FeU-Net: overcomplete representations with large kernels for edge detection

1st Federico Urli
Dept. DMIF
University of Udine
Udine, Italy
ORCID: 0000-0001-9861-9677

2rd Michele Somero
Dept. DMIF
University of Udine
Udine, Italy
ORCID: 0000-0001-5846-8827

3nd Lauro Snidaro
Dept. DMIF
University of Udine
Udine, Italy
ORCID: 0000-0003-3828-9017

4th Chad Johnson
Enabling Technologies
LimaCorporate SpA
Udine, Italy
chad.johnson@limacorporate.com

5th Tiziano Vallisa
Enabling Technologies
LimaCorporate SpA
Udine, Italy
tiziano.vallisa@limacorporate.com

6th Ingrid Visentini
Enabling Technologies
LimaCorporate SpA
Udine, Italy
ingrid.visentini@limacorporate.com

***Abstract*—In recent years, segmentation algorithms utilizing deep learning have achieved outstanding performance in medical image segmentation. However, accurately delineating small anatomical structures continues to be a challenging task, even for the most advanced methods that produce impressive results. This challenge might arise from the use of small kernels and downsampling operations, which often emphasize complex high-level features at the expense of low-level details like edges.**

Inspired by recent research highlighting this challenge, we developed a novel architecture that combines the standard U-Net with an additional branch harnessing the potential of large convolutional kernels. These large kernels are utilized in a decreasing-increasing manner over image features of the same size, guiding the network to focus on smaller parts.

The proposed method demonstrated strong potential in segmenting small anatomical structures, surpassing our baseline and matching the performance of a robust state-of-the-art network across various datasets and domains, all while maintaining a relatively small number of parameters.

***Index Terms*—Semantic segmentation, U-Net, Edge detection, Overcomplete networks, Large kernels, Features fusion, Medical imaging.**

I. Introduction

Segmenting medical images presents a significant challenge due to their nature, often characterized by problems such as blur, noise, or low contrast. Therefore, algorithms relying on edge detection, template matching, or traditional machine learning, while successful in specific scenarios, often encounter difficulties because they are based on hand-crafted features.

Deep learning techniques overcome this limitation since they automatically infer the best features to solve a specific task [13]. Numerous deep learning algorithms have been utilized in medical image segmentation. U-Net [8], for instance, has garnered widespread adoption and serves as a precursor to many more advanced architectures that have emerged over the years. A few examples, among many others, are [14] [9] [6] [3]. These architectures are frequently structured as encoder-decoder systems. Initially, the image undergoes processing by the encoder, which utilizes multiple convolutions and downsampling operations to extract essential features. Subsequently, the extracted features are forwarded to the decoder, which endeavors to restore the original image resolution and generates the segmentation mask as output.

This technique has achieved satisfactory results across various image segmentation datasets and domains, excelling notably in certain instances. Nevertheless, authors frequently report a decrease in accuracy along edges [8] [3], likely attributable to the nature of the architectures and their downsampling operations. Indeed, downsampling operations cause the convolution kernels of the network to expand their receptive field, enabling them to learn essential high-level features like objects, shapes, or blobs. However, a consequence of this is a diminished capacity for the convolution kernels to concentrate on fine details.

The authors of [12] [10] aptly describe this issue, asserting that in classical encoder-decoder architectures, only the shallow layers of the encoder are primarily focused on capturing low-level features such as edges. To address this challenge, they mitigate the issue by incorporating an extra branch into the network, which conducts upsampling operations on the image instead of downsampling. The rationale behind this approach lies in the fact that by upsampling the image, the receptive field of convolution kernels is reduced. Consequently, the neural network should concentrate on smaller patches, effectively capturing edges.

Because capturing edges plays a critical role in semantic segmentation, we experimented with a comparable method. Instead of employing upsampling operations on the image, we preserved its size while reducing the receptive field of kernels. This was achieved by initially utilizing large kernels and subsequently transitioning to smaller ones.

We believe this approach could be advantageous for two reasons. Firstly, the interpolation methods utilized in upsam-

pling operations might influence the appearance of the edges we intend to extract. Secondly, recent research [2] suggests that large kernels demonstrate a greater bias towards shape features, whereas small kernels are more inclined towards texture features [4]. Therefore, large kernels may be more suitable for edge extraction.

To prove the effectiveness of this approach, we developed a novel network with two branches similarly to [12]. The first branch follows the conventional encoder-decoder architecture, employing downsampling operations and small kernels in the encoder, followed by upsampling operations in the decoder. In contrast, the second branch maintains the image resolution at each stage. Meanwhile, in the encoder, the network begins with very large kernels, shifts to smaller kernels in deeper layers, and then progressively increases the kernel size in the decoder. Indeed, the two branches exchange features with each other through a novel block that combines and fuses features. In this way we obtained a significant improvement on Kvasir [7] and CVC-Clininc [1] datasets, outperforming on Dice metric a baseline with an equal number of parameters and competing against a state-of-the-art network with four times the parameters [3].

The main contributions of this work are:

- A novel architecture that focus on small anatomical structures
- An innovative fusion block that aims to integrate complementary features
- Extensive experimentation on different domains

The paper is organized as follows: Section II introduces the developed methods. The configuration of the experiment is presented in III, while final remarks and future works are discussed in Section IV.

II. Proposed method

A. Large kernels for overcomplete representations

In a standard encoder-decoder architecture, the receptive field of the filters expands with deeper layers in the network. This enlargement in receptive field size can be explained by two factors: (i) each convolution layer filter collects data from its neighboring window, and (ii) the inclusion of a max-pooling layer following every convolution layer. Essentially, the receptive field size doubles after each convolution layer due to the max-pooling layers. Consequently, apart from the initial layer, filters in subsequent layers exhibit diminished capacity to capture intricate features such as edges and their textures. Networks employing the conventional undercomplete architecture tend to generate less precise predictions around edges in tasks such as segmentation. In response to this challenge, we present a Focus Edge U-Net (FeU-Net), which implements an ensemble approach with two branches. In one branch, larger kernels are followed by smaller ones in the encoder, and smaller kernels are followed by larger ones in the decoder. The branch maintains the exact resolution of features. Meanwhile, the other branch adopts a standard U-Net architecture. The assumption behind the first branch is founded on the premise that through the reduction of the receptive field size of kernels, we drive the filters in the deeper layers to focus on acquiring intricate edge details [12], prioritizing smaller regions. Furthermore, we integrate a fusion layer after every convolution layer in both branches to facilitate the exchange of features between the two parts. The combined network leverages both the low-level fine edge capturing feature maps and the high-level shape feature maps.

B. Architecture

The input image is simultaneously passed through both branches. As we can see in Fig. 1, in both branches we have 4 residual convolution blocks in the encoder as well in the decoder.

In the standard U-Net branch, the initial resolution is reduced by downsampling operations performed by a 2x2 convolution block with stride 2. Features are then passed to a bottleneck layer and then processed by the decoder that restores feature maps to their original size using transposed convolutions. Through these downsampling and upsampling operations, 3x3 convolution blocks of the U-Net branch are able to infer hierarchical features and extract important high-level information from images.

In the FE-Net branch the resolution of feature maps of each stage is maintained equal to the original size of the image. Instead, the size of kernels is reduced at each stage of the encoder and increased at each stage of the decoder. We chose the kernel sizes to approximately halve the receptive field at each stage of the encoder. We selected the kernel sizes to approximately halve the receptive field at each stage of the encoder. Specifically, the chosen kernel sizes are 19, 13, 7, and 3 in their respective stages of the encoder, and they are then applied in reverse order in the decoder. In this way the network leaded to extract features of small parts. As this branch could potentially utilize a large number of parameters due to the use of large kernels, we decided to reduce its number of filters compared to the other one.

To enhance the performance of the two networks, we have devised a new approach called the Cross Squeeze and Excite Fusion Block (SEFB), visible in Fig. 2. This block effectively captures complementary features from both branches of the network and channels them to each branch accordingly. The SEFB comprises resizing layers to harmonize the features of one branch with the spatial dimensions of the other, 1x1 convolutions for adjusting features along the depth dimension, and squeeze and excitation blocks (SE) as introduced by [5]. Therefore, the resized features from each branch undergo regulation through 1x1 convolutions, followed by summation with features of the other branch, and are then passed through the SE block. The latter allows for the individual weighting of each channel of the fused features, enabling the network to emphasize complementary information. Finally, features are combined with the original input. SEFB can be described by formulas in the following way:

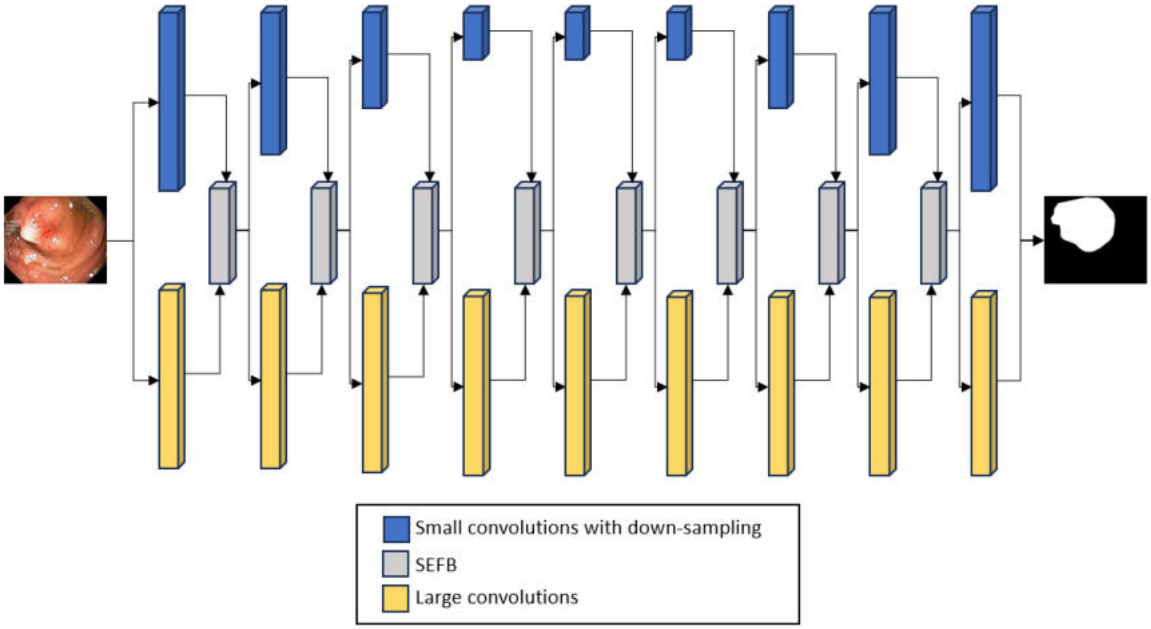

Fig. 1: Model architecture.

$$X_i = SE(X_i + Resize_{h_x \times w_x}(Conv_{1\times1\times C_x}(Y_i)))$$

$$Y_i = SE(Y_i + Resize_{h_y \times w_y}(Conv_{1\times1\times C_y}(X_i)))$$

where:

- X is the tensor of features of U-Net branch
- Y is the tensor of features of FE-Net branch
- i is the current stage of features in the network
- h and w are height and width of corresponding branch features
- C is the number of channels of corresponding branch features

Prior to the prediction layer, which consists of a 1x1 convolution layer with sigmoid activation, features from both branches are concatenated together.

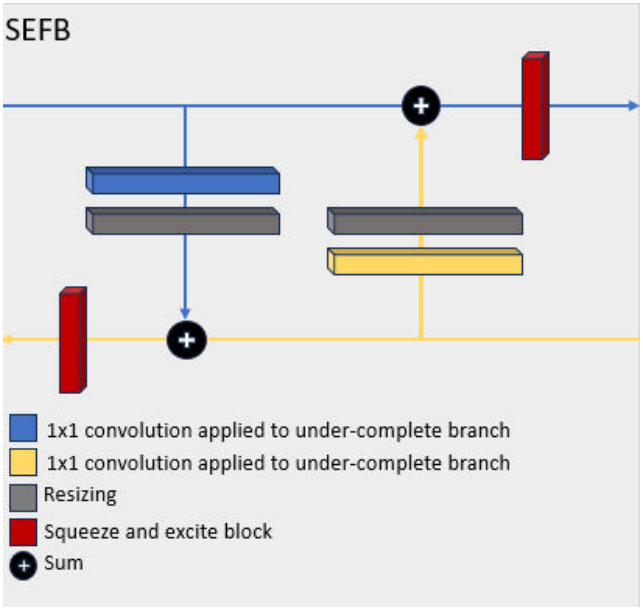

Fig. 2: Squeeze and excite fusion block.

III. Experiments

A. Datasets

We conduct experiments on four widely used datasets for polyp segmentation: Kvasir-SEG, CVC-ClinicDB.
The Kvasir-SEG comprises 1000 polyp images along with their respective ground truth annotations, exhibiting various resolutions spanning from 332 × 487 to 1920 × 1072 pixels. The CVC-ClinicDB consists of 612 polyp images paired with their ground truth annotations, all having a resolution of 384 × 288 pixels. Prior to processing, the resolution of each image of both datasets was adjusted to 256 × 256 pixels.

B. Evaluation methods

For the assessment of our method on Kvasir-SEG and CVC-ClinicDB datasets, we employed K-fold cross-validation with K=3. This involved evaluating the performance of each tested network design multiple times on various dataset splits, aiming to obtain a reliable estimate of the network's ability to generalize. As evaluation metrics, we utilized two widely recognized measures commonly employed in the assessment

of semantic segmentation algorithms: the Dice score and Intersection over Union (IoU). The Dice score and IoU are calculated for each fold using the best model specific to that fold. Subsequently, the average of these scores across all folds is computed to derive the final performance score.

C. Implementation details

The FeU-Net is trained using Dice loss with the RMSprop optimizer [11], employing a batch size of 4. The learning rate was configured to be 0.0001. This network architecture was implemented in the TensorFlow framework and trained on an Nvidia-RTX A5000 GPU. The training process lasted a total of 200 epochs for each fold. We utilize identical data augmentation techniques as described in [3], incorporating horizontal and vertical flips, shearing, rotations, and scaling to manipulate image shapes, along with adjustments to textures involving changes in brightness, contrast, saturation, and hue.

D. Comparison with other methods

As the principal objective of this study is to enhance the U-Net architecture with additional capabilities, we initially assess our method against U-Net. Subsequently, we compared it with a robust state-of-the-art architecture called Duck-Net [3] which is currently the CNN with highest performance on polyp segmentation. Table I demonstrates that the proposed method significantly outperforms U-Net on both datasets and achieves comparable performance with Duck-Net on the Kvasir-Dataset. On Kvasir-Dataset, the proposed method achieves a 2.3% improvement in Dice score and 3.6% in IoU compared to U-Net and matches the performance of [3].

We also tried the Fe-Net branch alone. Due to the small number of parameters, it does not achieve the results of the others. Moreover, as an architecture that relies solely on edges, its ability to identify complex objects suffers.

Network	Dice	IoU	N. Parameters
Fe-Net (ours)	0.715	0.566	1M
U-Net	0.844	0.730	10M
FeU-Net (ours)	0.867	0.766	10M
Duck-Net	0.869	0.769	39M

TABLE I: Results of networks on Kvasir dataset.

On the other hand, on CVC-ClinicDB, the proposed method achieves a 2% improvement in Dice score and 2.9% in IoU compared to U-Net. In this dataset, our network appears to have slightly inferior performance compared to [3].

As on the kvasir dataset, the Fe-Net branch alone does not achieve very good results, but is able to make its contribution in combination with the U-Net. All results are visible in Table II.

It is worth emphasizing that we designed both the U-Net and the proposed method FeU-Net to have the same number of parameters (10 million) in order to make a fair comparison, whereas Duck-Net comprises approximately 39 million parameters.

Network	Dice	IoU	N. Parameters
Fe-Net (ours)	0.686	0.541	1M
U-Net	0.866	0.764	10M
FeU-Net (ours)	0.885	0.793	10M
Duck-Net	0.899	0.811	39M

TABLE II: Results of networks on CVC-ClinicDB dataset.

E. Ablation study

We also conducted an ablation study in order to assess the potential of our architecture. In particular, we wanted to assess if our network is effective even without the SE block. Thus, we conducted k-fold validation over a network with this block removed. Fusion between features was performed through simple summation. The network achieved convincing results, surpassing U-Net by 1.6% on kvasir dataset and reaching a Dice score of 0.86. It demonstrated the ability to leverage features from both branches. However, the SE block appeared to assist the network in selecting complementary features more effectively, consequently leading to superior performance. Additionally, it is worth noting that this network has approximately 0.5 million parameters, which may marginally contribute to inferior performance.

F. Visual Inspection

We conducted a visual inspection in order to evaluate the effectiveness of our methods. So, we inspected all the prediction masks created by the FeU-Net and we compared them with those ones created by the U-Net. The sense of this operation is to check if the accuracy on borders is effectively increased by our experimental network.

As shown in Fig. 3, it is clear that FeU-Net exhibits superior effectiveness in detecting small, potentially hard to detect parts. However, we have also noticed that our method performs better when segmenting well identified areas. This discrepancy in performance could be attributed to the challenge U-Net faces in identifying objects, especially complex shapes, leading to the Fe-Net branch struggling to effectively detect their edges and small structures.

We additionally provide a comparison of the activation patterns of encoder filters between Fe-Net and U-Net, as illustrated in Fig. 4. It is evident that in U-Net, the activation of filters appears to be more sparse across the entire feature map. Conversely, in Fe-Net, filters appear to be more concentrated on specific regions.

G. Experiment on other modalities

To assess the generalizability of our approach, we conducted tests on a distinct and intricate domain. Specifically, we utilized an internal dataset focused on shoulder anatomy. Our training involved U-Net and FeU-Net models trained on 2D slices derived from splitting 3D CT scans of the shoulder. In this way, we obtained 25035 bi-dimensional images. Each mask contains pixels corresponding to the humerus and

(a) Ground Truth

(b) FeU-Net Predictions

(c) U-Net Predictions

Fig. 3: In this image, the ground truth from the Kvasir dataset is depicted in the first row, while the predictions of FeU-Net and U-Net are presented in the second and third rows, respectively.

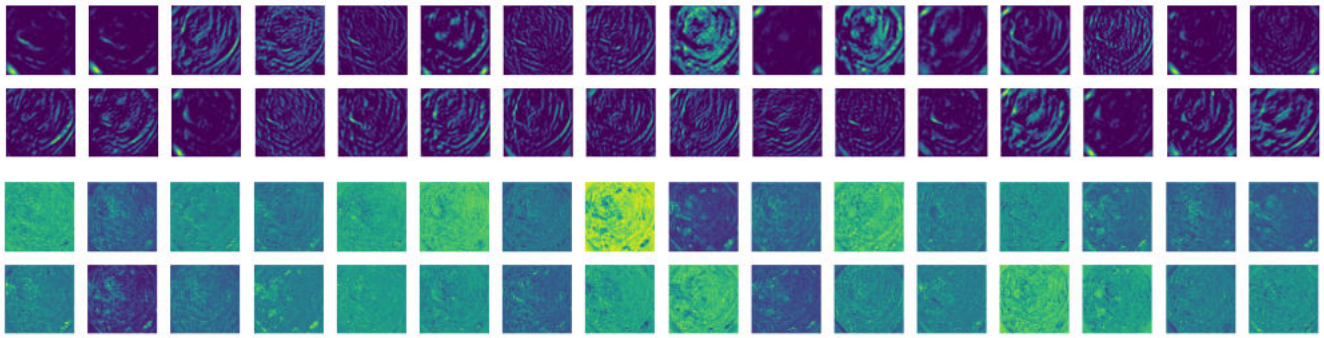

Fig. 4: An example of feature maps from a deep hidden layer of the proposed architecture. The first two rows display feature maps from a hidden layer of the Fe-Net branch, while the last two rows exhibit feature maps from a hidden layer of the U-Net branch.

scapula, with images having a resolution of 512x512 and represented as single-channel images with Hounsfield (HU) values ranging between -3024 and 3071. It is important to highlight that this experimental dataset exhibited significant differences in scale, shape, and magnitude of values compared to previous datasets.

Given the notably large resolution of images in this setup, we opted to include an additional downsampling operation as the initial step in the networks. Specifically, in the Fe-Net branch, this constitutes the sole downsampling operation at the beginning of encoder, which is then counterbalanced by a single upsampling operation at the end of the decoder.

Similar to previous experiments, we trained both networks on various dataset splits utilizing a k-fold validation strategy with 3 folds.

Moreover, in this experiment, the FeU-Net exhibited superior performance compared to the baseline, particularly notable in the case of the scapula, where we observed a 1% improvement in the Dice metric. Conversely, the improvement was more modest for the humerus, standing at just 0.3%. The results

obtained are summarized in Table III.

As observed in previous examples from the Kvasir dataset, we noticed that the improvements obtained were particularly noticeable along the borders of segmented structures which is the area of greatest interest in medical image segmentation. A clear illustration of our network's performance can be seen in Fig. 5. The segmentation masks generated by FeU-Net exhibit greater accuracy in capturing fine structures compared to U-Net. Furthermore, even when two anatomical structures are in close proximity, our experimental network demonstrates a greater ability to separate them accurately.

Network	Scapula	Humerus	N. Parameters
U-Net	0.959	0.970	10M
FeU-Net (ours)	0.969	0.973	10M

TABLE III: Results expressed through the Dice score metric on shoulder CT scan dataset.

IV. Conclusions

We introduced a novel network called FeU-Net, which combines the standard undercomplete architecture of U-Net with an overcomplete structure, Fe-Net. Fe-Net is specifically engineered to capture fine edges and small anatomical structures that are often overlooked by other methods. Unlike previous approaches, we achieved this objective while maintaining image resolution at each stage of the Fe-Net branch, while simultaneously reducing kernel sizes to decrease their receptive field.

Additionally, we proposed a novel fusion strategy involving resizing and a squeeze and excite block, enabling an effective exchange of complementary features between the two networks.

The proposed method not only surpasses the undercomplete baseline upon which it is built, but also matches the performance of a state-of-the-art network with four times the parameters. This improvement holds across different complex datasets containing both small and large segmentation masks. In future research, our goal is to further explore the capabilities of our network by refining the functionality of the fusion block and carefully fine-tuning the architecture. We aim to conduct a thorough comparison with existing methods at higher resolutions, with a specific emphasis on evaluating differences, particularly regarding small structures.

Additionally, we intend to integrate the Fe-Net branch into other architectures to assess its ability to enhance more complex networks.

Acknowledgements

This work was funded by the European Union – NextGenerationEU, National Recovery and Resilience Plan (NRRP) M4C2 Inv. 3.3 D.M. 352/2022. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

This research has been funded by the Ministero delle Imprese e del Made in Italy through a grant named Accordi per l'Innovazione, for the project titled 'SHIAB'

References

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[2] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022.

[3] Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. Using duck-net for polyp image segmentation. *Scientific Reports*, 13(1):9803, 2023.

[4] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[6] Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.

[7] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[9] Edward Sanderson and Bogdan J Matuszewski. Fcn-transformer feature fusion for polyp segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, pages 892–907. Springer, 2022.

[10] Aniruddh Sikdar, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. Deepmao: Deep multi-scale aware overcomplete network for building segmentation in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 487–496, 2023.

[11] Tijmen Tieleman and G Hinton. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Technical report*, 2017.

[12] Jeya Maria Jose Valanarasu, Vishwanath A Sindagi, Ilker Hacihaliloglu, and Vishal M Patel. Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 363–373. Springer, 2020.

[13] 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.

[14] Hasib Zunair and A Ben Hamza. Sharp u-net: Depthwise convolutional network for biomedical image segmentation. *Computers in Biology and Medicine*, 136:104699, 2021.

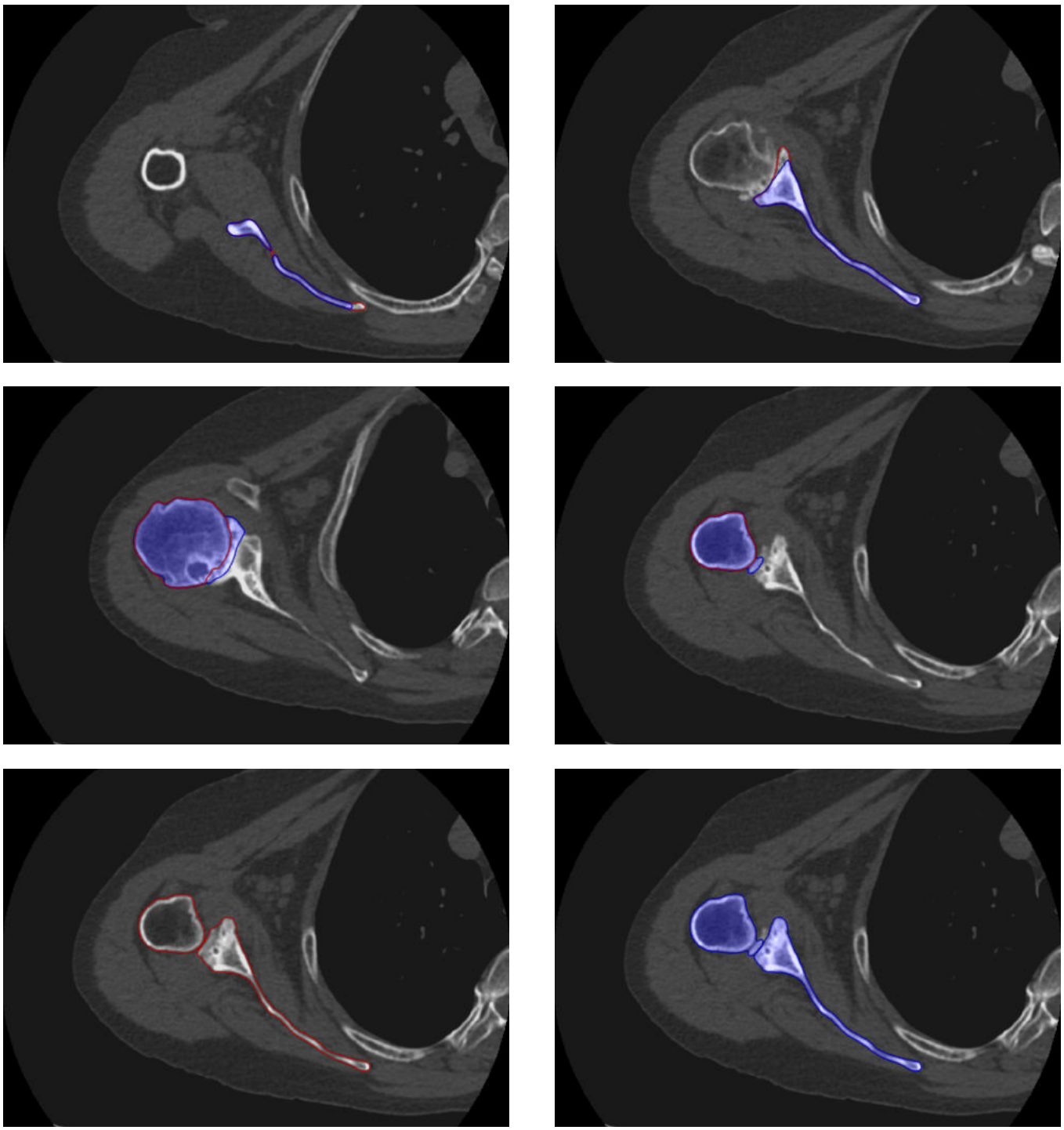

Fig. 5: Different performance of U-Net and FeU-Net on scapula and humerus (axial-view). The mask made by the U-Net is filled and displayed in blue, while the mask made by the FeU-Net is empty and displayed in red.