Ostbayerische Technische Hochschule Amberg-Weiden
Department of Electrical Engineering, Media and Computer Science

CONCATENATION AND SUMMATION IN UNET AND UNET++

Students

Xusheng Guo

Tran Thai Duc Hieu

Supervised by

Prof. Dr. Tatyana Ivanovska

July 3, 2025

1 Introduction

UNet (Ronneberger et al., 2015) is a convolutional neural network designed for image segmentation, featuring a symmetric encoder–decoder structure. The encoder captures context through downsampling, while the decoder enables precise localization through upsampling. UNet improves segmentation accuracy with skip connections that link corresponding encoder and decoder layers. At each level of the decoder, the feature map produced by upsampling is concatenated with the corresponding feature map from the encoder path at the same spatial resolution.

UNet++ (Zhou et al., 2018), also known as Nested UNet, is an extension of the original UNet architecture designed to reduce the semantic gap between encoder and decoder feature maps. It introduces intermediate convolutional layers (nested blocks) and denser skip connections that allow for more gradual and refined feature fusion. By redesigning the skip pathways, UNet++ encourages better alignment between encoder and decoder features, which helps the network learn more effectively. The architecture also supports deep supervision, allowing the model to generate auxiliary outputs at different depths to improve training stability and performance. Experimental results, particularly in medical image segmentation, have shown that UNet++ outperforms the original UNet in terms of accuracy and generalization.

In both cases, the concatenation operation is used in skip connections to combine feature maps from the encoder with those in the decoder at the same level. This helps the network use both low-level details and high-level context when making predictions.

In this experiment, we explore an alternative method called element-wise summation, which adds the feature maps together instead of joining them. We compare how concatenation and summation affect the performance of UNet and UNet++, and see if summation can make the models simpler while still keeping good segmentation results.

All pretrained models are placed on GitLab: link

2 Related Work

The U-Net architecture (Ronneberger et al., 2015) and its numerous variants have emerged as the most widely adopted frameworks, particularly in biomedical applications due to their encoder-decoder structure and efficient use of skip connections.

Several studies have explored the use of U-Net and its variants in high-resolution image segmentation. For instance, the work by Hoşbas et al. (2020) evaluated the performance of UNet and UNet++ in the context of high-resolution image change detection, demonstrating the importance of model depth and skip pathway design for accurate localization of changes in remote sensing imagery. Similarly, Rasti et al. (2022) compared different U-Net architectures for retinal layer segmentation in posterior segment optical coherence tomography (OCT), highlighting the impact of architectural adjustments on segmentation accuracy in fine-grained, layered anatomical structures.

In the medical domain, U-Net and its extensions, such as UNet++, Attention U-Net, and ResUNet, have been systematically reviewed and benchmarked for their performance across various imaging modalities. The study by Rashid et al. (2024) provides a detailed overview of how these architectures have evolved to address challenges like class imbalance, limited data, and varying image resolutions in medical segmentation tasks. Their work emphasizes improvements introduced by nested architectures, attention mechanisms, and multi-scale feature fusion.

Moreover, the review conducted by Ahsan et al. (2024) presents a comprehensive taxonomy of U-Net variants and evaluates their applications in both 2D and 3D medical imaging. Their findings show that while classical U-Net remains a strong baseline, newer variants significantly improve performance in complex tasks such as multi-organ segmentation, lesion detection, and modality adaptation.

Together, these studies underscore the versatility and continued evolution of the U-Net family in handling diverse medical imaging challenges, motivating further comparative analysis across multiple datasets and modalities.

3 Datasets

In this project, several public datasets were used to evaluate the performance of different variants of UNet and UNet++ in medical image segmentation tasks. The datasets span a range of medical imaging modalities and anatomical targets, offering a diverse test for assessing segmentation performance under varied conditions.

3.1 BUSI (Breast Ultrasound Images) dataset

The BUSI dataset contains ultrasound images of the breast, annotated for tumor segmentation tasks. It was published as part of the Hi-gMISnet project, which focuses on high-granularity medical image segmentation. The dataset consists of 780 images in total.

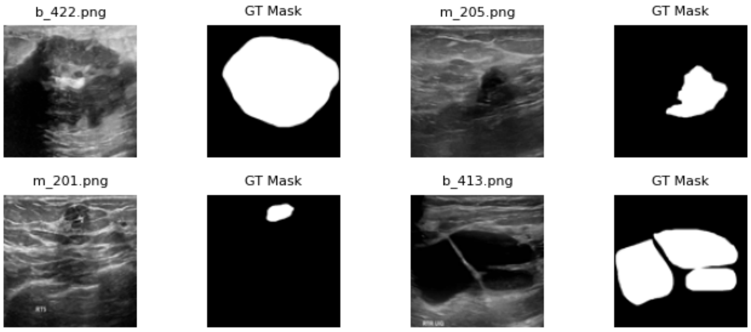

Figure 1: Example images in BUSI dataset

3.2 DynamicNuclearMSBench dataset

This dataset consists of high-resolution microscopic images designed for binary segmentation of nuclear regions. The dataset includes more than 7,000 labeled images featuring different types of cells and staining methods. It is challenging because the nuclei often overlap, the image brightness can change subtly, and the shapes of the cells are complex. These factors make it a tough test for segmentation algorithms. The dataset was sourced from the MedSegBench repository, which standardizes medical segmentation benchmarks.

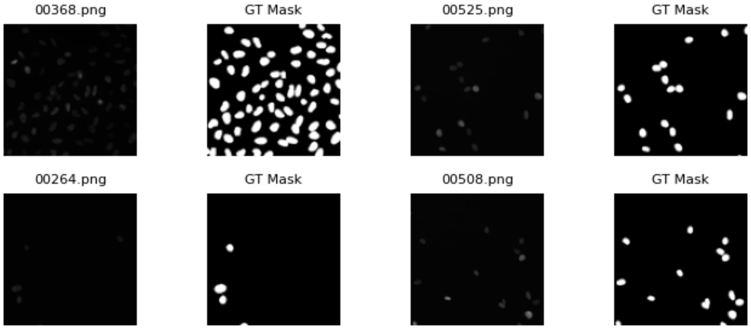

Figure 2: Example images from DynamicNuclear dataset

3.3 Covid19 Radiography dataset

Comprising chest X-ray images from COVID-19 and other pulmonary conditions (e.g., viral pneumonia, lung opacity, and normal cases). It contains a total of 21,165 images, with 14,814 for training, 2,115 for validation, and 4,236 for testing.

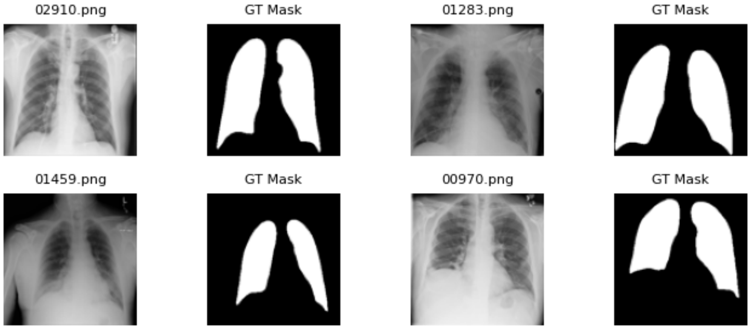

Figure 3: Example images from Covid19 Radiography dataset

3.4 UsforKidney dataset

The USForKidney dataset consists of CT-to-ultrasound image pairs annotated for kidney segmentation. It is particularly useful for evaluating model generalizability across modalities. The dataset includes over 4,500 labeled examples spanning training, validation, and test splits.

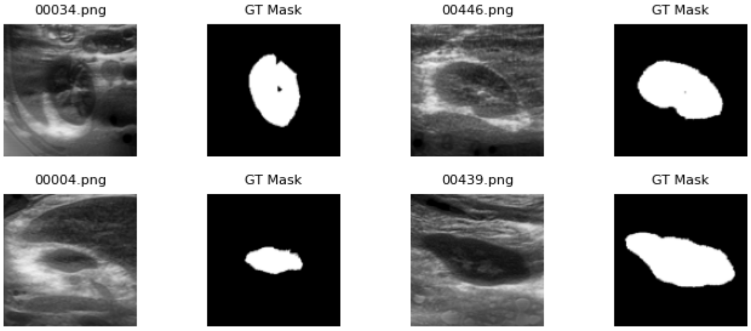

Figure 4: Example images in UsForKidney dataset

4 Model architectures

4.1 UNet

UNet (Ronneberger et al., 2015) is based on the Fully Convolutional Network (FCN) by Long et al. (2015). It was designed to work well with a small amount of training data and to give more accurate results in image segmentation tasks. The network has a U-shape structure with two main parts: a downsampling path that captures the context and an upsampling path that helps recover the details. One important feature of UNet is the "skip connections". These connections pass detailed information from the downsampling part to the upsampling part, helping the network keep important features that might otherwise be lost.

In this project, the UNet model was implemented based on the original design, with the addition of Batch Normalization layers to improve training speed and stability (Ioffe & Szegedy, 2015). One more modification in this experiment is that zero-padding is applied instead of no padding in the original paper. Each downsampling block consists of two convolutional layers with 3×3 kernels and 1-pixel padding, followed by Batch Normalization. The first convolution in each block receives input from the output of a max pooling layer, except in the first block, where the input is the combined RGB channels of two images. All max pooling operations use a 2×2 kernel with a stride of 2 to reduce spatial dimensions. With each downsampling block, the number of feature channels is doubled compared to the previous level, allowing the network to learn increasingly complex features (see Figure 5).

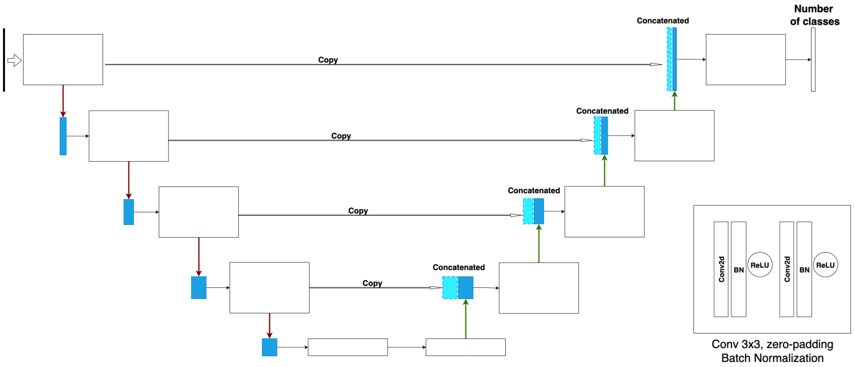

Figure 5: UNet reference architecture diagram

4.2 Proposed architecture: UNet Summation

The UNet Summation model is a modified version of the original UNet architecture proposed by the student group in this project. The most notable change in this UNet version is the replacement of the concatenation operation with an element-wise addition at the beginning of each upsampling block (see Figure 6).

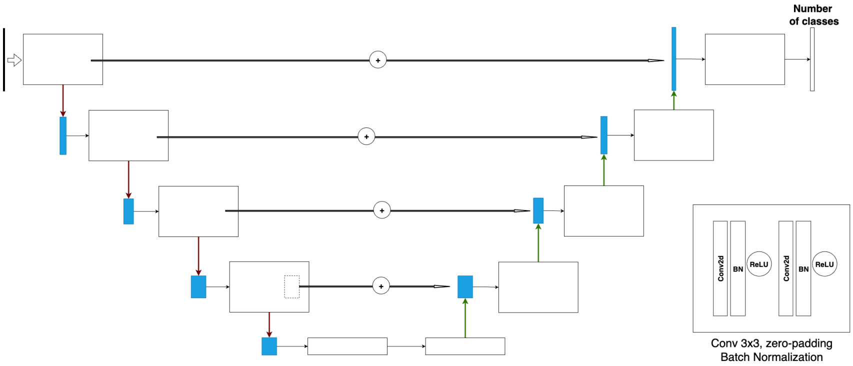

Figure 6: UNet Summation reference architecture diagram

In the standard UNet, concatenation increases the number of feature channels by combining the output of the corresponding downsampling block with the upsampled features. In contrast, the use of element-wise addition in UNet Summation maintains a constant number of channels after the skip connections, as both feature maps involved in the addition must have identical dimensions. As a result, the spatial and channel dimensions of the upsampling blocks remain consistent with those of the corresponding convolutional blocks before the skip connections.

4.3 UNet++ (Nested UNet)

Nested UNet, also known as UNet++, is an improved version of the original UNet model, introduced by Zhou et al. (2018) to achieve better segmentation accuracy. UNet++ keeps the same basic encoder–decoder structure as UNet but introduces more intermediate convolutional blocks (called "nested blocks") and denser skip connections between the encoder and decoder. The main idea is to gradually improve the high-resolution feature maps before combining them with the decoder output. This helps the model focus better on fine details in the image, as the features being merged are more semantically similar, meaning that they contain more related or meaningful information.

UNet++ was also proposed to use deep supervision in the original paper, which means that multiple intermediate outputs from different decoder depths can be used to compute the loss during training. This strategy encourages the network to learn discriminative features at various scales, making the training process more stable and improving segmentation performance across different object sizes. During inference,

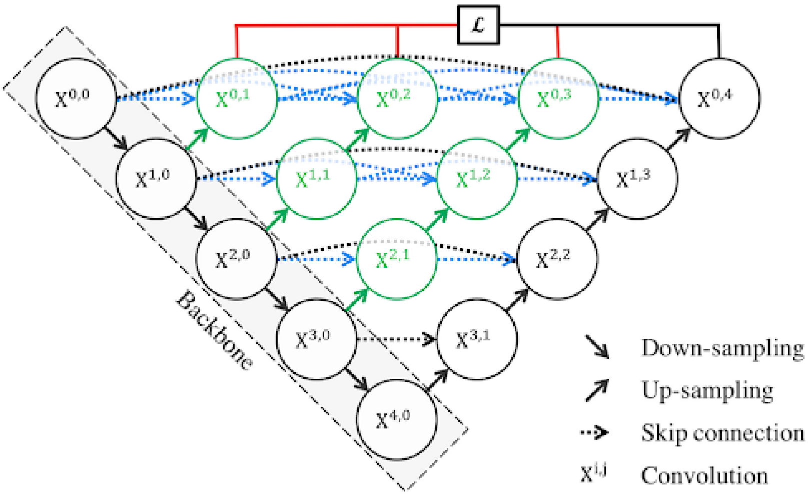

Figure 7: UNet++ reference architecture (Zhou et al., 2018)

the architecture can be pruned, meaning some decoder paths can be removed, reducing computational cost while maintaining high accuracy.

4.4 Proposed architecture: UNet++ Summation

With the same idea as UNet Summation, the most important feature of this architecture is the replacement of the concatenation operation with element-wise addition at the decoder stages. In the original UNet++ architecture, multiple nested skip connections are combined using concatenation, which increases the number of feature channels. In contrast, UNet++ Summation uses element-wise addition to merge feature maps. As a result, the channel dimensions remain consistent for encoder and decoder blocks on the same level. All other configurations, including the nested structure, dense skip connections, and optional deep supervision, remain unchanged from the original UNet++ design.

5 Experiments

To evaluate the effect of different skip connection strategies—namely concatenation and element-wise summation—on UNet and UNet++ architectures, a series of controlled experiments were conducted across all datasets.

5.1 Data preparation

Step 1: Downloading the BUSI Dataset
The BUSI dataset was obtained directly from the Hi-gMISnet project website. It contains annotated breast ultrasound images and was downloaded manually due to its public availability on Kaggle.

Step 2: Downloading other datasets via script
For the other datasets—DynamicNuclearMSBench, Covid19 Radiography, and UsForKidney—a Python script was developed to automate the downloading process. This script utilized the medsegbench library, which provides convenient access to standardized medical image segmentation benchmarks.

Step 3: Organizing folder structure
Once all datasets were downloaded, the files and annotations were organized into a consistent directory structure. The folder layout was modeled after the BUSI dataset's format to ensure compatibility with the data loading and preprocessing pipeline.

Step 4: Verifying consistency across datasets
To make sure all the datasets could be used together smoothly during training, we checked that their file names, image formats, and annotation files followed a consistent structure.

5.2 Training implementation

All models were implemented in PyTorch and trained using NVIDIA RTX 2080 GPUs. The training pipeline was standardized across experiments to ensure comparability. Each model was trained independently on each dataset

We adopted the Dice Loss function, given its robustness to class imbalance, especially in medical image segmentation. The Adam optimizer was used with an learning rate of 0.0001. Early stopping was employed with a patience of 5 epochs.

5.3 Evaluation metrics

We evaluated segmentation performance using metrics that capture both overlap quality and classification correctness at the pixel level. All our datasets are binary medical image segmentation problems (e.g., lesion vs. background), where target regions are often sparse and class imbalance is common. The following metrics were used:

- **Mean Intersection over Union (IoU):** Also known as the Jaccard Index, IoU measures the overlap between prediction and ground truth masks averaged over the test set:

  $$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}$$

  It penalizes both over-segmentation and under-segmentation and is widely used as a benchmark metric in segmentation challenges.

- **Mean Dice Coefficient:** The Dice score quantifies the similarity between two sets. It is defined as:

  $$\text{Dice} = \frac{2 \cdot |P \cap G|}{|P| + |G|}$$

  This is the same as the F1 score in binary classification, and it emphasizes correct overlap more than IoU in some edge cases.

- **Pixel Accuracy:** The ratio of correctly classified pixels (both foreground and background) over the total number of pixels:

  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  While easy to interpret, pixel accuracy can be misleading on imbalanced datasets where background pixels dominate.

- **Confusion Matrix:** We compute a 2×2 confusion matrix at the pixel level for binary classification. This summarizes true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), enabling further analysis.

- **IoU per Class:** To assess class-wise segmentation quality, we compute IoU separately for each class (foreground and background):

  $$\text{IoU}_i = \frac{TP_i}{TP_i + FP_i + FN_i}$$

5.4 Results and analysis

All pretrained models are placed on GitLab: link

Inference was performed on the test split of each dataset, with each model evaluated separately on its corresponding test set. For every dataset, the reported metrics represent the average performance across all test samples.

The primary evaluation metrics used in this study are the mean Dice score and the Foreground Intersection over Union (IoU), which reflect segmentation accuracy in general and how a specific model adapts to the foreground of the dataset.

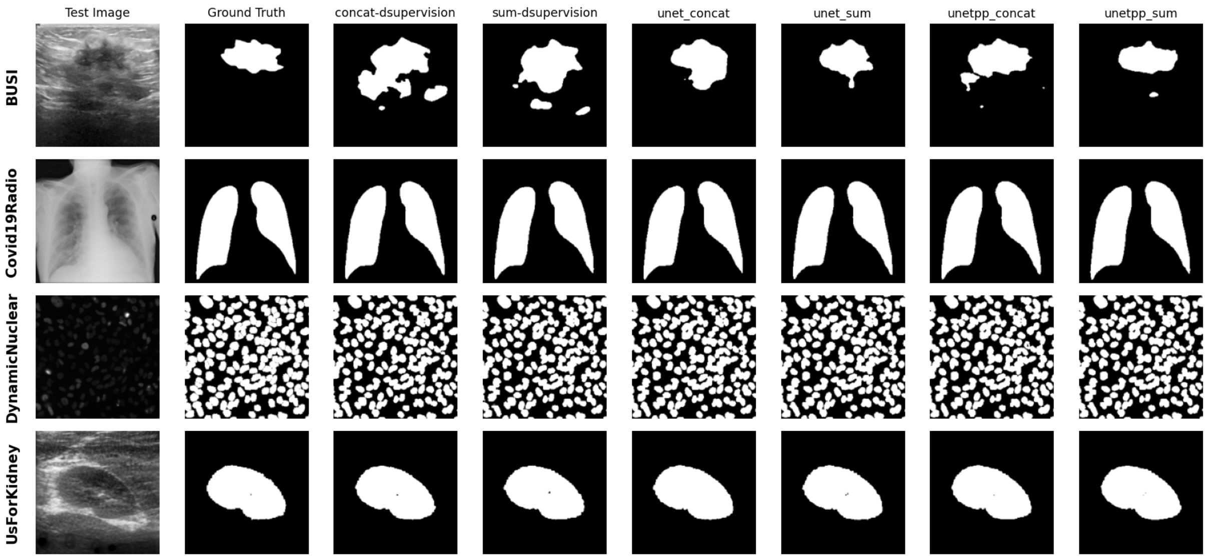

Figure 8: Example images predictions from each dataset and each model

The evaluation metrics for each dataset are shown in four tables below.

Table 1: Results on BUSI dataset

Model	Mean IoU	Mean Dice	Pixel Accuracy	IoU Background	IoU Foreground
concat-dsupervision	0.7883	0.7566	0.9541	0.9789	0.7337
sum-dsupervision	0.8050	0.7780	0.9598	0.9863	0.7249
unet_concat	0.8051	0.7797	0.9581	0.9876	0.6963
unet_sum	0.7799	0.7422	0.9522	0.9900	0.6179
unetpp_concat	0.8208	0.8020	0.9630	0.9878	0.7438
unetpp_sum	0.8230	**0.8058**	0.9636	0.9868	**0.7581**

Table 2: Results on UsForKidney dataset

Model	Mean IoU	Mean Dice	Pixel Accuracy	IoU Background	IoU Foreground
concat-dsupervision	0.9700	0.9753	0.9905	0.9941	0.9756
sum-dsupervision	0.9709	0.9760	0.9908	0.9938	0.9786
unet_concat	0.9721	0.9770	0.9911	0.9939	**0.9798**
unet_sum	0.9727	**0.9775**	0.9914	0.9942	0.9796
unetpp_concat	0.9713	0.9764	0.9909	0.9944	0.9765
unetpp_sum	0.9711	0.9762	0.9908	0.9945	0.9760

Table 3: Results on Covid19Radio Dataset

Model	Mean IoU	Mean Dice	Pixel Accuracy	IoU Background	IoU Foreground
concat-dsupervision	0.9858	0.9889	0.9949	0.9968	0.9890
sum-dsupervision	0.9853	0.9885	0.9947	0.9971	0.9872
unet_concat	0.9869	**0.9898**	0.9953	0.9972	0.9892
unet_sum	0.9855	0.9888	0.9948	0.9964	**0.9897**
unetpp_concat	0.9859	0.9891	0.9949	0.9967	0.9895
unetpp_sum	0.9856	0.9888	0.9948	0.9965	0.9894

Table 4: Results on DynamicNuclear dataset

Model	Mean IoU	Mean Dice	Pixel Accuracy	IoU Background	IoU Foreground
concat-dsupervision	0.9226	0.9249	0.9860	0.9944	0.9313
sum-dsupervision	0.9215	0.9241	0.9856	0.9937	0.9324
unet_concat	0.9257	0.9280	0.9866	0.9941	0.9375
unet_sum	0.9264	**0.9287**	0.9866	0.9946	0.9345
unetpp_concat	0.9231	0.9251	0.9864	0.9925	**0.9461**
unetpp_sum	0.9203	0.9235	0.9851	0.9929	0.9340

5.4.1 BUSI experiment analysis

For the BUSI dataset, our Summation UNet++ model achieved the highest scores in both Dice and IoU Foreground metrics, with **0.8058** and **0.7581** respectively. In contrast, the UNet++ model with Deep Supervision performed the worst in this experiment.

To better understand these results, we need to consider the characteristics of the BUSI dataset. It consists of ultrasound images, which are typically noisy and have low resolution. As a result, the boundaries between the foreground and background in the images are often blurry and unclear.

This blurriness poses a challenge for more complex architectures like UNet++ with Deep Supervision. Such models tend to be confused between foreground and background regions, leading to more False Positive predictions. This issue is evident in Figure 8 and Table 1, where the IoU Background metric for UNet++ with Deep Supervision is the lowest, at just **0.979**.

5.4.2 Covid19Radio and UsForKidney experiment analysis

The Covid19Radio and UsForKidney datasets share several similarities in image characteristics. Both datasets contain relatively clear images, with UsForKidney images being slightly more blurred; however, the boundaries of the foreground objects remain distinct and easily recognizable by human eyes. Additionally, the foreground regions in these datasets tend to be large and exist in only one or two objects (two sides of a lung).

This clarity and simplicity of the foreground objects appear to favor simpler architectures. In fact, as shown in Tables 3 and 2, models such as UNet and Unet Summation achieved the best results.
In the Covid19Radio dataset, the UNet model received the highest Dice score of **0.9898**, while the UNet Summation model has the highest IoU Foreground of **0.9897**.
For the UsForKidney dataset, the UNet Summation model obtained the highest Dice score of **0.9775**, while the UNet model has the highest IoU Foreground of **0.9798**

5.4.3 DynamicNuclearMSBench experiment analysis

The DynamicNuclearMSBench dataset presents a more challenging scenario due to the presence of overlapping nuclei, subtle intensity variations, and complex cell morphologies. These factors make accurate segmentation difficult, requiring models to effectively distinguish closely packed and irregularly shaped objects.

However, as shown in Table 4, the results are very surprising. Despite the dataset's complexity, metrics for UNet++ variants with Deep supervision were not the highest. The performance differences between models using concatenation and those using element-wise summation are quite small.

The UNet Summation model achieved the highest Dice score of **0.9287**. Meanwhile, the UNet++ with concatenation reached the best IoU Foreground score of **0.9461**

6 Conclusion and future work

In this work, we compared two ways of connecting the encoder and decoder in UNet and UNet++—by either concatenating their feature maps or by adding them element by element—using four public medical imaging datasets. As shown in the results tables, there is always at least one summation-based model that achieves the highest score in either Dice or IoU Foreground.

Based on our analysis in the report, we found that the performance of skip connection strategies - concatenation and element-wise summation, also depends on the specific dataset. While concatenation is still the most commonly used and reliable method, element-wise summation also achieved strong results, offering similar accuracy with a simpler architecture and fewer parameters.

These findings suggest that summation is a promising alternative, especially for scenarios where computational efficiency is important, without sacrificing much in segmentation performance.

Future work

Future directions for this work include:

- **Extension to other UNet variants:** Expanding the experiments to include more advanced UNet-family models, such as Attention UNet, ResUNet, and U2Net, to evaluate whether summation maintains its effectiveness across architectures with attention mechanisms or residual blocks.
- **3D Medical Image Segmentation:** Applying the same comparison between concatenation and summation on 3D UNet and UNet++ models for volumetric datasets (e.g., CT or MRI scans).
- **Different loss functions:** Compare with more loss functions such as Binary Cross Entropy (BCE) loss or combination of BCE and Dice loss
- **Cross-Dataset Generalization:** Training models on one dataset and testing on another to assess generalization ability under domain shift, which is critical for clinical deployment.

Overall, this project presents a meaningful step toward designing more efficient and adaptable U-Net-based architectures by rethinking the role and form of skip connections in medical image segmentation networks.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, 2018, pp. 3–11.

[3] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *International Conference on Machine Learning (ICML)*, 2015.

[4] H. Hoşbas et al., "High-resolution Image Change Detection Using UNet and UNet++," *Remote Sensing Letters*, vol. 11, no. 9, pp. 839–848, 2020.

[5] R. Rasti et al., "Comparison of U-Net Architectures for Retinal Layer Segmentation in OCT," *Biomedical Optics Express*, vol. 13, no. 2, pp. 890–905, 2022.

[6] M. Rashid et al., "A Review of U-Net Variants for Medical Image Segmentation," *Medical Image Analysis*, vol. 80, pp. 102526, 2024.

[7] M. Ahsan et al., "Taxonomy and Performance Evaluation of U-Net Variants in 2D and 3D Medical Imaging," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 123–139, 2024.

[8] T. Talukder Showrav and Md. K. Hasan, "Hi-gMISnet: generalized medical image segmentation using DWT based multilayer fusion and dual mode attention into high resolution pGAN," *Physics in Medicine and Biology*, 2024.

[9] Ozan Oktay, Jo Schlemper, Luke Le Folgoc, Matthew Lee, Maria Heinrich, Kazunari Misawa, Kensaku Mori, Stephen McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., *Attention U-Net: Learning Where to Look for the Pancreas*, arXiv preprint arXiv:1804.03999, 2018.

[10] Hao Zhang, Wei Wu, Zhili Li, Dan Zhu, *Road extraction by deep residual U-Net*, IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 749–753, 2018.