers and complex feature interactions, increasing both parameter count and training time.

3.3.2 Our Innovation: BEFUnet (EGF)

To reduce complexity and better suit the small-data regime, we propose a lightweight **Edge-Gated Fusion (EGF)** module, yielding the BEFUnet (EGF) variant.

Given edge features E_i and body features B_i at the same scale, EGF operates as follows:

1. Spatially align B_i to the resolution of E_i via bilinear interpolation.
2. Project E_i and B_i into a common channel dimension using 1×1 convolutions, producing E_i' and B_i'.
3. Concatenate E_i' and B_i' and feed them through a small gating network (1×1 convolutions followed by a sigmoid) to obtain a spatial gate $\alpha_i \in [0,1]^{H\times W}$.
4. Fuse features via convex combination:

$$F_i = \alpha_i \cdot E_i' + (1-\alpha_i)\cdot B_i'. \tag{1}$$

Intuitively, α_i learns to emphasize edge features where boundary cues are strong and to emphasize body features where region context is more reliable. Compared to LCAF+DLF, EGF:

- uses fewer convolutions and no multiplicative cross-interactions,
- reduces fusion overhead and parameter count,
- and empirically trains approximately twice as fast per batch.

3.4 Decoder

For both BEFUnet variants, the decoder receives fused features F_1, F_2, F_3, F_4 from the fusion modules. A U-shaped decoder progressively upsamples the deepest fused features and combines them with shallower fused features via skip connections, using 3×3 convolutions with batch normalization and ReLU. The final 1×1 convolution outputs a single-channel logit map, consistent with binary segmentation.

3.5 Loss and Metrics

All models are trained with a combination of binary cross-entropy (BCE) and Dice loss:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}. \tag{2}$$

We evaluate models using:

- Dice coefficient,
- Intersection over Union (IoU),
- Sensitivity (SE),
- Specificity (SP),
- Accuracy (ACC).

In the main tables, we focus on Dice, IoU, and SE as the most clinically relevant metrics, and report mean scores over the test set.

4 Dataset and Experimental Setup

4.1 Kvasir-SEG Dataset

We use the Kvasir-SEG dataset, which contains 1000 colonoscopy images with corresponding polyp segmentation masks. The images exhibit substantial variability in polyp size, shape, and appearance, and are thus suitable for evaluating boundary-aware segmentation.

We follow an 80/10/10 split:

- 80% for training (800 images),
- 10% for validation (100 images),
- 10% for testing (100 images).

This setup intentionally represents a small-data regime with only 800 labeled training samples.

4.2 Preprocessing and Augmentation

All images and masks are resized to 256×256 pixels. Masks are binarized using a fixed threshold. Data augmentation on the training set includes:

- random horizontal and vertical flips,
- random brightness and contrast adjustments.

These transformations are implemented using the Albumentations library.

4.3 Training Details

All models are trained for 40 epochs using:

- optimizer: AdamW,
- learning rate: 10^{-4},
- batch size: 16,
- loss: BCE+Dice,
- device: single GPU (e.g., NVIDIA T4 on Kaggle).

We measure training speed in terms of wall-clock time per batch of 16 images:

- U-Net: approximately 1 second per batch,
- BEFUnet (EGF): approximately 1 second per batch,
- BEFUnet (LCAF+DLF): approximately 2 seconds per batch.

Table 1: Mean segmentation performance on the Kvasir-SEG test set. U-Net outperforms both BEFUnet variants in the small-data regime. BEFUnet (EGF) achieves similar performance to BEFUnet (LCAF+DLF) while being faster to train.

Model	Dice ↑	IoU ↑	Sensitivity (SE) ↑
U-Net	0.822	0.727	0.826
BEFUnet (EGF) [ours]	0.761	0.651	0.812
BEFUnet (LCAF+DLF)	0.782	0.679	0.821

Table 2: Approximate training time per batch of 16 images. BEFUnet (EGF) matches U-Net in speed, whereas BEFUnet (LCAF+DLF) is approximately twice as slow.

Model	Time / batch (16 images)	Relative cost
U-Net	1 s	1×
BEFUnet (EGF) [ours]	1 s	1×
BEFUnet (LCAF+DLF)	2 s	2×

5 Results

5.1 Quantitative Metrics

Table 1 reports the mean Dice, IoU, and sensitivity scores on the test set for U-Net, BEFUnet (EGF), and BEFUnet (LCAF+DLF). All results are averaged over 100 test images.

Overall, U-Net achieves the best mean Dice and IoU scores, indicating superior generalization under limited training data. The full BEFUnet with LCAF+DLF performs slightly better than the EGF variant on average, but the differences are modest relative to the increased training cost.

5.2 Training-Time Comparison

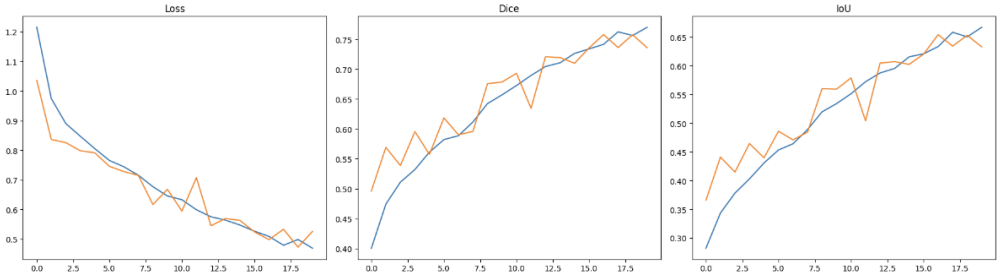
Table 2 summarizes the observed training time per batch for each model.

These results indicate that EGF provides a more favorable trade-off between performance and efficiency than LCAF+DLF in our setting.

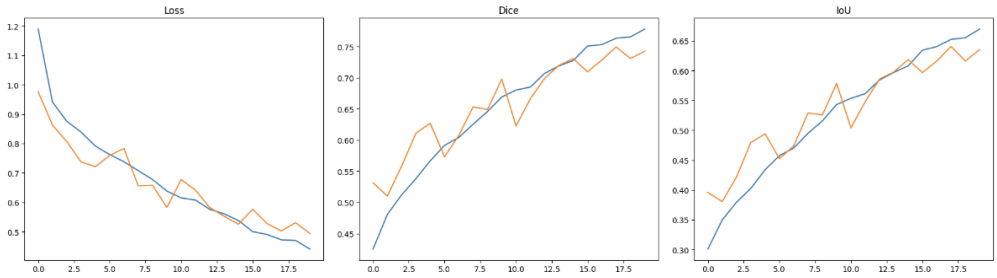
5.3 Training Curves

To analyze convergence behavior, we monitor training and validation loss, Dice, and IoU across epochs for all three models. Figure 1 illustrates the training dynamics. We observe that U-Net and BEFUnet (EGF) converge faster, reaching stable Dice and IoU values within the first 20 epochs, whereas BEFUnet (LCAF+DLF) requires more epochs to stabilize due to its higher capacity.

(a) U-Net: Loss, Dice, and IoU curves



(b) BEFUnet (LCAF+DLF): Loss, Dice, and IoU curves



(c) BEFUnet (EGF) [ours]: Loss, Dice, and IoU curves

Figure 1: Training and validation curves for U-Net (Row 1), BEFUnet (LCAF+DLF) (Row 2), and BEFUnet (EGF) [ours] (Row 3).

U-Net and BEFUnet (EGF) converge relatively quickly, achieving stable performance by ∼20 epochs, whereas the full BEFUnet (LCAF+DLF) requires more iterations to stabilize due to its higher capacity. After 20 epochs both U-Net and BEFUnet (EGF) didnt show any improvements in validation accuracy but BEFUnet (LCAF+DLF) showed slow improvements from start.

5.4 Distribution Analysis via Violin Plots

To go beyond mean metrics, we visualize the distribution of Dice, IoU, and sensitivity across the test set using violin plots (Figure 2). These plots reveal the variability and presence of outliers for each model.

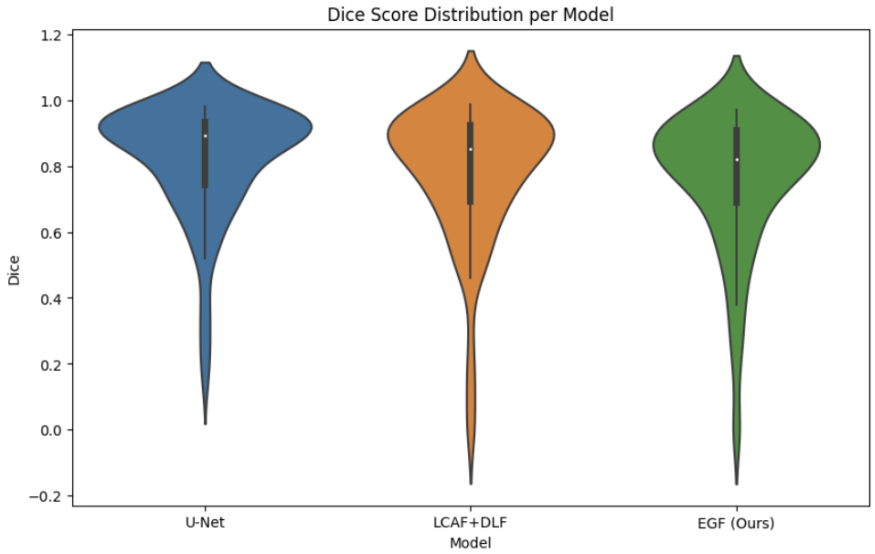
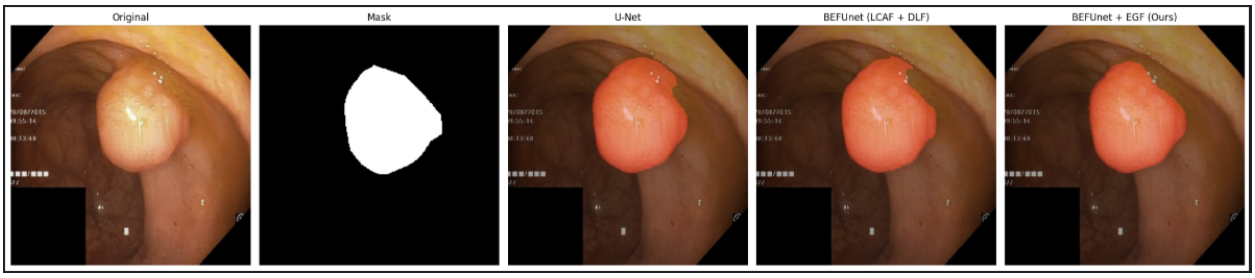


Figure 2: Violin plots of Dice, IoU, and sensitivity for U-Net, BEFUnet (EGF), and BEFUnet (LCAF+DLF) on the test set. U-Net exhibits strong central tendency with relatively tight distributions, whereas BEFUnet variants show slightly higher variance, reflecting their greater capacity and sensitivity to limited data.

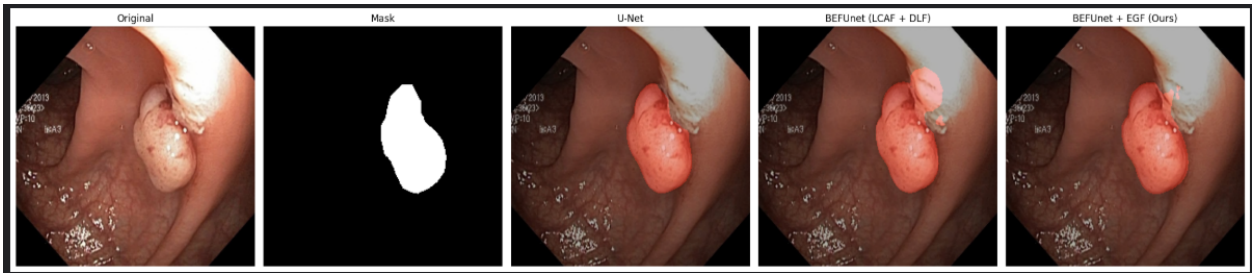
5.5 Qualitative Results

We present qualitative examples in Figure 3, comparing ground-truth masks with predictions from U-Net, BEFUnet (EGF), and BEFUnet (LCAF+DLF). These examples illustrate typical cases where:

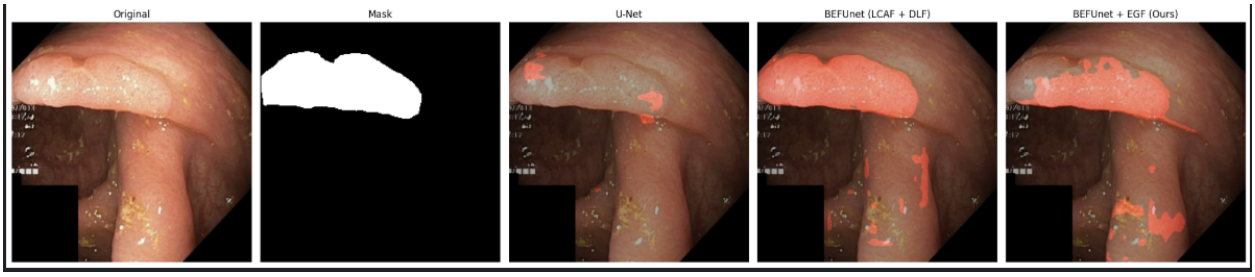
- U-Net captures the overall polyp shape with relatively clean boundaries,
- BEFUnet (EGF) produces comparable segmentations with slightly sharper edges,
- BEFUnet (LCAF+DLF) offers strong boundary detail but can exhibit over- or under-segmentation in certain difficult cases.



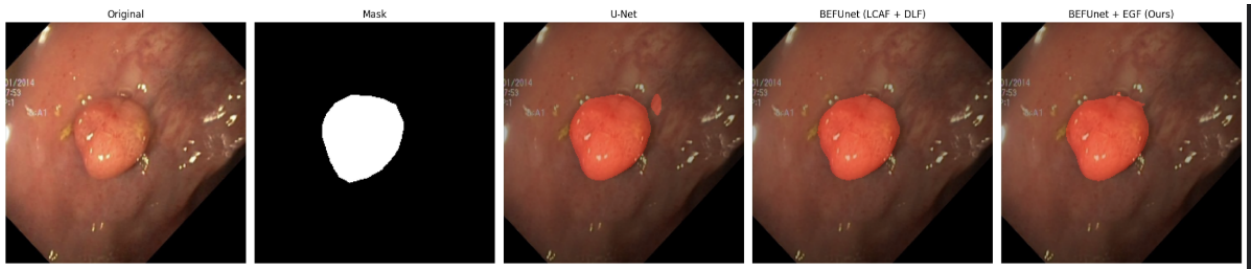
(a) Example 1.



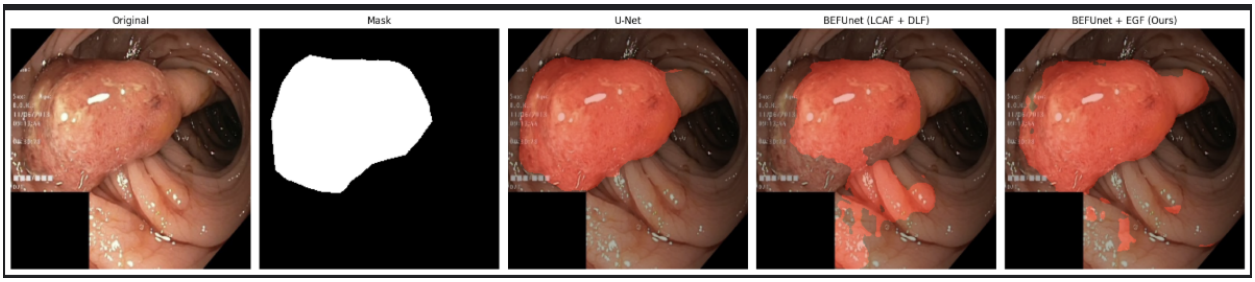
(b) Example 2.



(c) Example 3.



(d) Example 4.



(e) Example 5.

Figure 3: Qualitative segmentation results on Kvasir-SEG for five test images. Each row shows the input colonoscopy image, ground-truth mask, U-Net prediction, BEFUnet (LCAF+DLF) prediction, and BEFUnet (EGF) prediction.

6 Discussion

Our primary goal was to understand how BEFUnet-style architectures behave when trained on a relatively small dataset, and whether they continue to outperform U-Net in such a regime.

U-Net in the small-data regime. The results in Table 1 show that U-Net achieves the highest mean Dice and IoU among the three models. This highlights the strength of its convolutional inductive bias and relatively low capacity, which together promote robust generalization when only a few hundred training samples are available.

BEFUnet variants and overfitting. Both BEFUnet variants incorporate a transformer-based body encoder and additional fusion modules, increasing model capacity. In the small-data setting, this additional capacity does not translate into improved performance; instead, it makes the models more susceptible to overfitting. The violin plots in Figure 2 further suggest higher variance in their per-image scores.

Effectiveness of EGF. Our proposed Edge-Gated Fusion (EGF) module significantly simplifies the original LCAF+DLF fusion while maintaining boundary awareness. BEFUnet (EGF) achieves performance close to BEFUnet (LCAF+DLF) in Dice, IoU, and sensitivity, but with approximately half the training time per batch. This indicates that much of the benefit of boundary-aware fusion can be retained with a lightweight gating mechanism, at a fraction of the computational cost.

When to use BEFUnet. Our experiments suggest that BEFUnet-style architectures are better suited to settings where sufficient training data are available to leverage their capacity and more complex fusion mechanisms. In contrast, when annotated data are scarce, simpler architectures like U-Net may be preferable, both in terms of accuracy and efficiency.

7 Conclusion

We studied the behavior of BEFUnet-style hybrid CNN–Transformer architectures under a small-data regime using the Kvasir-SEG polyp segmentation dataset. Comparing a U-Net baseline, a full BEFUnet variant with LCAF+DLF fusion, and a lightweight BEFUnet (EGF) variant, we found that U-Net achieves the best mean Dice and IoU when trained on only 800 images. This result underscores the importance of aligning model complexity with dataset size: high-capacity hybrid architectures do not automatically outperform simpler baselines when annotated data are limited.

At the same time, our Edge-Gated Fusion (EGF) provides a practical improvement over the original fusion strategy. BEFUnet (EGF) achieves segmentation performance comparable to BEFUnet (LCAF+DLF) while training approximately twice as fast per batch, making it a more attractive choice under constrained computational budgets.

Future work includes exploring stronger regularization strategies, semi-supervised learning, and active learning to better exploit BEFUnet-style architectures in low-data settings, as well as extending our analysis to other medical imaging modalities and tasks.

References

[1] S. Wang, Y. Xiao, Y. Shi, Y. Chen, and Y. Gao. BEFUnet: Boundary-Enhanced Feature Fusion Network for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 2023.

[2] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[3] J. Chen, et al. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[4] D. Jha, et al. Kvasir-SEG: A segmented polyp dataset. In *MultiMed*, 2020.