Medical Image Segmentation Review: The Success of U-Net

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof

(Survey Paper)

***Abstract*—Automatic medical image segmentation is a crucial topic in the medical domain and successively a critical counterpart in the computer-aided diagnosis paradigm. U-Net is the most widespread image segmentation architecture due to its flexibility, optimized modular design, and success in all medical image modalities. Over the years, the U-Net model has received tremendous attention from academic and industrial researchers who have extended it to address the scale and complexity created by medical tasks. These extensions are commonly related to enhancing the U-Net's backbone, bottleneck, or skip connections, or including representation learning, or combining it with a Transformer architecture, or even addressing probabilistic prediction of the segmentation map. Having a compendium of different previously proposed U-Net variants makes it easier for machine learning researchers to identify relevant research questions and understand the challenges of the biological tasks that challenge the model. In this work, we discuss the practical aspects of the U-Net model and organize each variant model into a taxonomy. Moreover, to measure the performance of these strategies in a clinical application, we propose fair evaluations of some unique and famous designs on well-known datasets. Furthermore, we provide a comprehensive implementation library with trained models. In addition, for ease of future studies, we created an online list of U-Net papers with their possible official implementation.**

***Index Terms*—Convolutional neural network, deep learning, medical image segmentation, transformer, U-Net.**

I. Introduction

IMAGE segmentation, defined as the partitioning of the entire image into a set of regions, plays a vital role in a wide range of medical applications. Automated segmentation facilitates clinical workflows by reducing the data processing time and supports clinicians by providing quantitative measures of organs or pathologies. Semantic segmentation is a preparatory step in automatic image processing techniques and can further enhance the segmentation quality by removing unwanted objects and by detecting specific regions (localization) that are more relevant to the task on hand (e.g., organ/lesion boundary delineation) [1].

Image segmentation tasks can be classified into two categories: semantic segmentation and instance segmentation [2], [3]. Semantic segmentation is a pixel-level classification that assigns corresponding categories to all the pixels in an image, whereas instance segmentation also needs to identify different objects within the same category based on semantic segmentation. Designing segmentation methods to distinguish organ or lesion pixels requires task-specific image data to provide the appropriate critical ground truth. Common medical imaging modalities for acquiring data are X-ray, Positron Emission Tomography (PET), Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Ultrasound (US) [4]. Early traditional approaches to medical image segmentation mainly focused on edge detection, template matching techniques, region growing, graph cuts, active contour lines, machine learning, and other mathematical methods [5].

An initial attempt at semantic segmentation using a deep neural network was proposed in [6]. This approach passes the input images through the convolutional encoder to produce the latent representation. Then, on top of the generated feature maps, the fully connected layers are included to produce a pixel-level prediction. The main limitation of this architecture was the use of fully connected layers, which depleted the spatial information and consequently degraded the overall performance. Long et al. [7] proposed Fully Convolutional Networks (FCNs) to address this limitation. The FCN structure applies several

Manuscript received 15 May 2023; revised 12 May 2024; accepted 18 July 2024. Date of publication 21 August 2024; date of current version 5 November 2024. Recommended for acceptance by J.A. Schnabel. *(Reza Azad and Ehsan Khodapanah Aghdam contributed equally to this work.) (Corresponding author: Dorit Merhof.)*

Reza Azad is with the Faculty of Electrical Engineering, Information Technology, RWTH Aachen University, 52074 Aachen, Germany, and also with the Faculty of Informatics and Data Science, University of Regensburg, 93053 Regensburg, Germany.

Ehsan Khodapanah Aghdam is with Independent Researcher, Tabriz 51368, Iran.

Amelie Rauland and Yiwei Jia are with the Faculty of Electrical Engineering, Information Technology, RWTH Aachen University, 52074 Aachen, Germany.

Atlas Haddadi Avval is with the School of Medicine, Mashhad University of Medical Sciences, Mashhad 9177899191, Iran.

Afshin Bozorgpour and Sanaz Karimijafarbigloo are with the Faculty of Informatics and Data Science, University of Regensburg, 93053 Regensburg, Germany.

Joseph Paul Cohen is with the Center for Artificial Intelligence in Medicine & Imaging, Stanford University, Palo Alto, CA 94304 USA.

Ehsan Adeli is with Stanford University, Stanford, CA 94305 USA.

Dorit Merhof is with the Faculty of Informatics and Data Science, University of Regensburg, 93053 Regensburg, Germany, and also with the Fraunhofer Institute for Digital Medicine MEVIS, 28359 Bremen, Germany (e-mail: dorit.merhof@ur.de).

All information is gathered in a GitHub repository https://github.com/NITR098/Awesome-U-Net.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3435571, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3435571

convolutional blocks consisting of the convolution, activation, and pooling layers on the encoder path to capture semantic representation and similarly uses the convolutional layer along with the up-sampling operation in the decoding path to delineate the object of interest's boundary by providing a pixel-level prediction. The main motivation underlying the successive up-sampling process on the decoding path was to gradually increase the spatial dimension for a fine-grained segmentation result. Inspired by the architecture of FCNs and the encoder-decoder models, Ronneberger et al. develop the U-Net [8] model for biomedical image segmentation. It is tailored to practical use in medical image analysis and can be applied in a variety of modalities, including CT [9], [10], [11], MRI [12], [13], US [14], [15], X-ray [16], [17], Optical Coherence Tomography (OCT) [18], [19], and PET [20].

FCN networks, including the U-Net, can efficiently exploit a limited number of annotated samples by their symmetrical design with skip connection paths and leveraging data augmentation (e.g., random elastic deformation) to extract detailed features of images without the need for extremely large training datasets, resulting in good segmentation performance [21]. The U-Net network is composed of two parts. As Fig. 2 illustrates, the first part is the contracting path that employs the down-sampling module consisting of several convolutional blocks to extract semantic and contextual features. In the second part, the expansive path applies a set of convolutional blocks equipped with the upsampling operation to gradually increase the spatial resolutions of the feature maps, usually by a factor of two, while reducing the feature dimensions to produce the pixel-wise classification score. The most significant and important part of U-Net is the skip connections which copy the outputs of each stage within the contracting path to the corresponding stages in the expansive path. This novel design propagates essential high-resolution contextual information along the network, encouraging the network to re-use the high-level representation and image-level context for accurate localization. This novel structure has become the backbone in the field of medical image segmentation since 2015, and several variants of the model have been derived to progress the state of the art based on it. The auto-encoder design of U-Net makes it a unique tool for a variety of applications, e.g., image synthesis [22], [23], [24], image denoising [25], [26], [27], image reconstruction [28], [29], and image super-resolution [30]. Besides U-Net architecture, several distinctive architectures such as DeepLab families [31] and SegNet [32] are also demonstrated significant performance in semantic segmentation tasks. However, due to its simplicity, modular design, and potential, it achieved the community's remarkable attention and is considered standard in the medical image segmentation context.

Our review covers the most recent U-Net-based medical image segmentation literature and discusses more than a hundred methods proposed until September 2022. We provide a comprehensive review and explain various aspects of these methods, including network architecture enhancements concerning the vanilla U-Net, medical image data modalities, loss functions, evaluation metrics, and the practical use cases of each category. We propose a summary of highly cited approaches in our taxonomy according to the rapid developments in U-Net and its variants. We group the U-Net variants into the following categories:

1) Skip Connection Enhancements
2) Backbone Design Enhancements
3) Bottleneck Enhancements
4) Transformers
5) Rich Representation Enhancements
6) Probabilistic Design

The key contributions of this review paper can be outlined as follows:

- This review covers the most recent literature on U-Net and its variants for medical image segmentation problems and overviews more than 100 segmentation algorithms proposed till September 2022, grouped into six categories.
- We provide a comprehensive review and insightful analysis of different aspects of U-Net-based algorithms, including the refinement of base U-Net architectures, training data modality, loss functions, evaluation metrics, and their critical contributions.
- We provide comparative experiments of some reviewed methods on popular datasets and offer codes and pre-trained weights on GitHub.

II. Taxonomy

This section suggests a taxonomy that organizes different approaches presented in the literature to modify U-Net architecture for medical image segmentation. Due to the modular design of U-Net, we proposed our taxonomy to cope with the inheritance design of U-net rather than the conceptual taxonomies offered in [33]. Furthermore, this property makes it difficult to fit each study into only one group, so a method may belong to several groups of divisions. Fig. 1 depicts our structure for taxonomy, and we think this taxonomy helps the field be organized and even motivational for future research. In Section III, we will go through each concept of taxonomy. In the remainder of this section, we will first explain the naive 2D U-Net, and following that, we will introduce the 3D U-Net. Eventually, we will elaborate on the importance of the U-Net model from a clinical perspective.

A. 2D/3D U-Net

In 2015, Ronnebreger et al. [8] proposed a new architecture with respect to Long et al.'s [7] FCN framework in conjunction with *ISBI cell tracking challenge*, where they won the competition by a large margin. Fig. 2 shows the structure of the U-Net model. Their proposed method is a cornerstone in a few attitudes those days. First, it is based on a fully convolutional network in an encoder-decoder design with insufficient data than the DNNs instinct with some intuitive data augmentation techniques. Second, their model was reasonably fast and outperformed other methods in the challenge. The model architecture can be divided into two parts: The first part is the contracting path, also known as the encoder path, where its purpose is to capture contextual information. This path consists of repeated blocks, where each block contains two successive 3×3 convolutions, followed

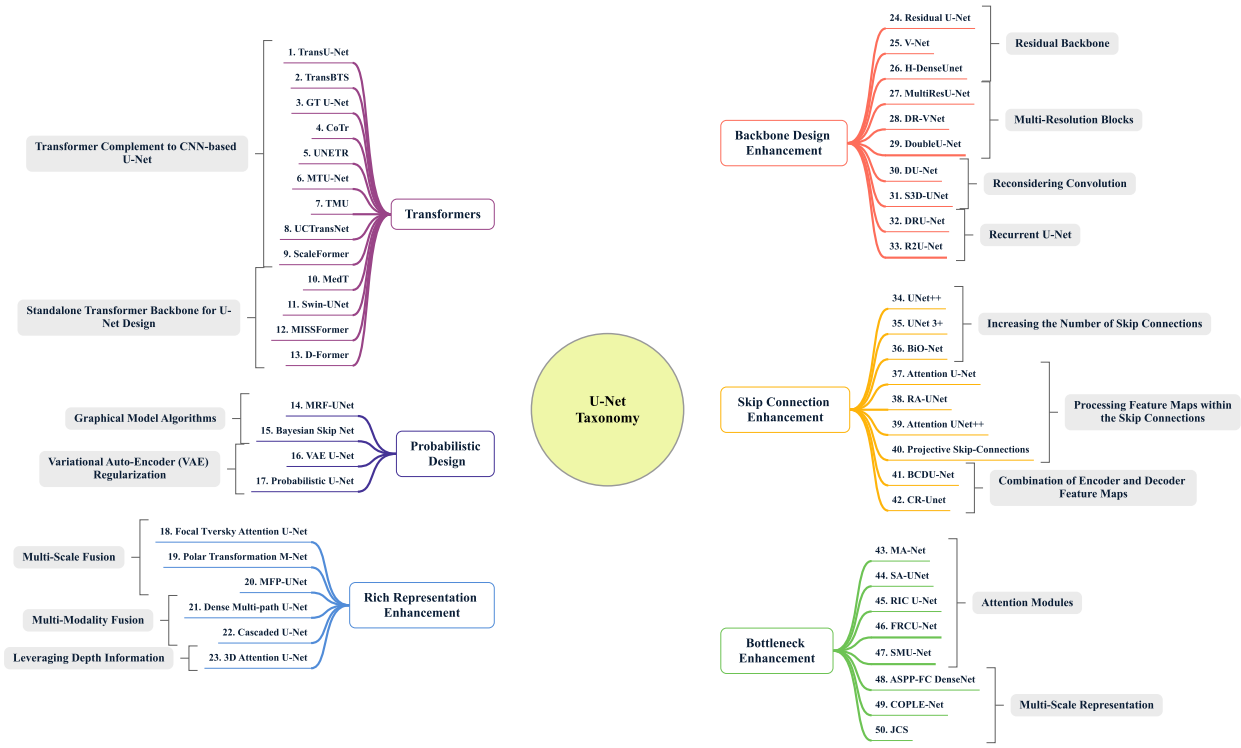


Fig. 1. The proposed U-Net taxonomy categorizes different extensions of the U-Net model based on their underlying design idea. More specifically, our taxonomy takes into account the modular design of the U-Net model and shows where the improvement happens (e.g., skip connection). Due to the clarification and unity in the studies' denomination, we may utilize some brevities. In this case, each prefix number denotes 1. [34], 2. [35], 3. [36], 4. [37], 5. [38], 6. [39], 7. [40], 8. [41], 9. [42], 10. [43], 11. [44], 12. [45], 13. [46], 14. [47], 15. [48], 16. [49], 17. [50], 18. [14], 19. [51], 20. [52], 21. [53], 22. [54], 23. [55], 24. [56], 25. [57], 26. [58], 27. [13], 28. [59], 29. [60], 30. [61], 31. [12], 32. [62], 33. [26], 34. [63], 35. [64], 36. [65], 37. [66], 38. [67], 39. [68], 40. [69], 41. [70], 42. [71], 43. [72], 44. [73], 45. [74], 46. [75], 47. [76], 48. [77], 49. [78], 50. [79].

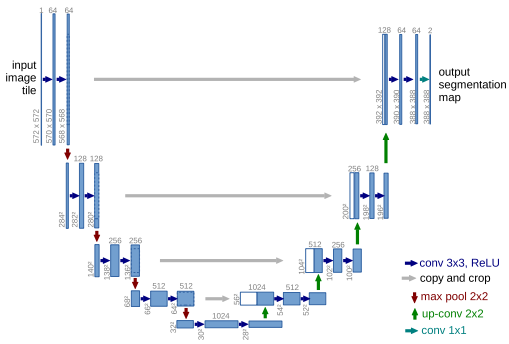


Fig. 2. The initial 2D U-Net architecture that is designed to cope with semantic segmentation challenge. Figure from [81].

by a ReLU activation function and max-pooling layers. The max pooling layer is also included to gradually increase the receptive field of the network without imposing an additional computational burden.

The second part is expanding the path, also called the decoder path, where it aims to gradually up-sample feature maps to the desired resolution. This path consists of one 2×2 transposed convolution layer (up-sampling), followed by two consecutive 3×3 convolutions and a ReLU activation. The connection path between encoder and decoder paths (also known as a bottleneck) includes two successive 3×3 convolutions followed by a ReLU activation. The successive convolutional operations included in the U-Net model enables the network's receptive field size to be increased linearly. This process makes a network gradually learn coarse contextual and semantic representation in deep layers compared to shallow layers. Learning high-level semantic features makes the network slowly lose localization of extracted features, where this aspect is essential to reconstruct segmentation results. Ronnebreger et al. presented skip connections from the encoder path to the decoder path on the same scales to overcome this challenge. The existential reason for these skip connections is to impose localization information of extracted semantic features at the same stage from the encoder. To this end, the connection module concatenates low-level features coming from the encoder path with high-level representation derived from the decoding path to enrich localization information. Eventually, the network uses a 1×1 convolution to map the final representation to the desired number of classes. To mitigate the loss of contextual information in the missing image's border pixels, the U-Net model uses an overlap tile strategy. In addition, to deal with insufficient training data a typical data augmentation technique such as rotation, and gray-level intensities invariance, elastic deformation is utilized. It should be noted that elastic deformation is a common strategy to make the model resistant to deformations, a common variation in tissues. From a practical perspective, the original U-Net model outperformed a sliding-window convolutional network [6] in warping error terminology

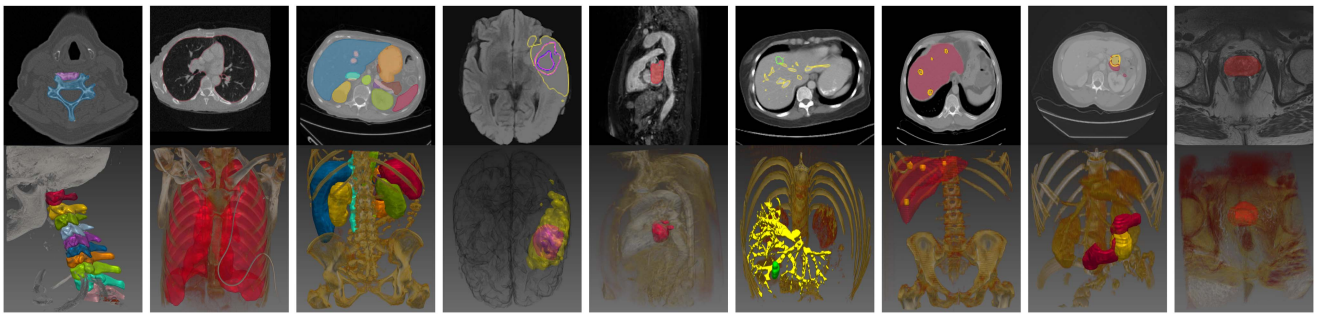


Fig. 3. Sample of the 3D medical dataset and a single selected 2D frame, where the target area (e.g., organ) is highlighted using the annotation mask. c.1) Cervical spine [82], c.2) Lung [83], c.3) Fourteen abdominal organs [84], c.4) Brain [85], [86], c.5) Heart [85], c.6) Hepatic vessel [85], c.7) Liver [85], c.8) Pancreas [85], c.9) Prostate [85].

in the *EM segmentation challenge* dataset [8]. This network also became a new state-of-the-art on two other cell segmentation datasets, *PhC-U373* and *DIC-Hela* cells, by a large margin of approximately 9% and 31% from the previous best methods in the *ISBI Cell Tracking Challenge 2015* by reporting Intersection over Union (IoU) metric [8].

Due to the abundance and representation power of volumetric data, most medical image modalities are three-dimensional. So, Çiçek et al. [80] proposed a 3D volumetric-based U-Net not only to pay attention to this need but also to overcome the time-consuming slice-by-slice annotation process for data. As it is noticeable that neighboring slices share the same information, there is no need for this much data redundancy. In [80], they replaced all 2D operations in U-Net architecture with the equivalent 3D companions and embedded a batch normalization layer for faster convergence after each 3D convolution layer. Fig. 3 shows samples of 2D and 3D medical image segmentation challenges designed for different tasks. It can be seen that the 3D data provides more comprehensive information regarding the tissue and tumors, however, compared to the 2D data it has a more computational cost.

B. Clinical Importance of U-Net

Automating redundant tasks that require an expert to review large images for small features is a very successful clinical application of the U-Net. In this direction, Fitzke et al. [87] applied a large-scale segmentation network to count specific cells in pathological images. They explicitly indicate that detecting cancerous cells from histopathological images is a challenging task that relies on the experiences of the expert pathologist. They found 21.9% of cases that used OncoPetNet [87] led to a change in tumor grading compared to human expert evaluation.

Regarding diagnostic support systems, the works of De Fauw et al. [88] as well as Bernard et al. [89] comprise important research contributions. De Fauw et al. [88] utilized a 3D U-Net for their two-stage network in risk assessment and referral suggestion network for retinal diseases as a segmentation network to produce tissue maps from dense sparse annotations. They address two critical issues with clinically developed networks' generalizability (concerning versatile imaging devices for optical coherence tomography and image acquisition processes) and inter-patient pathological differences by providing each scan's five different segmentation maps. The overall ensembled classification result for referral suggestion outperformed the eight clinical site specialists' performance.

Bernard et al. [89] analyzed the performance of ten architectures that were involved in the "Automatic Cardiac Diagnosis Challenge (ACDC)" workshop held in conjunction with the *2017 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Nine of the ten teams used U-shaped structures, while the team ranked first in terms of the Dice score and Hausdorff distance using a 2D U-Net architecture. The main point in this challenge is the utilization of some clinical parameters to assess the performance of a pipeline. The results also qualified with two independent experts' annotations, which demonstrated the models can outperform the clinician's performance in cardiac diagnosis.

Radiotherapy is a standard treatment for various cancers [90]. Thus, the efficacy and safety of organ-specific radiotherapy rely on the precise localization of organs at risk and tumors. However, because this process is mostly done manually by a radiographer or dosiometrist, the findings may be inconsistent and imperfectly correct, resulting in significant practitioner variance. Nikolov et al. [91] addressed these issues by applying an enhanced 3D U-Net structure on head and neck CT scans for automatic segmentation and compared the resulting segmentation contour with a human-crafted one in terms of new surface Dice similarity coefficient, which is a thresholded metric over standard Dice score. The difference between oncologist performance and the trained model endorsed that the U-Net-based structure can perform on par with a human expert while saving considerable time [92].

Tumor detection and growth monitoring are critical in medicine to support diagnosis, to evaluate cancer progression and inform treatment decisions. Automated segmentation may help physicians and oncologists to decide when to initiate or adjust treatments such as chemotherapy, radiation therapy, or surgery. Kickingereder et al. [93] studied brain tumor growth by a variant of 3D U-Net [94] compared with the conventional clinical procedure applied to MRI scans. They find that "[t]ime to progression from quantitative ANN-based assessment of tumor response was a significantly better surrogate endpoint than central RANO assessment." RANO constitutes a widely

recognized medical modality for evaluating treatment response in the context of neuro-oncology, particularly with regard to brain tumor therapies.

As a proxy for the future clinical impact of the U-Net, over the last few years, the number of international image analysis competitions have a high demand for automatic segmentation methods, accounting for 70% [95], which is mainly supported by the health startups and companies active in the field of health to utilize the successful algorithms for their need.

III. U-Net Extensions

U-Net is a ubiquitous network according to its approximately 48 thousand citations during its first release in 2015. This is evidence that it can handle diverse image modalities in broad domains and not only in medical fields. From our sight, the core advantage of U-Net is its modular and symmetric design, which makes it a suitable choice for broad modification and collaboration with diverse plug-and-play modules to increase performance. Therefore, by pursuing this cue, we infringe the Ronneberger et al. [8] network to modular improvable counterparts besides solid auxiliary modification for achieving SOTA or par with segmentation performances. In this respect, we offer our taxonomy (Fig. 1) and divide the diverse variants of U-Net modifications into systematic categories. This taxonomy aims to provide comprehensive and practical information for both vendors and researchers. In the following parts of this section, each category will be extensively discussed along with relevant papers.

A. Skip Connection Enhancements

Skip connections are an essential part of the U-Net architecture as they combine the semantic information of a deep, low-resolution layer with the local information of a shallow, high-resolution layer. This section provides a definition of skip connections and explains their role in the U-Net architecture before introducing extensions and variants of the classic skip connection used in the original U-Net. Skip connections are defined as connections in a neural network that do not connect two following layers but instead skip over at least one layer. Considering a two-layer network a skip connection would connect the input directly to the output, skipping over the hidden layer [96]. In image segmentation, skip connections were first used by Long et al. in [7]. At the time, the most common use of convolutional networks was for image classification tasks which only have a single label as output. In a segmentation task, however, a label should be assigned to each pixel in the image adding a localization task to the classification task. Long et al. [7] added additional layers to a usual contracting network using upsampling instead of pooling layers to increase the resolution of the output and obtain a label for every pixel. Since local, high-resolution information gets lost in the contracting part of the network it cannot be completely recovered when upsampling these volumes. To combine the deep, coarse semantic information with the shallow fine appearance information they add skip connections that connect up-sampled lower layers with finer stride with the final prediction layer.

In the original U-Net architecture by Ronneberger et al. [8] each level in the encoder path is connected to the corresponding same-resolution level in the decoder path by a skip connection to combine the global information describing what with the local information resolving where. The difference to the above approach is not only the higher number of skip connections but also the way in which the features are combined. Long et al. [7] up-sampled feature maps from earlier layers to the output resolution and added them to the output of the final layer. Ronneberger et al. [8] concatenated the features of the corresponding encoder and decoder level and process them together by passing them through two convolutional layers and an up-sampling layer together.

Li et al. [58] conducted an ablation study on skip connections by training a dense U-Net with and without skip connections. The results clearly show that the network with the skip connections generalizes better than the network without skip connections. Over the following years, many variants and extensions of the original U-Net architecture were developed concerning the skip connections [97], [98]. The following sections will present different types of extensions dealing with processing the encoder feature maps passed through the skip connections, combining the two sets of feature maps, and extending the number of skip connections.

1) Increasing the Number of Skip Connections: In 2020 Zhou et al. [63] introduced the UNet++ in which they redesign skip connections to be more flexible and therefore exploit multiscale features more effectively. Instead of restricting skip connections to only aggregate features that have the same scale in the encoder and decoder path, they redesign them in such a way that features of different semantic scales can be aggregated [63]. They argue that there has been no proof so far that encoder and decoder feature maps at the same scale are the best match for feature fusion and therefore design a more flexible setup. In their approach, they tackle two problems simultaneously. Since the optimal depth of a U-Net is unknown apriori and usually has to be determined through an exhaustive search, they incorporate U-Nets of different depths into one architecture. In this architecture, all the U-Nets share the same encoder but have their own decoder. Instead of only passing the same-scale encoder feature maps through the skip connections, each node in the decoder is also presented with the feature maps of the same-level decoders of the U-Nets with a lower depth. It can then be learned during training, which of the presented feature maps should ideally be used for the segmentation.

Huang et al. [64] take the dense skip connections introduced in the UNet++ one step further by introducing full-scale skip connections in their architecture the UNet 3+. They argue that both the original U-Net with plain skip connections between same-level encoder and decoder nodes and the UNet++ with the dense and nested skip connections do not sufficiently explore features from full scales making it challenging for the network to learn the position and boundary of an organ explicitly. To overcome this limitation they connect each decoder level with all encoder levels and all preceding decoder levels.

Since not all feature maps arriving at a decoder node through skip connections have the same scale, higher-resolution encoder

feature maps will be downscaled using a max-pooling operation and lower-resolution feature maps coming from intra-decoder skip connections will be upsampled using bilinear upsampling. Additionally, apart from the up- or down-sampling operation, each skip connection is equipped with a 3×3 convolutional layer calculating 64 output maps. The 64 feature maps arriving through each skip connection are stacked and the stack of feature maps is passed through another convolutional layer, followed by batch normalization and a ReLU activation before being further processed in the respective decoder node.

Instead of increasing the number of forward skip connections, Xiang et al. [65] add additional backward skip connections: Their Bi-directional O-Shape network (BiO-Net) is a U-Net architecture with bi-directional skip connections. This means that there are two types of skip connections:

- The forward skip connections are known from the original U-Net architecture, combining encoder and decoder layers at the same level. These skip connections preserve the low-level visual features from the encoder and combine them with the semantic decoder information.
- The backward skip connections pass decoded high-level features from the decoder back to the same level encoder. The encoder can then combine the semantic decoder features with its original input and flexibly aggregate the two types of features.

Together these two types of skip connections build an O-shaped recursive architecture that can be traversed multiple times to receive improved performance.

2) Processing Feature Maps Within the Skip Connections: In the attention U-Net established by Oktay et al. [66], attention gates (AGs) are added to the skip connections to implicitly learn to suppress irrelevant regions in the input image while highlighting the regions of interest for the segmentation task at hand. In biomedical imaging, when organs to be segmented show high inter-patient variation in terms of shape and size, a common approach is to use a cascaded network. The first network extracts a rough region of interest (ROI) including the organ to be segmented and the second network predicts the exact organ segmentation in this ROI. These approaches, however, suffer from redundant model parameters and high computational resources. Adding attention gates to the skip connections maintains a high prediction accuracy without the need for an external organ localization model. It is therefore trainable from scratch and introduces no significant computational overload and only a few additional model parameters. The output of an AG is the elementwise multiplication of the input feature maps with attention coefficients $\alpha_i \in [0, 1]$ as $\hat{\mathbf{x}}^l_{i,c} = \mathbf{x}^l_{i,c} \cdot \alpha^l_i$. For the computation of the attention coefficients both the input feature maps x, that have been passed through the skip connection from the encoder and the gating signal g are analyzed. Here, the gating signal is collected from a coarser scale for adding contextual information. The applied additive attention is formulated as follows:

$$q^l_{\text{att}} = \psi^T(\sigma_1(W_x^T \mathbf{x}^l_i + W_g^T g_i + b_g)) + b_\psi,$$
$$\alpha^l_i = \sigma_2(q^l_{\text{att}}(\mathbf{x^l_i}, g_i; \Theta_{\text{att}})), \tag{1}$$

where σ_1 and σ_2 are ReLU and sigmoid activations respectively, $W_x \in \mathbb{R}^{F_l \times F_{int}}$, $W_g \in \mathbb{R}^{F_g \times F_{int}}$ and $\psi \in \mathbb{R}^{F_{int} \times 1}$ are linear transforms and b_g and b_ψ are bias terms. Adding an AG to a skip connection, therefore, highlights the ROIs in the feature maps from the encoder path before they are concatenated with the feature maps of the decoder path. So in addition to adding higher resolution information, additional information on the location of the object(s) to be segmented is added, eliminating the need for cascaded multi-network approaches.

The attention UNet++ by Li et al. combines the attention U-Net with the UNet++ [68]. Attention gates as described in [66] are added to all the skip connections of the UNet++ with its nested U-Nets and dense skip connections. With similar motivation, Jin et al. [67] introduced a 3D U-Net with attention residual modules in the skip connections, called the RA-UNet. The network was developed for the task of segmenting tumors in the liver. The main difficulties of this task lie in the large spatial and structural variability, low contrast between liver and tumor, and similarity to nearby organs. The added attention residual learning mechanism in the skip connections improve the performance by focusing on specific parts of the image as claimed by the authors. The output of the attention module ($\mathbf{OA}$) in the RA-UNet structure is formulated as:

$$\mathbf{OA}(\mathbf{x}) = (1 + \mathbf{S}(\mathbf{x}))\mathbf{F}(\mathbf{x}), \tag{2}$$

where $\mathbf{S}(\mathbf{x})$ originates from the soft mask branch and has values in [0,1] to highlight important features and suppress noise and redundant features in the original feature maps $\mathbf{F}(\mathbf{x})$ passed through the trunk branch. The soft mask branch itself uses a residual encoder-decoder architecture to calculate its output.

To improve performance on the difficult task of the ovary and follicle segmentation from ultrasound images, Li et al. [71] added spatial recurrent neural networks (RNNs) to the skip connections of a U-Net. Since there are usually many small follicles in an image, it is very likely that the neighboring follicles are spatially correlated. In addition, there might be a possible spatial correlation between the follicles and the ovary. As the max-pooling operation in the original U-Net brings a loss of spatially relative information the spatial RNNs should improve the segmentation results by learning multi-scale and long-range spatial contexts.

Li et al. [71] built the spatial RNNs from plain RNNs with a ReLU activation. Each spatial RNN module takes feature maps as input and produces spatial RNN features as output. It uses four independent data translations to integrate local spatial information in up, down, left, and right directions. The maps from each direction are concatenated and passed through a 1×1 convolutional layer to produce feature maps where each point contains information from all four directions. The process is then repeated to extend the local spatial information to global contextual information. The final feature maps passed through the skip connection are a combination of the original encoder feature maps and the RNN features extracted from these maps. The authors claim that the architecture is especially strong at avoiding the segmentation of false positives and detecting and segmenting very small follicles. A limitation of the RNN modules is that they make training more difficult and computationally expensive. To

compensate for this, Li et al. added deep supervision. While most medical applications demand segmentations to be in the same dimension as the input image, there are also medical protocols that require segmentation of the image projection, e.g., Liefers et al. [99] studied the retinal vessel segmentation as a 2D $\rightarrow$ 1D retinal OCT segmentation task. This adds the problem of dimensionality reduction to the segmentation. Lachinov et al. [69] introduced a U-Net with projective skip connections to handle ND $\rightarrow$ MD segmentations, where $M < N$.

The encoder is a classic U-Net encoder with residual blocks. The decoder however only restores the input resolution for the M dimensions of the segmentation. The remaining reducible dimensions $M < d \leq N$ are left compressed. This means that the sizes of the encoder and decoder feature maps no longer match which is why Lachinov et al. [69] introduce the projective skip connections. The encoder feature maps passed along the projective skip connections are processed by an average pooling layer with varying kernel size so that the dimensions which are not present in the segmentation are reduced to the size they have in the bottleneck. This way they can be concatenated with the corresponding decoder feature maps. Global Average Pooling (GAP) and a convolutional layer are added after the last decoder level to calculate the final MD segmentation.

3) Combination of Encoder and Decoder Feature Maps: Another extension of the classic skip connections is introduced in the BCDU-Net by Azad et al. [70] where a bi-directional convolutional long-term-short-term-memory (LSTM) module is added to the skip connections. Azad et al. argue that a simple concatenation of the high-resolution feature maps from the encoder and the feature maps extracted from the previous up-convolutional layer containing more semantic information might not lead to the most precise segmentation output. Instead, they combine the two sets of feature maps with non-linear functions in the bi-directional convolutional LSTM module. Ideally, this leads to a set of feature maps rich in both local and semantic information. This method uses two ConvLSTMs, processing the input data in two directions in the forward and backward paths. The output will be determined by taking into consideration the data dependencies in both directions. In contrast to the approach by Li et al. [71], where only the encoder feature maps are processed by the RNN and then concatenated with the decoder features, this approach processes both sets of feature maps with the RNN.

B. Backbone Design Enhancements

Apart from adapting the skip connections of a U-Net, it is also common to use different types of backbones in newer U-Net extensions. The backbone defines how the layers in the encoder are arranged and its counterpart is therefore used to describe the decoder architecture.

In the original U-Net by Ronneberger et al. [8] each level in the encoder consists of two 3×3 convolutional layers with ReLU activation followed by a max pooling operation. The number of feature maps doubles at each level. Any 2D or 3D convolutional neural networks (CNN) image classifier can be used as an encoder in a U-Net, adding its mirrored counterpart as the decoder. Dozens of studies modified the vanilla U-Net main blocks to broaden the receptive fields of convolution operations and extract rich and fine-grained semantic representations for challenging multi-class problems, e.g., [60], [100], [101], [102], [103]. This section presents several prominent backbones used in the U-Net architecture and explains their benefits and downsides.

1) Residual Backbone: A very common backbone for the U-Net architecture is the ResNet initially developed by He et al. [104]. Residual networks enable deeper network architectures by tackling the vanishing gradient problem that often occurs when stacking several layers in deep neural networks, as well as a degradation problem that leads to first saturating and then degrading accuracy when adding more and more layers to a network. Residual building blocks explicitly fit a residual mapping by adding skip connections and performing an identity mapping that is added to the output of the stacked layers. In their implementation of a residual U-Net, Drozdzal et al. [56] refer to the standard skip connections in the U-Net as long skip connections and the residual skip connections as short skip connections, as they only skip ahead over two convolutional layers. Using residual blocks as the backbone in a U-Net, Drozdzal et al. [56] can build deeper architectures and find that the network training converges faster compared to the original U-Net. Milletari et al. [57] report the same findings in their 3D U-Net architecture using 3D residual blocks as the backbone.

A prominent adaption of the backbone is to exchange all 2D convolutions with 3D convolutions to process an entire image volume as can often be found in medical applications. When processing a 3D image in a slice-wise fashion using 2D convolutions, the contexts on the z-axis can not be captured and learned by the network. Using fully convolutional architecture with 3D convolutions elevates this drawback and can fully leverage the spatial information along all three dimensions. A drawback of using 3D convolutional layers as the backbone in a u-net is the high computational cost and GPU memory consumption, which limits the depth of the network and the filter's size, i.e., its field-of-view. Milletari et al. [57] fully convolutional volumetric, V-Net architecture uses 3D residual blocks as a backbone, thereby enabling fast and accurate segmentation in 3D images. The H-DenseUNet by Li et al. [58] uses two U-Nets, one with 2D-dense blocks as the backbone and the other with 3D-dense blocks as the backbone. This enables them to first extract deep intra-slice features and then learn inter-slice features in shallower volumetric architecture with a lower computational burden.

2) Multi-Resolution Blocks: To tackle the difficulty of analyzing objects at different scales, Ibtehaz et al. [13] introduce the MultiResUNet with inception-like blocks as a backbone. Inception blocks, introduced by Szegedy et al. [105], use convolutional layers with different kernel sizes in parallel on the same input and combine the perceptions from different scales before passing them deeper into the network. The two following convolutions with 3×3 kernels in the classical U-Net resemble one convolution with a 5×5 kernel. For incorporating a multi-resolution analysis into the network, 3×3 and 7×7 convolutions should be added in parallel to the 5×5 convolution. This can be achieved by replacing the convolutional layers

with inception-like blocks. Adding the additional convolutional layers increases the memory requirement and computational burden. Ibtehaz et al. [13], therefore, formulate the more expensive 5×5 and 7×7 convolutions as consecutive 3×3 convolutions. The final *MultiRes Block* is created by adding a residual connection. Instead of keeping an equal number of filters for all consecutive convolutions, the number of filters is gradually increased to further reduce the memory requirements. In the final architecture, the two consecutive 3×3 convolutions from the original U-Net are replaced by one MultiRes block, leading to faster convergence, improved delineation of faint boundaries, and higher robustness against outliers and perturbations.

Another well-known backbone for U-Net extensions is the DenseNet introduced by Huang et al. in [106]. Similarly to residual networks, the DenseNet also aims at fighting the vanishing gradient problem by creating skip connections from early layers to later layers. The DenseNet maximizes the information flow by connecting all layers with the same feature map size with each other. This means that every layer obtains concatenated inputs from all preceding layers. Contrary to what one might expect, a dense net actually requires fewer parameters compared to a traditional CNN because it does not have to relearn redundant feature maps and can therefore work with very narrow layers with e.g. only 12 filters and can learn multi-resolution features. The direct connection from each layer to the loss function implements implicit deep supervision which helps train deeper network architectures without vanishing gradients.

Karaali et al. [59] utilized Dense Residual blocks in the U-Net-like representation for retinal vessel segmentation. To this end, they were inspired by DenseNet [106], and ResNet [104] to design a Residual Dense-Net (RDN) block. In their architecture, the first sub-block comprises successive batch Normalization, ReLU, Convolution, and Dropout counterparts, which employs the dense connectivity pattern as in [106]. The following sub-block applies a residual connectivity pattern. Using a DenseNet-like backbone helps the U-Net architecture learn more relevant features using fewer parameters. The residual connectivity smooths the information flow across the layers to facilitate the optimization step.

3) Reconsidering Convolution: This direction aims to reduce the computational burden of the naive convolution operation by re-considering the alternative convolutional operations. Jin et al. [61] exchange each 3×3 convolutional layer in the original U-Net with a deformable convolutional block for the accurate segmentation of retinal vessels. Their architecture is named DUNet. The deformable convolutional blocks are inspired by the work on deformable convolutional networks by Dai et al. [107] and should adapt the receptive fields to adjust optimally to different shapes and scales of complicated vessel structures in the input features. In deformable convolutions, offsets are learned and added to the grid sampling locations normally used in the standard convolution. In a classic convolution the kernel sampling grid G would be defined as:

$$G = (-2, -2), (-2, -1), \ldots, (2, 1), (2, 2). \tag{3}$$

Considering this grid, every pixel m_0 in the output feature map $\mathbf{y}$ can be calculated as:

$$\mathbf{y}(m_0) = \sum_{m_i \in G} (w)(m_i) \cdot \mathbf{x}(m_0 + m_i), \tag{4}$$

from the input $\mathbf{x}$. In the deformable convolution, an offset Δm_i is added to the grid locations.

$$\mathbf{y}(m_0) = \sum_{m_i \in G} (w)(m_i) \cdot \mathbf{x}(m_0 + m_i + \Delta m_i). \tag{5}$$

Every deformable convolutional block consists of a convolutional layer, to learn the ideal offsets from the input. A deformable convolution layer applying the convolution with the adapted sampling points followed by batch normalization and ReLU activation. Since the calculated offset Δm_i is usually not an integer, the input value at the sampling point is determined using bilinear interpolation. Exchanging the simple convolutions with deformable convolutions helps the network adapt to different shapes, scales, and orientations but comes at a higher computational burden because an additional convolutional layer per block is needed to determine the offsets of the sampling grid.

When segmenting from 3D images it is important to make use of the full spatial information from the volumetric data. However, this is not possible with 2D convolutions and 3D convolutions are computationally very expensive. To address this problem, Chen et al. [12] used separable 3D convolutions as the backbone of the U-Net. The 3D convolution is divided into three branches where each branch represents a different orthogonal view so that the input is processed in axial, sagittal, and coronal views. Additionally, a residual skip connection is added to the separated 3D convolution. Using separable 3D convolutions as the backbone of the U-Net, Chen et al. [12] can take into consideration the full spatial information from the volumetric data in the U-Net architecture without the extremely high computational burden of standard 3D convolution.

4) Recurrent Architecture: Recurrent neural networks (RNN) are used frequently to process sequential data such as in speech recognition. Liang et al. [108] were among the first groups to design a recurrent convolutional neural network (RCNN) for images recognition. Although the input image, in contrast to sequential data, is static, each unit's activity is modulated by its neighbouring units' activities because the activities of RCNNs evolve over time. By unfolding the RCNN through time, they can obtain arbitrarily deep networks with fixed parameters. Using these RCNN blocks as the backbone of the U-Net architecture enhances the ability of the model to integrate contextual information. Alom et al. [109] used RCNN blocks as a backbone in their RU-Net architecture, ensuring better feature representation for segmentation tasks.

C. Bottleneck Enhancements

The U-Net architecture can be separated into three main parts: the encoder (contracting path), the decoder (expanding path), and the bottleneck which lies between the encoder and decoder. The bottleneck is used to force the model to learn a compressed representation of the input data which should only contain the

important and useful information needed to restore the input in the decoder. To this end, various modules are designed in multiple studies [75], [110] to recalibrate and highlight the most discriminant features. In the original U-Net, the bottleneck consists of two 3×3 convolutional layers with ReLU activation. More recent approaches however have extended the classic bottleneck architecture to improve performance.

1) Attention Modules: Several works apply attention modules in the bottleneck of their U-Net architecture. Fan et al. used a position-wise attention block (PAB) in their MA-Net to model spatial dependencies between pixels in the bottleneck feature maps with self-attention [72]. The feature maps passed into the bottleneck at the end of the encoder path are first processed by a 3×3 convolutional layer. The resulting outputs are then processed by three individual 1×1 convolutional layers producing A, B, and C. A and B are reshaped to form two vectors. A matrix multiplication of these two vectors passed through a softmax function yields the spatial feature attention map $P \in \mathbb{R}^{N \times N}$ in which the positions $p_{i,j}$ encode the influence of the i^{th} position on the j^{th} position in the feature map. Subsequently, a matrix multiplication is performed between the reshaped C and the spatial feature attention map P, and the resulting feature maps are multiplied with the input I' before being passed through a final 3×3 convolutional layer. The final output O is therefore defined as follows:

$$O_i = \alpha \sum_{i=1}^{N} (P_{ji} C_i) + I'_j, \tag{6}$$

where α is set to zero at the beginning of training and it is learned to assign more weight during the training process. Considering that the final output is the weighted sum of the feature maps across all positions and the original feature maps, it has a global contextual view and can selectively aggregate rich contextual information. Intra-class correlation and semantic consistency are improved because the PAB can consider long-range spatial dependency between features in a global view.

Guo et al. also add a spatial attention module to the bottleneck of their SA-UNet architecture [73]. The spatial attention module should enhance relevant features and compress unimportant features in the bottleneck. In their approach, the input feature maps are passed through an average pooling and a max pooling layer in parallel. Both pooling operations are applied along the channel dimension to produce efficient feature descriptors. The outputs are then concatenated and passed through a 7×7 convolutional layer and sigmoid activation to obtain a spatial attention map. By multiplying the spatial attention map with the original input features, the inputs can be weighted based on their importance for the segmentation task at hand. The attention module only adds 98 parameters to the original U-Net and is therefore computationally very lightweight.

In another work, Azad et al. [76] utilized the idea of a texture/style matching mechanism in the U-Net bottleneck for brain tumor segmentation. In their design, an attention agent is designed to distill the informative information from a full modality (four MRI modalities, T1, T2, Flair, and T1c) into a missing-modality network (only Flair). Further information regarding the missing-modality task can be found in [21]. A deep frequency attention module is proposed in [75] to perform a frequency recalibration process on the U-Net bottleneck. This attention block aims to recalibrate the feature representation based on the structure and shape information rather than texture representation to alleviate the texture bias in object recognition.

2) Multi-Scale Representation: The aim of this direction is to enhance the bottleneck design by including multi-scale feature representation, e.g. atrous convolution. The atrous convolutions are performed like standard convolutions, but with convolutional kernels with inserted holes in them. The holes are defined by setting the weight of the convolutional kernel to zero at the corresponding locations and the pattern for doing so is defined by the atrous sampling rate r. Considering a sampling rate r, this introduces $r-1$ zeros between consecutive filter values. A $k \times k$ convolutional kernel is thereby enlarged to a $k + (k-1) * (r-1) \times k + (k-1) * (r-1)$ filter. This way the receptive field of the layer is expanded without introducing any additional network parameters to be learned.

When the objects to be segmented are of very different sizes it is important for the network to extract multiscale information. Combining the ideas of spatial pyramid pooling and atrous convolutions, the feature maps in the bottleneck of the U-Net can be resampled in parallel by atrous convolutions with different sampling rates and then combined to obtain rich multiscale features.

Hai et al. [77] use atrous spatial pyramid pooling (ASPP) in the bottleneck of a U-Net architecture for the segmentation of breast lesions. The final feature maps of the encoder are passed in parallel through a 1×1 convolutional layer and three atrous 3×3 convolutional layers with atrous sampling rates of 6, 12, and 18, respectively. These four processed groups of feature maps are concatenated together with the original feature maps passed to the bottleneck and processed by a final 1×1 convolution before being passed to the decoder.

Wang et al. make use of ASPP in the bottleneck as well in their COPLE-Net for the segmentation of pneumonia lesions from CT scans of COVID-19 patients [78]. Here, four atrous convolutional layers with dilation rates of 1, 2, 4, and 6 respectively are used to process the bottleneck feature maps to capture multi-scale features for the segmentation of small and large lesions. Similarly, Wu et al. [79] proposed a multi-task learning paradigm, JCS, for COVID-19 CT image classification and segmentation. JCS [79] is a two branches architecture, which utilizes a Group Atrous (GA) module, in its segmentation branches bottleneck for feature modification. GA first applies 1×1 convolution operation to expand the channels of the feature map. Then the feature map is divided into four equal sets. Utilizing the atrous convolutions with different rates on these sets results in more global feature maps with diverse receptive fields. To fully extract more discriminant features from the final feature map, JCS adopts a squeeze and Excitation (SE) [111] block as an attention mechanism for recalibrating channel-wise convolution features.

D. Transformers

Inspired by the recent success of the Transformer models in Natural Language Processing (NLP), these models were further

extended to perform vision recognition tasks. More specifically, the Vision Transformer (ViT) model was introduced by Dosovitskiy et al. [112] to alleviate the deficiency of CNNs in capturing the long-range contextual dependencies.

Contrary to the Transformers in NLP tasks [113], the computer vision tasks usually contain more than one-dimensional data (e.g., 2D image, 3D video), which needs to be prepared for the transformer model. Hence, ViT's pipeline starts with the image sequentialization process to prepare the tokenized sequence for the encoder module. From now on, the words **patch** and **token** will be used interchangeably.

If $x \in \mathbb{R}^{H\times W\times D\times C}$ is a volumetric 3D image with a (H, W, D) spatial resolutions and C input channels, first the x is dividing into $N = \frac{H\times W\times D}{P^3}$ flattened uniform, non-overlapping patches $x_p^i \in \mathbb{R}^{N\times(P^3.C)}, i \in \{1, \ldots, N\}$ with (P, P, P) spatial resolution for each patch, therefore each patch is representing by a 1D sequence with a length of $1 \times (P^3.C)$. A linear layer applies on top of the sequence to map them to a K dimensional embedding space. Afterward, three representations were learned from the fed tensor to Transformer, namely **Q**eury (**Q**), **K**ey (**K**), and **V**alue (**V**) matrices. Mathematically, self-attention can be represented as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{K_h}}\right)\mathbf{V}, \qquad (7)$$

where $\sqrt{K_h}$ denotes a normalization factor to preserve the attention matrix (7) from the possible gradient vanishing or exploding through the training.

So far, we have concisely highlighted the ViT pipeline and the related mathematics. However, an in-depth analysis of the ViT and its functionality can be found in [114]. In the following sections, we will discuss the integration of the Transformer into the U-Net structure in medical segmentation. We categorized the presence of Transformers in U-shaped networks into two sub-categories: (a) Transformer as a complement to CNN-based U-Net-like structures and (b) U-shaped standalone Transformer architectures.

1) Transformer Complement to CNN-Based U-Net: The success of CNNs in diverse dense prediction tasks in the vision domain, e.g., segmentation, is noticeable. Their performance is underlined in their multi-scale representation and ability to capture local semantic and texture information. However, the local representation derived from the CNN architecture might not be robust enough to capture geometrical and structural information existing in the medical data. Therefore, there is a need for a mechanism to capture inter-pixel long relations to extend the performances of the existing CNN-based U-Net variants suffering from the limited receptive field of convolutional operations. Chen et al. [34] proposed one of the first studies that utilized the ViT in the U-Net structure to compensate for the U-Net's disability in long-range modeling dependencies, namely TransUNet. The stacked Transformers in the encoder path feed with the tokenized paths from abstract features extracted within the primal input to extract global contexts. The decoder path upsamples the encoded features combined with the high-resolution CNN feature maps to enable precise localization.

TransBTS [35] as a 3D U-Net-like architecture, modeling local and global information in spatial and slice/depth dimensions. TransBTS utilizes the 3D CNN backbone for its encoder and decoder path to capture local representation across spatial and depth dimensions. It also unleashes the high computational burden for the Transformer counterpart while using stacked ViT blocks in the bottleneck.

Li et al. [36] proposed the Group Transformer (GT) U-Net structure to address the low performance of previous segmentation methods in fuzzy boundaries while keeping the computational complexity low within the hybrid structure of CNN and the Transformer in a U-Net-like paradigm. All the main counterparts of U-Net are based on Group Transformer (GT) to dispense the quadratic computational complexity within these successive parallel convolution, Multi-Head Self-Attention (MHSA), convolution modules in each stage to gradually increase receptive filed and extracting long local-dependencies. So far, the presence of a Transformer in the segmentation tasks is crucial because if a network wants to provide an efficient prediction mask, it should be able to minimize the miss-classifying of the background and foreground pixels that leads to a reduction in False Positives (FP). Therefore learning long-range contextual features is as essential as fuzzy boundaries resulting from object overlappings or variations in exposure to medical imaging devices. To mitigate the occurrence of this miss predicting in boundary levels, GT U-Net utilizes a Fourier descriptor loss term within binary cross entropy to impose the prior shape knowledge.

CoTr [37] is a U-Net-like structure with CNN-based 3D residual blocks for encoder and decoder paths with the amalgamation of Deformable Transformer (DeTrans) for multi-scale fusion, besides the conventional skip connections from the encoder to the decoder for better localization information and faster convergence. TransUNet [34] suffers from parameters overload within MHSA, which treats all image tokenization positions equally. Therefore, CoTr instantiates the deformation concept from [107], [115] into the deformable self-attention mechanism in Transformer to decrease the computation complexity and prepare the ground for using Transformer to process multi-scale and high-resolution feature maps.

UNETR [116] is a 3D segmentation network that directly utilizes volumetric data incorporating ViT solely at the encoder stage to capture global multi-scale contextual information in a 3D volumetric style which is usually of paramount importance in medical image segmentation domain. Using a CNN-based decoder is since transformers can not capture spatial localization information well despite their excellent capability of learning global information.

In the medical image segmentation task, neighboring information of a specific region tends to be more correlated than far regions. To this end, Wang et al. [39] proposed the MT-UNet network utilized with the Mixed Transformer Module (MTM) to capture long-range dependencies wisely concerning the most neighboring contextual information. Therefore, MTM has an External Attention (EA) counterpart to address this concern. MTM is used in conjunction with a U-Net-like structure accompanied by CNN blocks. In the case of small medical datasets, CNN blocks are used to reduce the computational overhead by

downsampling the input feature maps and introducing a structure prior to the model.

Azad et al. [40] proposed a contextual attention network, namely TMU, for adaptively synthesizing the U-Net produced local feature with the ViT's global information for enhanced overlap boundary areas in medical images. TMU is two branches pipeline, wherein the first stream utilizes a U-Net-like block without a segmentation head (ResNet backbone [104]) to extract high semantic features and object-level boundary heatmap interaction representation. In the next branch, the ViT-based Transformer module applies to non-overlapping patches' of input images to extract long-range dependencies. Whereas the objective of segmentation differs from one subject to another data, as mentioned before, TMU aims to merge the local and global information adaptively. To this end, Azad et al. [40] proposed a contextual attention mechanism to produce image-level contextual information and highlight the most discriminative regions within importance coefficients delivered by attention weights from Transformer.

Skip connections in the U-Net-based model are used to transfer high spatial information from the encoder to the decoder for accurate localization, while the successive downsampling operations suffer from the loss of spatial information. However, Wang et al. [41] studied the effectiveness of the preliminary U-Net skip connections and stated that the naive skip connections suffer from the highly semantic gap, such as semantic gaps among multi-scale encoder features and between the encode-decoder stages. They proposed UCTransNet [41] that alleviates these mentioned issues from the channel perspective with an attention mechanism, namely Channel Transformer (CTrans). CTrans is a modification for skip connections and consists of two sub-counterparts, Channel Cross fusion with Transformer (CCT) and Channel-wise Cross-Attention, for aggregating multi-scale features adaptively and guiding the fused multi-scale channel-wise features to decoder effectively, respectively. CCT fuses multi-scale encoder features to adaptively compensate for the semantic gap between different scales with the advantage of long-range dependency modeling.

2) Standalone Transformer Backbone for U-Net Designs: So far, multiple studies incorporating the Transformer concept and conventional CNN modules have been reviewed in Section III-D1. In this section, we investigate the usage of a Transformer as a standalone main counterpart for designing backbones for U-Net-like structures. One of the first structures in this domain was proposed by Valanarasu et al. [43], namely MedT. However, Transformer's performance (also ViTs) has a strong bond with the fed data scale to the Transformer module [112], which on the medical scale, could be degraded more, and a high amount of data could not be available. Therefore MedT [43] proposes a gated axial-attention mechanism to control the information flow by positional embeddings to query, key, and value [117] in a multi-axis attention operation [118]. In [117], the accurate relative positional encoding learned on large-scale datasets rather than small-scale datasets caused MedT to introduce a gating parameter to control the amount of positional bias in capturing non-local information in hindering non-accurate positional embedding.

Transformers are well capable of capturing long-range dependencies through data, however, they suffer from severe and inevitable handicaps that impede them from their versatile use in vision tasks. In the vision tasks, Swin Transformer [119] plays a critical role as an efficient and linear Transformer with the capability of supporting hierarchical architectures. This intuition and the U-Net-like structure success emerged Swin-Unet [44] structure in the medical segmentation field. Cao et al. [44] used the Swin Transformer block as the main counterpart of their U-shaped network. 2D medical images split into non-overlapping patches, and each patch fed into the encoder path comprised of Swin blocks. The contextual features from the bottleneck output upsample in the decoder path with patch expanding layer (contrary to path merging layer) end couples with the multi-stage features from the encoder via skip connections to restore the spatial information.

MISSFormer network with an enhanced Transformer block as a primary entity in the network. One of the Transformer's drawbacks mentioned earlier is its un-suitability for capturing local context [120], [121], which comes with the solution for lessening the computational complexity with patching operation. However, the local contextual information plays a pivotal role in high-resolution vision tasks, therefore, some studies in the vision Transfomer domain tackle this problem by embedding convolution operations in their attention module, e.g., PVTv1 [122], PVTv2 [123], and Uformer [124]. Huang et al. [45] argues this methodology and conclude that direct usage of convolution layers in Transformer blocks limits the discrimination of features. For an input image, MISSFormer applies a 4×4 convolutions with the stride size instead of 4 (overlapping windows) for preserving local continuity in building the patches stage. The encoder path molds hierarchical representation with the help of an Enhanced Transformer Block. Afterward, the Enhanced Mix-Feed Forward Network (FFN) sub-module (modified clone of Mix FFN from [125]) aligns features and makes discriminant representation with 3×3 depth-wise convolutions for capturing the local context efficiently. Analogous to [41], MISSFormer rethinks the skip connection design, utilizes the Enhanced Transformer Context Bridge module for multi-scale information fusion, and hinders the gap between encoder and decoder feature maps. This module captures the local and global correlations between different scale features.

We reviewed multiple studies that utilized the Transformer in their U-Net pipeline in various methods. With such evidence and stunning growth in the Transformer field, we still hear about the outstanding collaboration between Transformers and U-Net-like networks so often.

E. Rich Representation Enhancements

To obtain a rich representation, the common approaches applied to medical image segmentation are multi-scale and multi-modal methods, e.g., [126], [127], [128]. The key objective is to enhance the performance of the trained models by utilizing all available information from multi-modal or multi-scale images while retaining the most desirable and relevant features.

The multi-scale method, also referred to as the pyramid method originated from the Laplace pyramid method proposed by Burt et al. [129]. The approach converts the source input image by resizing it into a series of images with decreasing spatial resolutions. This scheme allows encoders of models to directly access the features of the enhanced images of different sizes and thus learn the respective features.

The study of the organs of interest requires their specific imaging modality to provide targeted information. However, each imaging technique has its limitations and can only reveal partial details about the organ, which may lead to inaccurate clinical analysis. Therefore, a fusion of images from various imaging modalities can be conducted to supplement each other's information by integrating complementary information retrieved from several input images.

The powerful structural design of the UNet network with the encoder and decoder allows the network to mine salient features at multiple input levels and enables effective feature fusion of different modalities. Lachinov et al. [54] evaluate the performance of the Cascaded U-Net with multiple encoders processing each modality respectively to demonstrate the improvement due to the extraction multi-modal representation. The results indicate that the architecture taking the multiple modalities into account outperforms the network only relying on one single modality. The following classifications will illustrate the modality fusion proposed to learn richer representations.

1) Multi-Scale Fusion: Image pyramid input or side output layers are aggregated into U-Net structures to fuse the multi-scale information in the encoder or decoder stage.

Abraham et al. [14] propose the Focal Tversky Attention U-Net with a generalized focal loss function that modulates the Tversky index [130] to address the issue of data imbalance and improve precision and recall balance in medical image segmentation. Furthermore, they incorporate multi-scale image inputs into the attention U-Net model with deep supervision output layers [66]. The novel architecture facilitates the extraction of richer feature representations and results in 3% dice score improvement on multi-class CT abdominal segmentation task. Compared to the commonly used Dice loss, the Tversky similarity index introduces a specific weight for each class, which is inversely proportionate to the label occurrences. The authors develop a focal Tversky loss function (FTL) for better small regions-of-interest (ROIs) segmentation by forcing the function to shift the focus on less accurate and misclassified predictions.

$$FTL_c = \sum (1 - TI_c)^{1/\gamma}, \tag{8}$$

where TI_c shows the FTL and γ is set in the interval [1,3] to enable the loss function to concentrate more on incorrectly classified predictions that are less accurate. The reason is that when $\gamma > 1$, the FTL is almost unaffected if a pixel is misclassified with a high Tversky index. And if a pixel is incorrectly classified with a small Tversky index, the FTL will be high and forces the model to focus on hard samples. They use Soft Attention Gates (AGs) to prune features and propagate only relevant spatial information to the decoding layers to enhance the balance between precision and recall at a structural level. In addition, an input image pyramid injected into each of the max pooling layers in the encoder and deep supervision module enriches feature learning at different scales.

To better segment the Optic Disc (OD) and Optic Cup (OC) for accurate diagnosis of glaucoma from fundus images, Fu et al. [51] introduce the polar transformation into the U-shape convolutional network with multi-scale input layers to build the Polar Transformation M-Net, aims to extract the richer context representation of the original image in the polar coordinate system. Compared to prior work, which treats OD and OC individually while ignoring their interdependency and overlap, the proposed Polar transformation M-Net considers both OD and OC simultaneously and formulates the segmentation as a multi-label task. Moreover, a novel loss function built on the Dice coefficient is employed to address the data imbalance between OD and OC in the fundus images. The model aggregates the side-output layers serving as early classifiers that generate prediction maps at different scales.

The original U-Net may not fully exploit all contributions of the semantic strength since it only generates output segmentation maps from the final layer of the decoder path. Moreover, the output of the layers in different steps cannot be connected to one another, which blocks feature sharing and leads to redundant parameters. To address the mentioned issue, Moradi et al. proposed MFP-UNet which allows the output of all the blocks in different stages to be fed to the last layer [131]. Their architecture is composed of two pathways, the "bottom-up pathway" and the "top-down pathway". The encoder of the U-Net with dilated convolution filter serves as the "bottom-up pathway". The dilated convolutional kernel can increase the receptive fields of the module by the dilation factor. Besides, the expansion path of U-Net acts as the FPN top-down pathway. Each step of the top-down pathway provides prediction maps where lower-resolution semantically stronger features can be processed for transfer to higher resolutions. Additional convolution layers are included for processing the feature maps at different scales to one fixed resolution compared to the decoder path of the original U-Net, which can boost the accuracy and improve the resolution of each stage. According to the experiment results, the novel model provides a robust and powerful architecture regarding the capabilities of feature representation in a pyramid, which shows robustness to large and rich training sets.

2) Multi-Modality Fusion: In this section, we summarize the U-Net variant models with multimodal fusion modules, where a single encoder of U-Net is extended to multiple encoders to receive medical images in different modalities. The branches of encoders are connected by their respective strategies of aggregation, thus sharing information in different modalities, extracting richer representations, and complementing each other.

The Dense Multi-path U-Net architecture proposed by Dolz et al. [53] enhances traditional U-Net models regarding rich representation learning in two key aspects: modality fusion and inception module extension. Two typical strategies are employed to deal with multi-modal image segmentation tasks. The early fusion merges the low-level features of inputs of multiple imaging modalities at the very early stage. As for the late fusion strategy,

the CNN outputs of different modalities are fused at a later point. Nevertheless, these previous strategies cannot thoroughly model the highly complex relation of the image information across different paths of modalities. To alleviate the limitation, the proposed HyperDenseNet adopts the strategy where each stream receives inputs of image data of one modality, and the layers in the same and different paths are densely connected. The encoding path contains N streams, each responsible for one imaging modality. The Dense Multi-path U-Net supports Hyper-dense connections both within a single path and between several paths. In a densely-connected network, the output of the l^{th} layer is produced by the mapping H_l of the concatenation of all the feature layers.

$$x_l = H_l([x_{l-1}, x_{l-2}, \ldots, x_0]). \tag{9}$$

HyperDenseNet integrates the outputs among different paths based on the densely-connected network to obtain richer feature representation from combined modalities. In addition, the permuting and interleaving operations are applied to the concatenation to improve performance. Considering the case of two modalities with streams 1 and 2 denoted as x_l^1 and x_l^2 respectively, the output of the l^{th} layer of HyperDenseNet can be expressed as follows:

$$x_l^s = H_l^s(\pi_l^s([x_{l-1}^1, x_{l-1}^2, x_{l-2}^1, x_{l-2}^2, \ldots, x_0^1, x_0^2])), \tag{10}$$

where π_l^s represents the shuffling function acting on the feature maps.

Lachinov et al. [54] propose a deep cascaded variant of U-Net, Cascaded Unet, to process multi-modal input for better performance regarding brain tumor segmentation. Despite the feasibility of the original U-Net to handle multi-modal MRI image input, it fuses the feature information of all the modalities which is processed in an identical manner. Based on the original U-Net, the proposed Cascaded Unet employs multiple encoders in parallel for better exploiting feature representations for each specific modality. The encoder path contains separate subpaths where every subpath utilizes a convolution group to process one input modality and generate feature maps. Then elementwise maximum operation acts on the multiple feature maps per stage to obtain the resulting features. The output of the feature map is afterward joined with the corresponding feature map of the larger-scale block, which boosts the information flow between the feature maps at different scales. The decoder of Cascaded Unet produces output at each level depending on the output at the same scale and the output of the decoder block at the previous stage. This strategy encourages the model to iteratively improve the results from earlier iterations.

3) Leveraging Depth Information: Some methods modify the U-Net into a 3D model and design modules to extract the information across channels in order to fully exploit the structural information of the third-dimension medical images. For improving automatic brain tumor prognosis, Islam et al. adapt the U-Net architecture to a 3D model and integrate the 3D attention strategy to perform image segmentation [55]. Compared to only skip connections, the introduced 3D attention model is aggregated into the decoder part of U-Net that includes channel and spatial attention in parallel with skip connections. The additional 3D attention layers encourage the module to encode richer spatial features from the original images.

In this approach, the 3D attention U-Net is composed of a 3D encoder, the decoder, and skip connections combined with the channel and spatial attention mechanism. In the path for 3D spatial attention, the authors perform $1 \times 1 \times C$ convolution on the input feature maps to obtain the result of the $H \times W \times 1$ dimension. In parallel, the input feature maps are passed through an average pooling and then fed to the fully-connected layer to get the $1 \times 1 \times C$ sequential channel correlation. Since the two paths capture features parallelly, the inconsistency and sparsity caused by the two excitations can be alleviated by fusing skip connections. Furthermore, the integration of skip connections can enhance the performance of segmentation prediction which can be inferred from the experiments on the BraTS 2019 dataset.

F. Probabilistic Design

Another type of U-Net extension combines the classic U-Net with different types of probabilistic extensions. Depending on the task that should be achieved or the process that should be enhanced, different types of extensions from bayesian skip connections, over variational auto-encoders to Markov random fields are used, which are introduced in the following.

1) Variational Auto-Encoder (VAE) Regularization: In medical image segmentation tasks, different raters often produce different segmentations. Most of these different segmentations are plausible as many medical images contain ambiguities that can not be resolved considering only the image at hand. Taking this into consideration, Kohl et al. learn a distribution over segmentations from an ambiguous input to produce an unlimited number of possible segmentations instead of just providing the most likely hypothesis [50]. In their approach, they combine a U-Net, for producing reliable segmentations, with a conditional variational autoencoder (CVAE), which can model complex distributions and encodes the segmentation variants in a low-dimensional latent space.

Each position in the latent space encodes a different segmentation variant. Passing the input image through the prior net will determine the probability of the encoded variants for the given input image. For each possible segmentation to be predicted the network is applied to the same input image. A random sample from the prior probability distribution is drawn and broadcast to an N-channel feature map with the same shape as the segmentation map. It will then be concatenated with the final feature maps of the u-net and processed with successive 1×1 convolutions to produce the segmentation map corresponding to the drawn point from the latent space. Only the combination needs to be recalculated in each iteration, as the last feature maps of the U-Net and the output of the prior net can be reused for each hypothesis.

Apart from the standard training procedures for conditional VAEs and deterministic segmentation models, it has to be learned how to embed the segmentation variants in the latent space in a useful way. This is solved by the posterior net. It learns to recognize a segmentation variant and map it to a certain position in the latent space. A sample from its output

posterior distribution combined with the activation map of the u-net must result in a segmentation identical to the ground truth segmentation. From this, it follows that the training data set must include a set of different but plausible segmentations for each input image.

Myronenko [49] adds a VAE branch to a 3D U-Net architecture to address the problem of limited training data for brain tumor segmentation. In their architecture, the U-Net is used for the segmentation of the tumor, and the VAE is used for the reconstruction of the image sharing the same encoder. For the VAE, the output of the encoder is reduced to a lower dimensional space and a sample is drawn from the Gaussian distribution with the given mean and standard derivation (std). The sample is then reconstructed to the input image using an architecture similar to that of the U-Net decoder but without any skip connections. The total loss to be minimized during training is made up of three terms:

$$\mathbf{L} = \mathbf{L}_{\text{dice}} + 0.1 \cdot \mathbf{L}_{\text{L2}} + 0.1 \cdot \mathbf{L}_{\text{KL}}, \tag{11}$$

where $\mathbf{L}_{\text{dice}}$ is a soft dice loss between the predicted segmentation of the u-net and the GT segmentation. $\mathbf{L}_{\text{L2}}$ and $\mathbf{L}_{\text{KL}}$ are the losses for the VAE where $\mathbf{L}_{\text{L2}}$ describes how well the reconstructed image matches the input image and $\mathbf{L}_{\text{KL}}$ is the Kullback-Leibler (KL) divergence between the estimates normal distribution and a prior distribution $\mathcal{N}(0, 1)$. Using the VAE brach helps to better cluster the features at the end of the encoder. This helps to guide and regularize the shared encoder for small training set sizes. Adding the additional VAE branch, therefore, improved the performance and led to stable results for different random initializations of the network.

2) Graphical Model Algorithm: While the classic U-Net performs well on data from the same distribution as the training data, its accuracy decreases on out-of-distribution data.

To address this problem, Brudfors et al. [47] combine a U-Net with Markov random fields (MRFs) to form the MRF-Unet. The low-parameter, first-order MRFs are better at generalization because they encode simpler distributions which is an important quality to fit out of distribution data. The very accurate U-Net predictions make up for the fact that the MRFs are less flexible. As the combination of the U-Net and MRF distribution is intractable by calculating the product of the two, an iterative mean-field approach is used to estimate the closest factorized distribution under the Kullback-Leibler divergence. A detailed mathematical derivation of the process can be found in the work by Brudfors et al. [47]. Experiments showed that the combination of MRF and U-Net improved performance on in- and out-of-distribution data. The lightweight MRF component, which does not add any additional parameters to the architecture, serves as a simple prior and therefore learns abstract label-specific features.

Klug et al. [48] use a Bayesian skip connection in an attention-gated 3D U-Net to allow a prior to bypass most of the network and be reintegrated at the final layer in their work to segment stroke lesions in perfusion CT images. The skip connection provides the prior to the final network layer and should reduce false-positive rates for small and patchy segmentations of varying shapes. As a prior, the segmentation of the ischemic core obtained by a standard thresholding method is used. Klug et al. [48] evaluated two ways to combine the prior and the output of the U-Net to calculate the final output segmentation: Addition and convolution of the two maps. Superior results were achieved by using convolution for combination in all experiments.

The input to the U-Net is the concatenation of the 3D perfusion CT image and the prior. When comparing a 3D attention gate U-Net to the same architecture with the bayesian skip connection additionally reintegrating the prior at the end of the network, the latter achieves a better performance in terms of dice score with faster convergence. It is worth mentioning that we have found excessive papers in which probabilistic design is integrated into the U-Net in applications such as for brain tumor [132], and skin lesion segmentation [133].

G. Comparative Overview

In this section, we briefly review the recent works regarding U-Net variants presented in Section III-A – III-F for medical image segmentation in Table I. It lists the related network list in each direction along with information about the core ideas and the practical use cases. As detailed in Table I, the modification dealing with skip connections is one of the directions for the extension of the U-Net structure. Some works redesign skip connections by increasing the number of forward skip connections or aggregating some modules within the skip connections for processing feature maps. Some methods also apply bi-directional LSTM to combine the feature maps from Encoder and Decoder. The novel skip connections in these mentioned methods enable the models to be more flexible and therefore explore local and semantic features more efficiently and from different scales. However, it also means a more complex design of the network architecture which leads to a larger number of parameters and more expensive computation.

Instead of altering skip connections, some proposed methods use other types of backbones apart from the original U-Net by Ronneberger et al. [8]. Residual mapping structure, inception modules, and dense-connections are incorporated into the architectures respectively. Such designs alleviate the vanishing gradient and degradation problems, which facilitates the faster convergence of the training process. Several approaches focus on adapting convolution operations such as using deformable convolutional kernels to make the network more general and robust. Nevertheless, a higher computational cost is needed due to the additional convolution layers.

In the third strategy, some approaches utilize other mechanisms in the bottleneck aiming at enhancing feature extraction and compressing more useful spatial information. Attention modules are employed to model long-range spatial dependencies between pixels in the bottle-neck feature maps. Atrous spatial pyramid pooling (ASPP) with different sampling rates can resample the compressed feature maps in the bottleneck separately, which helps to gain the output of bottleneck from different sizes of reception fields.

The rise of the Transformer, which is prevalent in the field of NLP, inspires the development of computer vision. The proposed ViT facilitates the new trend of the combination of U-Net and

TABLE I
THE REVIEW OF U-NET-LIKE MODELS FOR MEDICAL IMAGE SEGMENTATION BASED ON THE PROPOSED TAXONOMY, FIG. 1

Strategy	Networks	Core Ideas	Practical Use Cases
SCE	Attention U-Net [66] UNet++ [11], [63] RA-Unet [67] BCDU-Net [70] UNet 3+ [64] BiO-Net [65]	Skip connections are defined as connections in a neural network that do not connect two following layers but instead skip over at least one layer. This strategy initially aimed to encourage feature reusability and compensate for gradient vanishing in a deeper network. This modification introduces the feasibility of transferring high spatial localization features from the encoder to the decoder for better segmentation maps. In addition, some use cases used skip connections as hierarchical multi-scale fusion paths for feature enrichment in diverse U-Net stages. Furthermore, skip connection could efficiently decrease the semantic feature gaps between different layers and scales.	• Exploit multiscale features [63] • Robust boundary representation [64] • Bi-directional feature representation [65] • Feature recalibration [66] • Suppress irrelevant regions besides feature reusability [67] • Enrich semantic representation [70]
BDE	Residual U-Net [56], [57], [104] Multi-Res U-Net [13], [105] Dense U-Net [106], [136] H-DenseUNet [58] DUNet [61], [107] S3D U-Net [12]	The backbone defines how the layers in the encoder are arranged and its counterpart is therefore used to describe the decoder architecture. Ideally, the strong backbone design (e.g., inception model) with pre-trained weight can further improve the model generalization capability.	• Converges to lower loss [58] • Addressing gradient vanishing [104] • Faster convergence rate [56] • Feature reusability [57] • Multi-scale encoding [13] • Better boundary representation [105] • Fine-grained feature set [106] • Cross-modality representation [58] • Reducing computation burden of multi-scale representation [61] • Efficient multi-scale computation [12]
BE	ASPP [77] MA-Net [72] COPLE-Net [78] SA-UNet [73] FRCU-Net [75] JCS [79] MS-Net [110]	The network bottleneck contains the compressed representation of the input data and provides necessary information (e.g., semantic, texture, shape features) to reconstruct the segmentation map. Any improvement in the bottleneck design can further improve the prediction result.	• Frequency recalibration [75] • Spatial attention [73] • Feature pyramid [77], [78] • Imposing attention mechanism [79]
T	TransUNet [34] TransBST [35] Swin-Unet [44] UNETR [116] TMU [40] UCTransNet [41]	Transformers' critical point is to compensate for CNN's limited receptive field. Extracting long-range contextual information with intuition to look at the whole image at once is a promotional function of Transformers. Due to the 1D sequence mapping functionality of Transformers, they can use as a play-and-plug module at different parts of U-Net-like structures. However, due to the quadratic computational complexity nature, utilizing efficient Transformers' design even from the NLP field within vision architectures is beneficial. Since Transformers calculate the affinities between different parts of input data adaptively, utilizing them is a wise solution for multi-scale feature amalgamation.	• Improving CNN bottleneck's feature discriminancy [34] • Capture inter-slice affinities from 3D data [35] • Hierarchical Efficient Transformer-based design [44] • Modeling 3D volumetric data with global multi-scale information [116] • Feature re-calibration / degrading the boundary maps erroneous [40] • Decreasing the gap between multi-scale semantic features [41]
RRE	Focal Tversky Attention U-Net [14] PT M-Net [51] Dense Multi-path U-Net [53] MFP-Unet [52] Cascaded Unet [54]	The key objective is to enhance the performance of the trained models by utilizing all available information from multi-modal or multi-scale images while retaining the most desirable and relevant features. Some methods also operate directly on volumetric images to take full advantage of depth information.	• Improved precision and recall balance [14] • Hierarchical representation learning [51] • Richer feature representation from combined modalities [53] • Robust architecture regarding the capabilities of feature representation in a pyramid [52] • Boosted information flow between the different scales [54]
PD	Probabilistic U-Net [50] MRF U-Net [47] VAE Regularization [49] Bayesian Skip [48]	In medical image segmentation tasks, different graders often produce different segmentations. Most of these different segmentations are plausible as many medical images contain ambiguities that can not be resolved considering only the shot at hand. Probabilistic models aim to model the nature of uncertainty in medical data.	• Modeling annotation uncertainty [50] • Addressing out-of-distribution [47] • Imposing regularization to address data limitation [49] • Encouraging feature-reusability and reduce FP rate [48]

{SCE, BDE, BE, T, RRE and PD} stands for skip connection enhancement, backbone design enhancement, bottleneck enhancement, transformer, rich representation enhancement and probabilistic design, respectively.

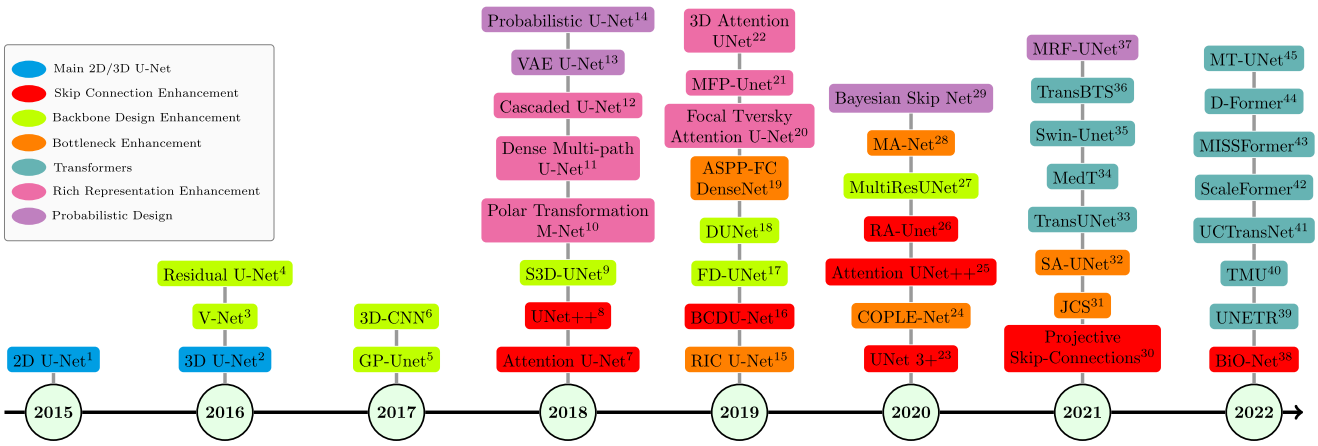


Fig. 4. The timeline of prominent U-Net-based methods proposed in medical semantic segmentation literature, from 2015 to 2022. The superscripts in ascending order denote the 1. [8], 2. [80], 3. [57], 4. [56], 5. [134], 6. [135], 7. [66], 8. [11], 9. [12], 10. [51], 11. [53], 12. [54], 13. [49], 14. [50], 15. [74], 16. [70], 17. [136], 18. [61], 19. [77], 20. [14], 21. [52], 22. [55], 23. [64], 24. [78], 25. [68], 26. [67], 27. [13], 28. [72], 29. [48], 30. [69], 31. [79], 32. [73], 33. [34], 34. [43], 35. [44], 36. [35], 37. [47], 38. [65], 39. [116], 40. [40], 41. [41], 42. [42], 43. [45], 44. [46], 45. [39], respectively.

Transformer. The tokenized input images are passed through Transformer to extract global information that will supplement U-Net. Despite the state-of-the-art (SOTA) performance that the Transformer-based U-Net-like models have achieved, the large number of parameters of models sometimes cause a long time for convergence. Some models are also highly dependent on the pretrained weights.

The last direction combines the original U-Net with various types of probabilistic extension modules resulting in new types of variants. Probabilistic U-Net integrates the U-Net structure with a conditional variational autoencoder (CVAE) to generate an unlimited number of plausible prediction results when the inputs are obscure. Furthermore, the network with Markov Random Fields (MRF) has the advantage of preventing

the model from overfitting with precise segmentations, which significantly improves the performance on out-of-distribution data. The methods in this direction demonstrate more or less robustness to defective input datasets, such as those of limited size or containing ambiguous images. Fig. 4 demonstrates the timeline of typical U-Net-based methods proposed in medical semantic segmentation literature from 2015 to 2022. As shown in the timeline, the U-Net structure has continued to be appealing in recent years regarding the task of medical image segmentation. The direction for the extension of U-Net is prominently influenced by Transformer after 2021 as a result of the emergence of ViT [112].

IV. Conclusion

In this study, we presented a thorough review of the literature on U-Net and its variants, which have become increasingly popular in the field of medical image segmentation over the years. We examined the main taxonomy of U-Net and its extensions, and highlighted the use cases of each category. Additionally, we provided a benchmark of performance and speed, as well as open problems, in the supplementary file. To structure the wide variety of U-Net extensions, we grouped them according to the type of change made to the architecture. Adaptions to the skip connections are presented in Section III-A, which comprises methods that increase the number of skip connections, apply additional processing to the feature maps in the skip connections to focus on the areas of interest, or improve the fusion of the encoder and decoder feature maps combined through the skip connections. Section III-B introduces different types of backbones used in the U-Net architecture, such as deeper network architectures, processing of 3D images, or multi-resolution feature extraction for high inter-patient variability in terms of the size of the object(s) of interest. In Section III-C, a variety of extensions of the bottleneck of the original U-Net is examined, including approaches that adapt the bottleneck for multi-scale representation of the bottleneck feature maps or position-wise attention to model spatial dependencies between pixels in the bottleneck. The transformer variants of the U-Net architecture, introduced in Section III-D, enable the networks to capture inter-pixel long-range dependencies and compensate for the otherwise limited receptive field of the convolutions in the original U-Net. Section III-E presents appraoches that adapt the U-Net architecture to use information from multiple modalities and/or scales for a rich representation of features. Finally, Section III-F presented methods to model the uncertainty in annotations of medical data or out-of-distribution samples.

References

[1] M. Antonelli et al., "The Medical Segmentation Decathlon," *Nat. Commun.*, vol. 13, no. 1, pp. 1–13, 2022.

[2] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, 2021.

[3] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[4] E. Wu, K. Wu, R. Daneshjou, D. Ouyang, D. E. Ho, and J. Zou, "How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals," *Nat. Med.*, vol. 27, no. 4, pp. 582–584, 2021.

[5] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH image to ImageJ: 25 years of image analysis," *Nat. Methods*, vol. 9, no. 7, pp. 671–675, 2012.

[6] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2843–2851.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2015, pp. 234–241.

[9] Q. Huang, J. Sun, H. Ding, X. Wang, and G. Wang, "Robust liver vessel extraction using 3D U-Net with variant dice loss function," *Comput. Biol. Med.*, vol. 101, pp. 153–162, 2018.

[10] A. Kazerouni et al., "Diffusion models for medical image analysis: A comprehensive survey," 2022, *arXiv:2211.07804*.

[11] Z. Zhou, M. M. Rahman 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*. Berlin, Germany: Springer, 2018, pp. 3–11.

[12] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-Unet: Separable 3D U-Net for brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2018, pp. 358–368.

[13] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.

[14] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag.*, 2019, pp. 683–687.

[15] H. Zhao and N. Sun, "Improved U-Net model for nerve segmentation," in *Proc. Int. Conf. Image Graph.*, Springer, 2017, pp. 496–504.

[16] M. Frid-Adar, A. Ben-Cohen, R. Amer, and H. Greenspan, "Improving the segmentation of anatomical structures in chest radiographs using U-Net with an ImageNet pre-trained encoder," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Berlin, Germany: Springer, 2018, pp. 159–168.

[17] D. Waiker, P. D. Baghel, K. R. Varma, and S. P. Sahu, "Effective semantic segmentation of lung X-Ray images using U-Net architecture," in *Proc. 4th Int. Conf. Comput. Methodol. Commun.*, 2020, pp. 603–607.

[18] J. I. Orlando et al., "U2-Net: A Bayesian U-Net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans," in *Proc. IEEE 16th Int. Symp. Biomed. Imag.*, 2019, pp. 1441–1445.

[19] R. Asgari, S. Waldstein, F. Schlanitz, M. Baratsits, U. Schmidt-Erfurth, and H. Bogunović, "U-Net with spatial pyramid pooling for drusen segmentation in optical coherence tomography," in *Proc. Int. Workshop Ophthalmic Med. Image Anal.*, Springer, 2019, pp. 77–85.

[20] H. Wang et al., "ICA-UNet: An improved U-Net network for brown adipose tissue segmentation," *J. Innov. Opt. Health Sci.*, vol. 15, no. 03, 2022, Art. no. 2250018.

[21] R. Azad, N. Khosravi, M. Dehghanmanshadi, J. Cohen-Adad, and D. Merhof, "Medical image segmentation on MRI images with missing modalities: A review," 2022, *arXiv:2203.06217*.

[22] P. Costa et al., "Towards adversarial retinal image synthesis," 2017, *arXiv:1701.08974*.

[23] H. Wu, X. Jiang, and F. Jia, "UC-GAN for mr to CT image synthesis," in *Proc. Workshop Artif. Intell. Radiat. Ther.*, Springer, 2019, pp. 146–153.

[24] B. Sun, S. Jia, X. Jiang, and F. Jia, "Double U-Net CycleGAN for 3D MR to CT image synthesis," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 18, pp. 149–156, 2022.

[25] M. P. Reymann et al., "U-Net for spect image denoising," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf.*, 2019, pp. 1–2.

[26] S. Nasrin, M. Z. Alom, R. Burada, T. M. Taha, and V. K. Asari, "Medical image denoising with recurrent residual U-Net (R2U-Net) base auto-encoder," in *Proc. IEEE Nat. Aerosp. Electron. Conf.*, 2019, pp. 345–350.

[27] S. Lee, M. Negishi, H. Urakubo, H. Kasai, and S. Ishii, "MU-Net: Multi-scale U-Net for two-photon microscopy image denoising and restoration," *Neural Netw.*, vol. 125, pp. 92–103, 2020.

[28] S. Guan, K.-T. Hsu, M. Eyassu, and P. V. Chitnis, "Dense dilated UNet: Deep learning for 3D photoacoustic tomography image reconstruction," 2021, *arXiv:2104.03130*.

[29] J. Feng, J. Deng, Z. Li, Z. Sun, H. Dou, and K. Jia, "End-to-end Res-UNet based reconstruction algorithm for photoacoustic imaging," *Biomed. Opt. Exp.*, vol. 11, no. 9, pp. 5321–5340, 2020.

[30] D. Qiu, Y. Cheng, and X. Wang, "Progressive U-Net residual network for computed tomography images super-resolution in the screening of COVID-19," *J. Radiat. Res. Appl. Sci.*, vol. 14, no. 1, pp. 369–379, 2021.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[32] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[33] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021.

[34] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[35] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2021, pp. 109–119.

[36] Y. Li et al., "GT U-Net: A U-Net like group transformer network for tooth root segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, Springer, 2021, pp. 386–395.

[37] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "COTR: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2021, pp. 171–180.

[38] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2022, pp. 272–284.

[39] H. Wang et al., "Mixed transformer U-Net for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 2390–2394.

[40] A. Reza, H. Moein, W. Yuli, and M. Dorit, "Contextual attention network: Transformer meets U-Net," 2022, *arXiv:2203.01932*.

[41] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2441–2449.

[42] H. Huang et al., "ScaleFormer: Revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation," 2022, *arXiv:2207.14552*.

[43] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2021, pp. 36–46.

[44] H. Cao et al., "Swin-UNet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 205–218.

[45] X. Huang, Z. Deng, D. Li, and X. Yuan, "MissFormer: An effective medical image segmentation transformer," 2021, *arXiv:2109.07162*.

[46] Y. Wu et al., "D-Former: A U-shaped dilated transformer for 3D medical image segmentation," 2022, *arXiv:2201.00462*.

[47] M. Brudfors et al., "An MRF-UNet product of experts for image segmentation," in *Proc. Conf. Med. Imag. Deep Learn.*, PMLR, 2021, pp. 48–59.

[48] J. Klug, G. Leclerc, E. Dirren, M. G. Preti, D. V. D. Ville, and E. Carrera, "Bayesian skip net: Building on prior information for the prediction and segmentation of stroke lesions," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2020, pp. 168–180.

[49] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2018, pp. 311–320.

[50] S. Kohl et al., "A probabilistic U-Net for segmentation of ambiguous images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6965–6975.

[51] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[52] S. Moradi et al., "MFP-UNet: A novel deep learning based approach for left ventricle segmentation in echocardiography," *Physica Medica*, vol. 67, pp. 58–69, 2019.

[53] J. Dolz, I. Ben Ayed, and C. Desrosiers, "Dense multi-path U-Net for ischemic stroke lesion segmentation in multiple image modalities," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2018, pp. 271–282.

[54] D. Lachinov, E. Vasiliev, and V. Turlapov, "Glioma segmentation with cascaded UNet," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2018, pp. 189–198.

[55] M. Islam, V. Vibashan, V. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain tumor segmentation and survival prediction using 3D attention UNet," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2019, pp. 262–272.

[56] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Berlin, Germany: Springer International Publishing, 2016, pp. 179–187.

[57] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[58] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[59] A. Karaali, R. Dahyot, and D. J. Sexton, "DR-VNet: Retinal vessel segmentation via dense residual UNet," in *Proc. Int. Conf. Pattern Recognit. Artif. Intell.*, Springer, 2022, pp. 198–210.

[60] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst.*, 2020, pp. 558–564.

[61] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, 2019.

[62] C. Kou, W. Li, W. Liang, Z. Yu, and J. Hao, "Microaneurysms segmentation with a U-Net based on recurrent residual convolutional neural network," *J. Med. Imag.*, vol. 6, no. 2, 2019, Art. no. 025008.

[63] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[64] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 1055–1059.

[65] T. Xiang, C. Zhang, D. Liu, Y. Song, H. Huang, and W. Cai, "BiO-Net: Learning recurrent bi-directional connections for encoder-decoder architecture," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2020, pp. 74–84.

[66] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv: 1804.03999*.

[67] Q. Jin, Z. Meng, C. Sun, H. Cui, and R. Su, "RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans," *Front. Bioeng. Biotechnol.*, vol. 8, 2020, Art. no. 605132.

[68] C. Li et al., "Attention UNet++: A nested attention-aware U-Net for liver CT image segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 345–349.

[69] D. Lachinov, P. Seeböck, J. Mai, F. Goldbach, U. Schmidt-Erfurth, and H. Bogunovic, "Projective skip-connections for segmentation along a subset of dimensions in retinal oct," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2021, pp. 431–441.

[70] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with densley connected convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 406–415.

[71] H. Li et al., "CR-UNet: A composite network for ovary and follicle segmentation in ultrasound images," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 974–983, Apr. 2020.

[72] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-Net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179 656–179 665, 2020.

[73] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "SA-UNet: Spatial attention U-Net for retinal vessel segmentation," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2021, pp. 1236–1242.

[74] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "RIC-UNet: An improved neural network based on UNet for nuclei segmentation in histology images," *IEEE Access*, vol. 7, pp. 21 420–21 428, 2019.

[75] R. Azad, A. Bozorgpour, M. Asadi-Aghbolaghi, D. Merhof, and S. Escalera, "Deep frequency re-calibration U-Net for medical image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3274–3283.

[76] R. Azad, N. Khosravi, and D. Merhof, "SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2022, pp. 48–62.

[77] J. Hai et al., "Fully convolutional densenet with multiscale context for automated breast tumor segmentation," *J. Healthcare Eng.*, vol. 2019, 2019, Art. no. 8415485.

[78] G. Wang et al., "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2653–2663, Aug. 2020.

[79] Y.-H. Wu et al., "JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 3113–3126, 2021.

[80] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2016, pp. 424–432.

[81] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[82] RSNA 2022 cervical spine fracture detection, 2022. [Online]. Available: https://www.kaggle.com/competitions/rsna-2022-cervical-spine-fracture-detection

[83] K. S. Mader, "Finding and measuring lungs in CT data," Apr. 2017. [Online]. Available: https://www.kaggle.com/kmader/finding-lungs-in-ct-data

[84] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015, Art. no. 12.

[85] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, *arXiv: 1902.09063*.

[86] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2014.

[87] M. Fitzke et al., "Oncopetnet: A deep learning based AI system for mitotic figure counting on H&E stained whole slide digital images in a large veterinary diagnostic lab setting," 2021, *arXiv:2108.07856*.

[88] J. De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, 2018.

[89] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[90] R. Baskar, K. A. Lee, R. Yeo, and K.-W. Yeoh, "Cancer and radiation therapy: Current advances and future directions," *Int. J. Med. Sci.*, vol. 9, no. 3, 2012, Art. no. 193.

[91] S. Nikolov et al., "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," 2018, *arXiv: 1809.04430*.

[92] J. S. Mikeljevic et al., "Trends in postoperative radiotherapy delay and the effect on survival in breast cancer patients treated with conservation surgery," *Brit. J. Cancer*, vol. 90, no. 7, pp. 1343–1348, 2004.

[93] P. Kickingereder et al., "Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: A multicentre, retrospective study," *Lancet Oncol.*, vol. 20, no. 5, pp. 728–740, 2019.

[94] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge," in *Proc. Third Int. Workshop Brainlesion Glioma Mult. Scler. Stroke Traumatic Brain Injuries*, Quebec City, QC, Canada, Springer, Sep. 14, 2018, pp. 287–297.

[95] L. Maier-Hein et al., "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nat. Commun.*, vol. 9, no. 1, pp. 1–13, 2018.

[96] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. Berlin, Germany: Springer, 2006.

[97] S. Banerjee et al., "Ultrasound spine image segmentation using multi-scale feature fusion skip-inception U-Net (SIU-Net)," *Biocybern. Biomed. Eng.*, vol. 42, no. 1, pp. 341–361, 2022.

[98] M. Asadi-Aghbolaghi, R. Azad, M. Fathy, and S. Escalera, "Multi-level context gating of embedded collective knowledge for medical image segmentation," 2020, *arXiv: 2003.05056*.

[99] B. Liefers, C. González-Gonzalo, C. Klaver, B. van Ginneken, and C. I. Sánchez, "Dense segmentation in selected dimensions: Application to retinal optical coherence tomography," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, PMLR, 2019, pp. 337–346.

[100] D. Li, Y. Peng, Y. Guo, and J. Sun, "MFAUNet: Multiscale feature attentive U-Net for cardiac MRI structural segmentation," *IET Image Process.*, vol. 16, no. 4, pp. 1227–1242, 2022.

[101] T. Nguyen, B.-S. Hua, and N. Le, "3D-UCaps: 3D capsules UNet for volumetric image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2021, pp. 548–558.

[102] W. Weng and X. Zhu, "INet: Convolutional networks for biomedical image segmentation," *IEEE Access*, vol. 9, pp. 16 591–16 603, 2021.

[103] R. LaLonde, Z. Xu, I. Irmakci, S. Jain, and U. Bagci, "Capsules for biomedical image segmentation," *Med. Image Anal.*, vol. 68, 2021, Art. no. 101889.

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[105] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[106] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[107] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[108] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3367–3375.

[109] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," *J. Med. Imag.*, vol. 6, no. 1, 2019, Art. no. 014006.

[110] B. Zhang et al., "Multi-scale feature pyramid fusion network for medical image segmentation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 18, pp. 353–365, 2023.

[111] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[112] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv: 2010.11929*.

[113] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[114] R. Azad et al., "Advances in medical image analysis with vision transformers: A comprehensive review," 2023, *arXiv:2301.03505*.

[115] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable {DETR}: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–16.

[116] A. Hatamizadeh et al., "Unetr: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 574–584.

[117] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 108–126.

[118] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," 2019, *arXiv: 1912.12180*.

[119] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.

[120] X. Chu et al., "Conditional positional encodings for vision transformers," 2021, *arXiv:2102.10882*.

[121] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "LocalViT: Bringing locality to vision transformers," 2021, *arXiv:2104.05707*.

[122] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[123] W. Wang et al., "PVT V2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.

[124] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "UFormer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17 683–17 693.

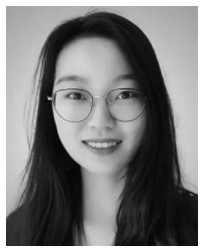
[125] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12 077–12 090.

[126] J. Wang, P. Lv, H. Wang, and C. Shi, "SAR-U-Net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in computed tomography," *Comput. Methods Prog. Biomed.*, vol. 208, 2021, Art. no. 106268.
[127] P. Zhao, J. Zhang, W. Fang, and S. Deng, "SCAU-Net: Spatial-channel attention U-Net for gland segmentation," *Front. Bioeng. Biotechnol.*, vol. 8, 2020, Art. no. 670.
[128] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[129] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," in *Readings in Computer Vision*. Amsterdam, The Netherlands: Elsevier, 1987, pp. 671–679.
[130] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection," *IEEE Access*, vol. 7, pp. 1721–1735, 2018.
[131] S. Moradi et al., "A novel deep learning based approach for left ventricle segmentation in echocardiography: MFP-UNet," 2019, *arXiv: 1906.10486*.
[132] C. Savadikar, R. Kulhalli, and B. Garware, "Brain tumour segmentation using probabilistic U-Net," in *Proc. Int. MICCAI Brainlesion Workshop*, Springer, 2020, pp. 255–264.
[133] X. Chen, Y. Zhao, and C. Liu, "Medical image segmentation using scalable functional variational Bayesian neural networks with Gaussian processes," *Neurocomputing*, vol. 500, pp. 58–72, 2022.
[134] F. Dubost et al., "GP-UNet: Lesion detection from weak labels with a 3D regression network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Springer, 2017, pp. 214–221.
[135] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 8, pp. 409–417, 2017.
[136] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 568–576, Feb. 2020.

Reza Azad received the MSc degree in artificial intelligence and robotics from the Sharif University of Technology (SUT), Tehran, Iran, in 2017. He is currently working toward the PhD degree with the Faculty of Electrial Engineering and Information Technology, RWTH Aachen University, Germany. His research interests include deep learning, medical image processing, and computer vision.

Ehsan Khodapanah Aghdam received the MSc degree in electrical engineering, communication systems from Shahid Beheshti University, Tehran, Iran, in 2019. Since 2021, he has been actively conducting independent and volunteer research either alone or with vision and medical imaging laboratories. His research interests include deep learning, medical image analysis, and computer vision.

Amelie Rauland received the BSc and MSc degrees in electrical engineering from RWTH Aachen University with a focus on computer engineering, in 2018 and 2021, respectively. Since 2022, she is currently working toward the PhD degree as a member of the IRTG 2150 and the Brain and Behaviour group of the Institute of Neurosciences and Medicine at Forschungszentrum Jülich, Jülich, Germany.



Yiwei Jia received the BSc degree in electrical engineering from Sun Yat-sen University, Guangzhou, China, in 2020. She is currently working toward the MSc degree and student research assistant with the Institute of Imaging & Computer Vision, RWTH Aachen University, Aachen, Germany. Her research interests include (but are not limited to) computer vision, deep learning, and medical image analysis.

Atlas Haddadi Avval is currently working toward the MD degree and a medicine student with the Mashhad University of Medical Sciences (MUMS), Mashhad, Iran. She is a researcher currently affiliated with the Department of Radiology, MUMS and has published and presented more than 30 peer-reviewed articles and conference abstracts, mainly in the fields of radiomics and radiogenomics, quantitative medical imaging analysis, and the use of deep learning in radiology.

Afshin Bozorgpour received the MSc degree in information technology with a specialization in network science from the University of Tehran, Iran, in 2021. Currently, he is pursuing his doctoral research with the Faculty of Informatics and Data Science, University of Regensburg, Germany. His primary research interests lie in the intersection of deep learning, generative models, medical image processing, and computer vision.

Sanaz Karimijafarbigloo received the MSc degree in electrical engineering from Shiraz University, Shiraz, Iran, 2019. She is currently working toward the PhD degree with the Faculty of Mathematics, Computer Science and Natural Sciences, RWTH Aachen University, Aachen, Germany, in collaboration with the Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany. Her research interests mainly cover medical image analysis, computer vision, and deep learning.

Joseph Paul Cohen received the PhD degree in computer science and machine learning from the University of Massachusetts, Boston, USA. He is a applied scientist with Amazon, USA. He previously held a postdoctoral fellowship with the Center for Artificial Intelligence in Medicine & Imaging, Stanford University, USA, as well as with Mila, the Quebec AI Institute, Canada. He is currently focusing on the limits of AI in medicine concerning computer vision, genomics, and clinical data. He received a U.S. National Science Foundation Graduate Fellowship, as well as an IVADO Postdoctoral Fellowship.

Ehsan Adeli Currently holds the assistant professor position with Stanford University, USA. Before joining Stanford, he was a postdoctoral research fellow with the Biomedical Research Imaging Center (BRIC), the University of North Carolina, Chapel Hill, NC, USA, and a research scholar with the Robotic Institute, Carnegie Mellon University, Pittsburgh, USA. His research lies at the intersection of machine learning, computer vision, healthcare, and computational neuroscience.

Dorit Merhof received the PhD degree in biomedical image analysis from the University of Erlangen-Nuremberg, Germany, in 2007. After two years with Siemens Molecular Imaging, Oxford, U.K., she joined the University of Konstanz, Germany, as an assistant professor for Visual Computing. From 2013-2022, she was a full professor (W3) with RWTH Aachen university. Since 2022, she is full professor (W3) and head of the Institute of Image Analysis and Computer vision, Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany. Her research interests include image analysis and computer vision for biological and medical image data.