Current and emerging trends in medical image segmentation with deep learning

Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek Kumar Singh, Vincent Jaouen, Dimitris Visvikis

Current and emerging trends in medical image segmentation with deep learning

Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek Kumar Singh, Vincent Jaouen and Dimitris Visvikis

Invited Paper

***Abstract*—In recent years, the segmentation of anatomical or pathological structures using deep learning has experienced a widespread interest in medical image analysis. Remarkably successful performance has been reported in many imaging modalities and for a variety of clinical contexts to support clinicians in computer-assisted diagnosis, therapy or surgical planning purposes. However, despite the increasing amount of medical image segmentation challenges, there remains little consensus on which methodology perform best. Therefore, we examine in this paper the numerous developments and breakthroughs brought since the rise of U-Net inspired architectures. Especially, we focus on the technical challenges and emerging trends that the community is now focusing on, including conditional generative adversarial and cascaded networks, medical Transformers, contrastive learning, knowledge distillation, active learning, prior knowledge embedding, cross-modality learning, multi-structure analysis, federated learning or semi-supervised and self-supervised paradigms. We also suggest possible avenues to be further investigated in future research efforts.**

***Index Terms*—artificial intelligence, semantic segmentation, deep neural networks, vision Transformers, medical imaging**

I. Introduction

THE increased volume of medical data to be interpreted by clinicians for diagnosis, therapeutic or surgical planning purposes has encouraged the development of computer-aided image analysis tools to benefit from precise, fast, repeatable and objective measurements made by computational resources. Among existing analysis tasks, medical image segmentation whose goal is to extract the boundaries of anatomical or pathological structures from medical images remains crucial. Also commonly used in computed vision [1], semantic segmentation is a key step for many medical imaging workflows since the information arising from the resulting voxel-wise localization can greatly help clinicians to diagnose disorders, assess disease progression, plan therapeutic interventions or monitor treatment effects. Core feature of many computer-aided detection and diagnosis schemes, image segmentation is involved in the analysis of many imaging modalities including computed tomography (CT), magnetic resonance (MR), positron emission tomography (PET) or ultrasound (US).

Delineating anatomical or pathological structures from medical images is traditionally performed manually. This task is exceedingly time-consuming and requires suitable clinical expertise to get clinically-relevant contours. This is therefore not applicable to large volumes of data typically produced in clinical routine or research studies. Given the potential fatigue of human experts and the wide variations in expertise, manual segmentation is prone to strong intra- and inter-expert variability [2]. Irregularities of the targeted structures, morphological variations or pathological deformities between patients as well as the potential lack of clearly visible boundaries with the surrounding anatomy further affect the non-agreement between operators. To ease the process, intra-subject semi-automatic techniques consisting of ascending and descending non-linear registration steps applied to manually-drawn masks can be applied to obtain volumetric results [3, 4]. Although more affordable than 3D volume annotations, such propagation schemes from sparse manual delineations to remaining slices still need interactions and may require manual refinements.

Mathematical models and low-level image processing have been extensively exploited for segmentation purposes before the rise of learning techniques. In particular, model-based segmentation incorporating statistical shape models has been followed in various clinical contexts [5]. These models have been further improved by exploiting prior knowledge of shape information, for instance by relying on internal shape fitting and auto-correction to guide the delineation process [6]. Conversely, aligning and merging manually segmented images into a specific atlas coordinate space have been developed as a reliable alternative to statistical shape models. In this context, various single- and multi-atlas methods have been proposed relying on non-linear registration [7]. Some hybrid methods relying on statistical shape models constrained with probabilistic atlases have also been investigated. Medical image segmentation has been also performed through Bayesian approaches using expectation-maximization [8], possibilistic clustering [9], histogram-based thresholding followed by region growing [10], active contours [11] and more recently machine learning (ML) [12] techniques.

However, the previously described methodologies are not perfectly suited for high inter-subject shape variability, weak boundaries and significant differences in tissue appearance. In most cases, their robustness is not up to the inherent limitations of medical images such as noise, non-uniform contrasts or motion artifacts. Moreover, many of these methods are semi-automatic and hence require prior knowledge, associated with high computational costs. This has strongly motivated the development of deep learning (DL) techniques to exploit image characteristics (e.g. contrast variation, orientation, shape, texture patterns) in a more efficient data-driven manner.

In recent years, artificial intelligence (AI) and more particularly DL models have reached impressive performance in

This work did not involve human subjects or animals in its research. P.-H. Conze and V. Jaouen are with IMT Atlantique, LaTIM UMR 1101, Inserm, Brest, France, pierre-henri.conze@imt-atlantique.fr. V. K. Singh is with Queen's University Belfast, Belfast, United Kingdom. G. Andrade-Miranda and D. Visvikis are with Inserm, LaTIM UMR 1101, Brest, France.

medical image segmentation, becoming the new state-of-the-art [13]. The transition between systems that use handcrafted features (i.e. ML) to systems that learn features directly from data (i.e. DL) is now acted. The most widely-used models to date are based on convolutional neural networks (CNN). Especially, the U-Net convolutional architecture [14] has been widely adopted in the community thank to its ability to output detailed delineations using a relatively low amount of training data. Nevertheless, new paradigms recently appear based on attention mechanisms, accompanying the emerging interest devoted to pure and hybrid Transformers-based models [15].

Despite the availability of review articles summarizing existing DL approaches in medical image segmentation [13, 16], a more exhaustive, updated and comprehensive panorama is needed to allow a wide audience (e.g. researchers, clinicians, radiologists) to benefit from the latest and emerging trends in the field. Some of these existing reviews only focus on specific aspects such as anatomy-aided techniques [17], multi-organ segmentation [18], 3D CNN models [19] or U-Net and its variants [20] whereas others [15, 16, 21] address a broader overview of medical image analysis with DL by describing various tasks including classification, detection, segmentation or registration. In addition, despite the increasing amount of medical image segmentation challenges, there remains little consensus on which methodology performs the best. In this context, this paper aims at examining the numerous developments and breakthroughs brought since the rise of U-Net [14] inspired architectures with a novel and specific focus on the technical challenges and emerging scenarios that the community is now focusing on, including knowledge distillation, contrastive learning, medical Transformers, prior knowledge embedding, cross-modality analysis, federated, active, self- or semi-supervised learning. The recurring motivation for these new approaches lies in the difficulty to obtain large annotated datasets, as compared to other imaging tasks (e.g. classification) or fields (e.g. computer vision).

This paper starts with a description of both background (Sect.II) and clinical needs (Sect.III) before providing an in-depth overview of current trends (e.g. prior knowledge embedding) in medical image segmentation with DL (Sect.IV). By bringing the light to medical Transformers, multi-task learning, segmentation uncertainty, contrastive learning and knowledge distillation, emerging trends are then motivated and explained in Sect.V. Sect.VI focuses on emerging applications including cross-modality learning, multi-domain segmentation, semi-supervised, active and federated learning. To help readers navigate through the paper, an overview of the targeted current and emerging trends is provided in Tab.I. Sect.VII finally concludes this paper by summarizing and discussing the possible avenues to be further investigated in the future.

TABLE I
CURRENT AND EMERGING TRENDS IN MEDICAL IMAGE SEGMENTATION WITH DEEP LEARNING, PROVIDED WITH A REPRESENTATIVE REFERENCE.

Topic	Sect.	Description	Ref.
Current trends			
Conditional generative adversarial networks	IV-A	Discriminator assesses if segmentation is synthetic or real	[22]
Cascaded networks	IV-B	Series of convolutional encoder-decoders	[23]
Prior knowledge embedding	IV-C	Regularization term added into the loss function	[24]
Deep supervision	IV-D	Objective functions at some hidden layers	[25]
Attention mechanisms	IV-E	Channel attention, spatial attention, branch channel...	[26]
Multi-structure analysis	IV-F	Incorporating inter-structure relationships	[27]
Learning frameworks	IV-G	Unified frameworks, neural architecture search	[28]
Emerging trends			
Medical Transformers	V-A	Hybrid CNN-Transformers, full Transformers...	[29]
Multi-task learning	V-B	Taking advantage of information shared among tasks	[30]
Segmentation uncertainty	V-C	Uncertainty modelling to improve performance	[31]
Constrative learning	V-D	Learning a disentangled feature representation	[32]
Knowledge distillation	V-E	Distilling information between models	[33]
Emerging applications			
Cross-modality segmentation	VI-A	Exploiting complementary between modalities	[34]
Multi-domain segmentation	VI-B	Dealing with multiple intensity domains	[35]
Self-supervised learning	VI-C	Self-supervised contrastive learning, pre-text tasks	[36]
Semi-supervised learning	VI-D	Mean teacher, pseudo labeling, auxiliary task...	[37]
Federated learning	VI-E	Distributed training between clinical institutions	[38]
Active learning	VI-F	Assisting annotators in the annotation process	[39]
Lightweight networks	VI-G	Trade-off between trainable parameters and performance	[40]

II. BACKGROUND

Over the last few years, CNN models have become state-of-the-art in medical image segmentation due to their ability to learn hierarchical representations of image features in a data-driven fashion [16]. Before going into current and emerging trends, this section aims at explicitly formulating what is medical image segmentation using DL (Sect.II-A) and describing seminal works (Sect.II-B) until the development of U-Net (Sect.II-C). The last part on data augmentation (Sect.II-D), especially useful to address data scarcity issues frequently encountered in the field, completes the panorama.

A. Problem formulation

Let $\mathcal{X}$ be a set of images $\boldsymbol{x} \in \mathbb{R}^{H \times W \times D}$ where H, W and D are the image dimensions in x-, y- and z- dimension respectively while the annotation set $\mathcal{Y} \subset [0,1]^{H \times W \times D \times C}$ contains for each $\boldsymbol{x} \in \mathcal{X}$ a map $\boldsymbol{y}$ of $H \times W \times D$ one-hot vectors indicating the ground truth classes for all voxels. In a fully-supervised setting, a deep segmentation network ϕ aims at approximating a mapping function $\phi : \boldsymbol{x} \rightarrow \phi(\boldsymbol{x}; \boldsymbol{\Theta}_\phi) = \hat{\boldsymbol{y}}$ between intensity $\boldsymbol{x}$ and class labels $\boldsymbol{y}$ images from N training samples $\{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{1 \leq n < N}$ by optimizing a loss function $\mathcal{L}_\phi(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N} \sum_{i=0}^{N} \ell_\phi(\boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)$ with $\hat{\boldsymbol{y}}_n = \phi(\boldsymbol{x}_n)$ through a stochastic optimizer. The parameters of ϕ, namely $\boldsymbol{\Theta}_\phi$, are optimized during the training process. A stochastic gradient descent scheme, from standard gradient descent to Adam [41]

or more elaborated optimizers, aims at finding the optimal weights $\boldsymbol{\Theta}_\phi^*$ such that $\boldsymbol{\Theta}_\phi^* = \arg\min_{\boldsymbol{\Theta}_\phi} \mathcal{L}_\phi(\boldsymbol{y}, \hat{\boldsymbol{y}})$. In practice, the network weights are iteratively updated in the direction of the steepest descent to reach the local minimum, following:

$$\boldsymbol{\Theta}_\phi \leftarrow \boldsymbol{\Theta}_\phi - \alpha \nabla_{\boldsymbol{\Theta}_\phi} \mathcal{L}_\phi \tag{1}$$

where the learning rate α is a hyper-parameter controlling the step size at each iteration. Tuning α is of paramount importance to find a good trade-off between convergence speed and stable optimization. Back-propagation deals with gradient computation, while the gradient descent algorithm, based on this gradient, aims at performing the learning procedure. ℓ_ϕ is a per-image loss function which is usually the cross-entropy loss defined, in a multi-class scenario with C classes, as follows:

$$\ell_{CE}(\boldsymbol{y}_n, \hat{\boldsymbol{y}}_n) = \frac{1}{|\mathscr{C}||\Omega|} \sum_{c \in \mathscr{C}} \sum_{u \in \Omega} -\boldsymbol{y}_{n,c,u} \log(\hat{\boldsymbol{y}}_{n,c,u}) \tag{2}$$

where Ω is the image grid and $c \in \mathscr{C}$ a given class with $\mathscr{C} = \{0, ..., C\}$ indexing the different structures of interest as well as the background. As reviewed in [42], a wide variety of loss functions exist including distribution-based (e.g. cross-entropy), region-based (e.g. Dice), compound (e.g. DiceCE) and boundary-based (e.g. Hausdorff distance) losses.

B. Seminal works

The simplest and early attempts to perform segmentation using CNN consisted in classifying each pixel individually in a patch-based manner [43]. Since input patches from neighboring pixels have large overlaps, the same convolutions were computed many times. By replacing fully-connected layers with convolutional layers, fully convolutional networks (FCN) gave the opportunity to take entire images as inputs and produce likelihood maps instead of single-pixel outputs. It removed the need to select representative patches and eliminated redundant calculations due to patch overlaps. In order to avoid outputs with far lower resolution than input shapes, FCN were applied to shifted versions of input images [44]. Multiple resulting outputs were thus stitched together to get results at full resolution. Further improvements were then proposed with architectures comprising a regular FCN to extract features, followed by an up-sampling part that enables the recover the input resolution using up-convolutions [16]. Compared to patch-based or shift-and-stitch methods, precise localization was possible in a single pass while taking into account the full image context. This motivated the strong interest devoted to convolutional encoder-decoders among which U-Net (Sect.II-C) is the most commonly used representative.

C. U-Net

Among existing convolutional encoder-decoders architecture, most DL-based medical image segmentation models are based on U-Net [14] and its 3D counterpart V-Net [45]. U-Net and V-Net consist of symmetrical architectures comprising an encoder that gradually reduces the spatial dimension using pooling layers, a decoder progressively recovering object details and initial resolution as well as skip-connections (i.e. long-range shortcuts) which concatenate features between contracting and expanding paths to help in improving localization accuracy and convergence speed. The contracting path encoder of a standard U-Net (resp. V-Net) architecture consists of sequential layers including 3×3 (resp. $3\times3\times3$) convolutional layers followed by batch normalization (BN) and rectified linear unit (ReLU) activations (Fig.1). Spatial size is reduced using 2×2 (resp. $2 \times 2 \times 2$) max-pooling layers. The first convolutional layer typically generates 32 or 64 channels and this number doubles after each pooling as the network deepens. The encoder finally projects each input greyscale image $\boldsymbol{x}_n$ to a latent representation (denoted as $\boldsymbol{z}_n$ in Fig.1). On its turn, the decoder branch is built symmetrically with respect to the encoder, except that max-pooling layers are replaced by up-sampling operations (e.g. bi/tri-linear interpolation, transpose convolution). Depending on the binary or multi-class nature of the segmentation issue at hand, a final 1×1 (resp. $1\times1\times1$) convolutional layer with sigmoid or softmax activation achieves pixel-wise segmentation $\hat{\boldsymbol{y}}_n = \phi(\boldsymbol{x}_n)$, at native resolution. V-Net-inspired models may more suffer from high computational cost and GPU memory usage than their 2D counterparts.

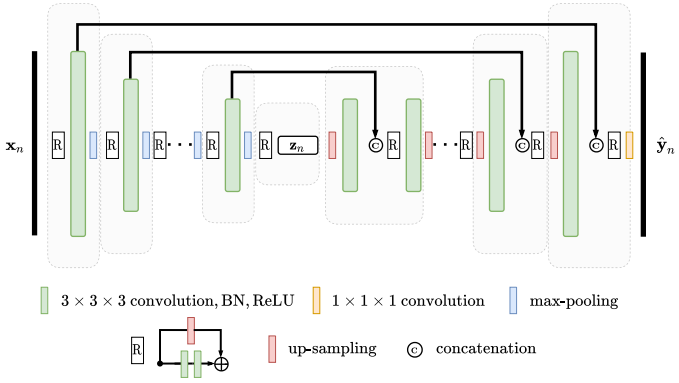

Fig. 1. Residual V-Net inspired convolutional encoder-decoder architecture for medical image segmentation purposes. Refer to Sect.II-C for further details.

Numerous refinements to the U-Net encoder-decoder architecture have been proposed including, to name a few, models which embed encoders pre-trained on large non-medical imaging databases (e.g. ImageNet) to leverage low-level features typically shared between different image types [46], sequential models exploiting residual convolutions [45] (Fig.1) or pyramidal atrous convolutions (instead of pooling operations) [47] as well as alternative attention models (Sect.IV-E) such as attention U-Net [26] which integrates attention gates on skip-connections to focus on salient features. As an extension to vanilla U-Net, U-Net++ [48] relied on re-designed skip-connections through intermediate convolution layers as well as deep supervision (Sect.IV-D). By aggregating features of varying semantic scales at the decoder branch, nested and dense skip-connections act as a flexible feature fusion scheme.

D. Data augmentation

Deep segmentation models are most often trained with extensive on-the-fly data augmentation, towards improved generalization properties. By comprising random geometric

transformations (e.g. translation, rotation, scaling, shear, flipping) and random intensity modifications (e.g. normalization, blurring, contrast adjustment), data augmentation can be seen as a clever way to artificially increase the amount of available data, with slightly modified copies of already existing images. In practice, geometric transformations are applied to both greyscale images and ground truth masks whereas intensity transformations only modify source images. Data augmentation enables to teach DL models desired invariance, covariance and robustness properties and to strongly reduce over-fitting.

More recently, it was shown that DL models could further benefit from more elaborated data augmentation techniques such as MixUp which exploits convex combinations of pairs of samples and associated labels to train neural networks [49]. MixUp regularizes the neural network to favor simple linear behavior in-between training examples. Originally proposed for image classification tasks, its extension to image segmentation is straightforward and efficient, as proven in [50] where improved generalization with MixUp as a data augmentation technique is reached for knee MR segmentation purposes. This success at the input data space further inspired the use of MixUp in the latent feature space [51], in a setting referred to manifold MixUp. As reviewed in [52], synthetic augmentation based on image synthesis methods, for instance exploiting generative adversarial networks (GAN), is another powerful alternative to standard data augmentation since it samples the manifold on which the original training set resides. Although effective, especially in extreme data scarcity scenarios, synthetic augmentation is more demanding to implement. Standard and more sophisticated data augmentation systems are obviously not mutually exclusive and can be used together.

While data augmentation is typically employed during training, using it at test time recently started to get special attention. Strongly linked to the way model uncertainty can be quantified [31] (Sect.V-C), test-time data augmentation consists in performing the inference both on original and augmented versions of images, followed by a merging procedure. Gains in performance are reported in various clinical contexts such as lesion segmentation from whole-body PET-CT images [53].

III. Clinical needs and applications

An ever-increasing number of research studies have illustrated the numerous applications of medical image analysis with DL, targeting a large number of pathologies and imaging modalities [16, 21]. On its side, medical image segmentation plays a key role in many medical imaging workflows tailored for diagnosis, disease progression assessment, surgery or therapeutic planning, follow-up, survival analysis, treatment response evaluation, dosimetry and many other applications.

Clinical needs deal, first of all, with organ delineation from anatomical CT or MR imaging given that clinical parameters (e.g. volume, shape, inner textures) can be exploited as biomarkers to diagnose or quantify disease progression, as in cardiac [54], brain [55] or prostate [56] disorders. Hepatic pathologies with primary or secondary liver lesions are also concerned, thus making fully-automatic liver segmentation [57] particularly useful and requested in clinical routine. Regarding pure organ volumetry, a good example is an automated assessment of the total kidney volume (TKV) from MR images in patients with polycystic kidney disease since TKV is the main image-based biomarker to follow PKD progression [58]. Segmenting healthy organs (e.g. liver, kidneys) to obtain a measurement of volume, size or shape is also a relevant use-case in the context of transplant surgery planning [59, 60].

In other works, whole or sub-structure organ segmentation is managed as a first step toward lesion detection and delineation. The main related challenge in this context deals with class imbalance as most voxels usually belong to the non-diseased class. In particular, there are numerous research works aimed at delineating skin lesions from dermatological images [30, 61], brain tumors from MR images [47, 62, 63], liver lesions from CT scans [57], head and neck primary tumors, lymph nodes and organs at risk from radiotherapy computed tomography (RT-CT) or combined PET and CT images [64–66], breast masses in mammograms [67] or ultrasound images [68], cystic kidney tissues in MR modality [69] or lesions in whole-body images [53]. In oncology, PET and CT imaging held a special place for disease characterization since they contain complementary information about the metabolic or biochemical function of tissues and organs as well as the anatomy of cancer [70]. Inner-lesion tissue segmentation is also increasingly targeted as in [71] where both active and necrotic tissues are identified inside liver tumors for patients with hepatocellular carcinoma in dynamic contrast-enhanced CT or in [72] where low and high-grade gliomas are decomposed into several tissue types comprising necrotic and enhancing cores, non-enhancing tumor and oedema.

Another emerging application deals with automatically extracting blood vessels (e.g. retinal, brain, liver vessels) from medical images [73, 74]. Apart from class imbalance and appearance similarity with non-vascular tissues, vascular segmentation brings additional limitations: complex multi-scale geometry with decreasing diameters and contrast along tree-like networks, inter-patient variability in branching patterns...

Medical image segmentation is also involved in plenty of radiomics pipelines where it has been for a very long time the bottleneck in both time and automation. Thus, extracting radiomic features in an automated and high-throughput way from relevant lesion areas is requested to quantify the characteristics of medical images, comprehensively characterize objects (e.g. tumors, organs, tissues) and finally provide useful guidance for clinicians. Although initially designed to process CT and functional PET images, the radiomics approach can be applied to any imaging modality or radio-tracer [75]. This includes works involving automated DL-based segmentation towards patient outcome prediction such as survival analysis [70] or chemotherapy response assessment and prediction [76].

More marginal applications can also be mentioned, as for the management of musculoskeletal diseases where patient-specific information related to the degree of muscle atrophy across joints is needed to plan interventions and predict interventional outcomes. In particular, DL-based shoulder muscle segmentation from MR images [46] can be employed to analyze the shoulder strength balance, which is particularly important given that a clear relationship between muscle atrophy and strength loss [77] has been established.

IV. Current trends

A. Conditional generative adversarial networks

A network design based on conditional generative adversarial networks (GAN) has been proposed as a general-purpose solution for image-to-image translation [78]. The goal is not only to learn the mapping from input to output images but also to learn a loss function to train this mapping. This kind of strategy is obviously suitable for segmentation purposes and among the possible applications, its feasibility for medical image segmentation has been demonstrated in several recent works [22, 60, 61, 79]. In practice, conditional GAN architectures comprise a generator aiming at providing segmentation masks through encoding and decoding layers as well as a discriminator which assesses if a given segmentation mask is synthetic or real. The adversarial network learns to discriminate real from synthetic delineations, i.e. ground truth masks versus those arising from the generator. This enforces the generative part to create increasingly plausible segmentation masks. During the training process, the generated delineations are gradually close to the ground truth, to the point of being able to deceive the discriminator. Unlike standard post-processing schemes, such iterative refinement performed through adversarial learning [79] is conducted in an end-to-end manner.

As generator ϕ, conditional GAN pipelines may use any type of U-Net [14] inspired model, from simple [22] to extended (using dense dilated convolution) [61] and cascaded (Sect.IV-B) [60] ones. The inputs of the discriminator D are the concatenation of source images and ground truth or predicted masks to be evaluated. Defined between 0 (fake) and 1 (plausible or real), the output of D is an array where each value corresponds to the degree of segmentation likelihood for a given image crop and its associated segmentation mask. Let $\phi(\boldsymbol{x})$ and $D(\boldsymbol{x}, \phi(\boldsymbol{x}))$ be the outputs of ϕ and D respectively. The loss function $\mathcal{L}_\phi(\boldsymbol{\Theta}_\phi, \boldsymbol{\Theta}_D)$ for the generator ϕ can be defined as the following combination:

$$\mathcal{L}_\phi(\boldsymbol{\Theta}_\phi, \boldsymbol{\Theta}_D) = \frac{1}{N}\sum_{n=1}^{N} \ell_{CE}(\phi(\boldsymbol{x}_n), \boldsymbol{y}_n) + \lambda \ell_{adv}(\phi(\boldsymbol{x}_n), \boldsymbol{y}_n) \tag{3}$$

where λ is an empirically set weighting factor and $\boldsymbol{\Theta}_D$ the trainable parameters of D. The adversarial term $\ell_{adv}(\phi(\boldsymbol{x}_n), \boldsymbol{y}_n)$ equals to $-\log(D(\phi(\boldsymbol{x}_n), \boldsymbol{x}_n))$. Minimizing ℓ_{CE} tends to provide rough predictions whereas maximizing $\log D(\phi(\boldsymbol{x}_n), \boldsymbol{x}_n)$ aims at improving contour delineations. Conversely, the optimizer typically fits D through cross-entropy using both estimated and ground truth masks. The loss function $\mathcal{L}_D(\boldsymbol{\Theta}_\phi, \boldsymbol{\Theta}_D)$ for D is therefore defined as:

$$\begin{aligned}\mathcal{L}_D(\boldsymbol{\Theta}_\phi, \boldsymbol{\Theta}_D) = \frac{1}{N}\sum_{n=1}^{N} & -\log(D(\boldsymbol{y}_n, \boldsymbol{x}_n)) \\ & -\log(1 - D(\phi(\boldsymbol{x}_n), \boldsymbol{x}_n))\end{aligned} \tag{4}$$

The above equation maximizes the loss values for ground truth (i.e. $\log(D(\boldsymbol{x}, \boldsymbol{y}))$) and minimizes loss values for generated masks (i.e. $-\log(1 - D(\boldsymbol{x}, \phi(\boldsymbol{x})))$). The optimization process is performed sequentially by alternating gradient descents on ϕ and D at each batch [81]. To further improve the ability of conditional GAN architectures to extract the contours of the targeted anatomy or abnormalities, investigations on more robust generators than traditional U-Net [14] is an interesting research avenue [60]. Condition GAN can also be employed as a tool combined with other constraints (Sect.IV-C) such as anatomical shape priors, as proposed in [24].

B. Cascaded networks

Managing long-range spatial context from medical images is an important feature to improve the automatic delineation process. However, increasing the network depth over and over to exploit larger receptive fields is not suitable for memory and computational reasons, especially given the volumetric nature of most medical imaging data. In addition, too many high-resolution details are discarded when the number of down-sampling operations in the encoder branch is significant. To address these challenging limitations, standard scale-space pyramid [82] and auto-context [83] ideas influenced the development of cascaded strategies exploiting series of convolutional encoder-decoders [23, 60]. Instead of simultaneously exploiting several pathways working at various scales [84], cascaded approaches consist in using a scale-space pyramid to perform segmentation at a higher resolution while also considering contextual information from lower resolutions.

By considering two convolutional encoder-decoders in cascade, the most common setup could consist in training a low-resolution model and using its weights as initialization of a high-resolution model through transfer learning and fine-tuning. Although this strategy can significantly speed up convergence, the ability of the high-resolution segmentation model to extract long-range contextual features remains limited. The idea of stacking (at least) two convolutional encoder-decoders to integrate multi-level information more directly came naturally [23] and made use of auto-context [83] such that posterior probabilities resulting from a given model can be used as features for the following one [55]. The models can be trained separately [55] but this prevents refining low-resolution models from the high-resolution ones during back-propagation. An end-to-end training is better suited to exploit simultaneous multi-level segmentation refinement [23, 60]. A cascade of deep modules exploiting tissue-specific geodesic distance maps as contextual information was employed in [85] to gradually improve the delineation accuracy. Combined with top-down reasoning, such bottom-up strategies could better handle texture information and region discontinuities.

C. Prior knowledge embedding

Regularization plays a key role in DL since it tends to increase both robustness and generalizability of a deep model when applied to unseen data. One common strategy consists in adding a regularization term into the loss function to get more accurate and plausible results [86]. The regularization term $\mathcal{R}_\phi$ deals with adding some prior knowledge to the model ϕ and its regularization effect is achieved by adding the scaled regularizer $\lambda\mathcal{R}_\phi$ to the loss function $\mathcal{L}_\phi$ to ensure further consistency between both predictions $\phi(\boldsymbol{x})$ and targets $\boldsymbol{y}$. The resulting loss function can be expressed as follows:

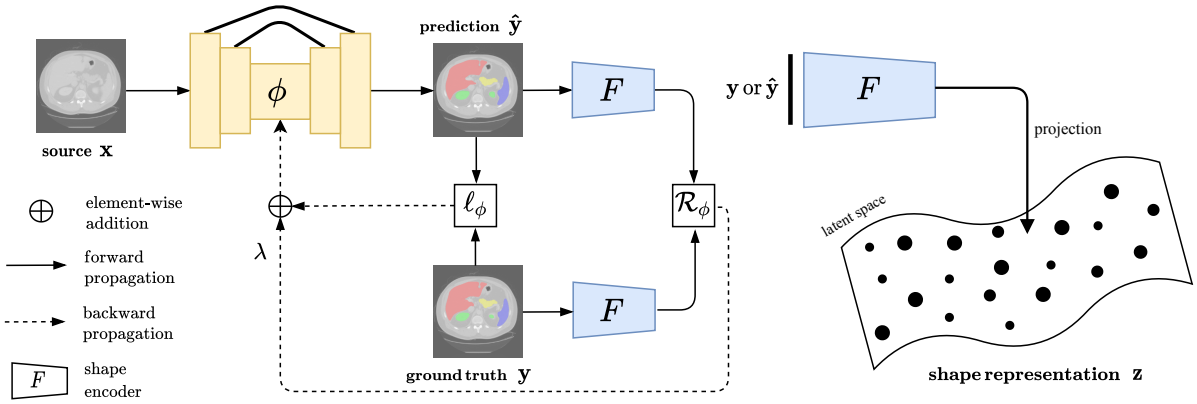

Fig. 2. Integration of anatomical shape priors into a deep segmentation pipeline [24, 26, 35, 80]. Shape priors-based regularization is performed using a shape encoder arising from a convolutional auto-encoder previously optimized on ground truth segmentation masks. Refer to Sect.IV-C for further details.

$$\mathcal{L}_\phi(\mathbf{\Theta}_\phi) = \frac{1}{N}\sum_{n=1}^{N} \ell_\phi(\phi(\boldsymbol{x}_n), \boldsymbol{y}_n) + \lambda \mathcal{R}_\phi \tag{5}$$

Many different types of information can be incorporated as prior knowledge into DL frameworks such as shape constraints [26], topology specifications [73], edge polarity [87] or adjacency rules between regions [88]. Nevertheless, integrating shape priors remains one of the most commonly used strategies toward anatomically meaningful predictions. In particular, the use of convolutional auto-encoder (AE) to learn anatomical shape variations from medical images has been demonstrated in multiple applications [24, 26, 35, 80]. Specifically, a convolutional AE is a deep network made of an encoder $F : \boldsymbol{y} \mapsto F(\boldsymbol{y}; \mathbf{\Theta}_F)$ and a decoder $G : F(\boldsymbol{y}; \mathbf{\Theta}_F) \mapsto G(F(\boldsymbol{y}; \mathbf{\Theta}_F); \mathbf{\Theta}_G)$ where $\mathbf{\Theta}_F$ and $\mathbf{\Theta}_G$ correspond to the learnable parameters of F and G respectively. F maps the input to a low-dimensional feature space whereas G reconstructs the original input from the compact representation. To avoid the AE to copy the input, F is usually designed to be undercomplete such that the latent space is much smaller than the input dimension. By penalizing the reconstruction $G \circ F(\boldsymbol{y})$, the cross-entropy loss function can be employed to optimize the AE following:

$$\mathbf{\Theta}_F^*, \mathbf{\Theta}_G^* = \underset{\mathbf{\Theta}_F, \mathbf{\Theta}_G}{\arg\min} \frac{1}{N}\sum_{n=1}^{N} \ell_{CE}((G \circ F)(\boldsymbol{y}_n), \boldsymbol{y}_n) \tag{6}$$

where $\mathbf{\Theta}_F^*$ (resp. $\mathbf{\Theta}_G^*$) are the optimal weights of the encoder (resp. decoder). After having trained the AE, its encoder component acts as a non-linear shape model and can project any predicted or ground truth segmentation masks to a shape manifold space [80]. The encoder produces a feature map $F(\boldsymbol{y})$ which compactly encodes the most salient characteristics of $\boldsymbol{y}$.

Once the AE trained, its encoder component can be integrated into the segmentation pipeline (Fig.2). A regularizer $\mathcal{R}_\phi$ that penalizes the deviation between predicted and ground truth segmentation masks fed as inputs of the learned shape model F is included in the global loss (Eq.5). A Euclidean distance between both latent shape representations [26] is usually used:

$$\mathcal{R}_\phi(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{N}\sum_{n=1}^{N} \|F(\boldsymbol{y}_n; \mathbf{\Theta}_F^*) - F(\phi(\boldsymbol{x}_n); \mathbf{\Theta}_F^*)\|_2^2 \tag{7}$$

Nevertheless, a cosine distance between predicted and ground truth masks in low-dimensional space is also of interest [74].

$$\mathcal{R}_\phi(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{N}\sum_{n=1}^{N} 1 - \cos(F(\boldsymbol{y}_n; \mathbf{\Theta}_F^*), F(\phi(\boldsymbol{x}_n); \mathbf{\Theta}_F^*)) \tag{8}$$

As reported in [80], shape is just one of the geometric attributes of anatomical or pathological structures one can exploit. Much more meaningful priors such as texture, topology or size can be embedded into training objectives towards stronger robustness and stability of DL segmentation networks.

D. Deep supervision

Introduced in the context of holistically-nested edge detection [89], additional convolutional operations can be applied at different levels of the decoder branch in order to exploit a deep supervision mechanism (Fig.3) able to boost the segmentation performance [15]. Companion objective functions are estimated at some hidden layers of the network and added to the output loss. In practice, feature maps as outputs of each intermediate decoder blocks can be up-sampled to the size of the input image using bi- or tri-linear interpolation, depending of the 2D or 3D nature of the segmentation problem. Similarly to [25], a convolutional operation (e.g. with 3×3 kernel) can be applied to create feature maps at each level of the decoder (16 in Fig.3). These maps then go through deep supervision modules to improve the gradient flow and encourage learning more useful representations [25]. After having performed the concatenation of these intermediate outputs, two convolutional layers including a final 3×3 one with softmax activation finally achieve pixel-wise segmentation (multi-label in Fig.3). In this context, the overall loss function $\mathcal{L}_\phi$ can be defined as the weighted sum of the cross-entropy losses (or any other losses) estimated at different decoder levels involving supervision:

$$\mathcal{L}_\phi(\mathbf{\Theta}_\phi) = \sum_{j=1}^{M} w_j \mathcal{L}_{CE}^j + w_f \mathcal{L}_{CE}^f \tag{9}$$

where w_j and $\mathcal{L}_{ce}^j$ denote the weight and loss for the points of supervision at level j of the decoder, w_f and $\mathcal{L}_{CE}^f$ the weight and loss computed at the final network output (where

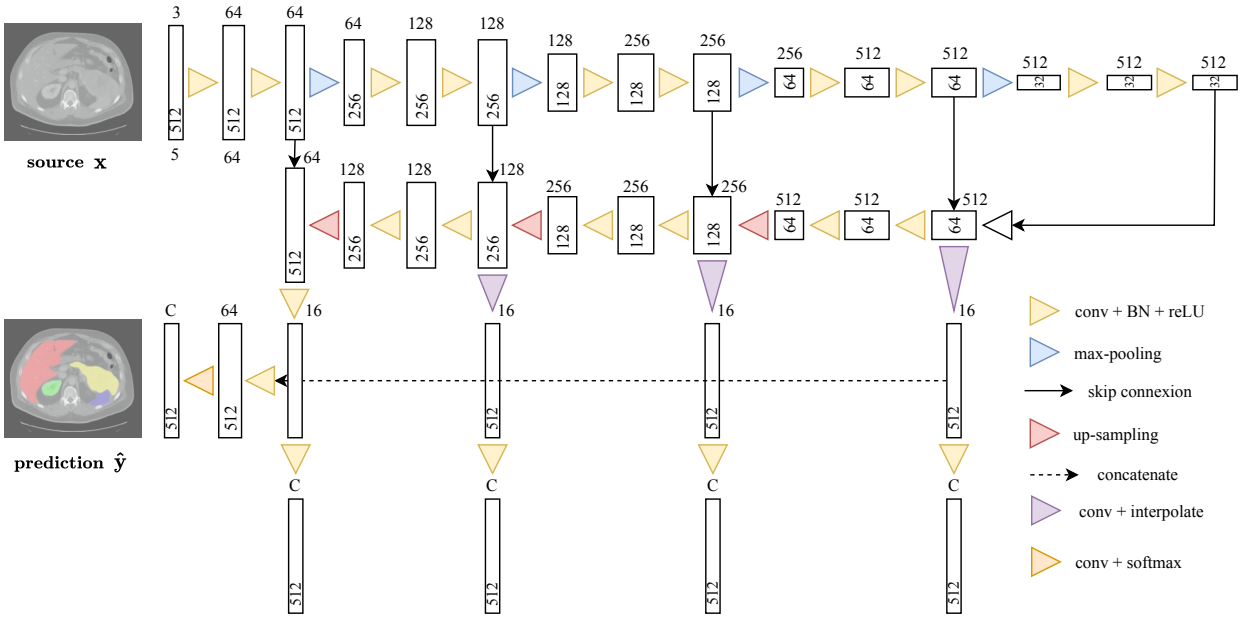

Fig. 3. Convolutional encoder-decoder architecture with deep supervision. The overall loss function is the weighted sum of losses estimated at different decoder levels [25]. C is the number of classes. Refer to Sect.IV-D for further details.

f stands for final). When following a VGG-13 architecture [90] as depicted in Fig.3, $M = 4$ intermediate decoder levels can be considered. Following [25], one can use $w_1 = 0.8$, $w_2 = 0.7$, $w_3 = 0.6$, $w_4 = 0.5$ and $w_f = 1$ where level $j = 1$ is closer to the network ending part than level $j > 1$. However, how to balance the hyper-parameter setting among different loss components remains a matter of concern. Instead of empirically defining weights, a relative weighting can be learned from the data using homoscedastic uncertainty [91].

E. Attention mechanisms

The human visual system can concentrate and focus actively on a tiny portion of highly relevant perceptible information while disregarding other irrelevant perceivable stimuli. Attention mechanisms were introduced in DL frameworks to imitate this aspect of how the human visual system processes information. In general, they can be regarded as a dynamic selection process where the features extracted from images are adaptively weighted to pay attention to the more salient ones, i.e. the feature needed for accurately solving a specific image analysis task. Attention mechanisms, particularly in image segmentation, can suppress feature responses in irrelevant background regions, hence reducing the rate of false-positive predictions. This is particularly true for the challenging instances of small objects with high shape variability.

The attention problem is usually formulated using three vectors: query, key and value. Conceptually, we can think of key and value as a look-up table in which the query is matched to a key, and the value associated with that key is returned. In image segmentation, it is equivalent to mapping the features of the structure to segment (query) against a collection of plausible target features (keys), then presenting the best-matched regions (values). Mathematically, let us consider that

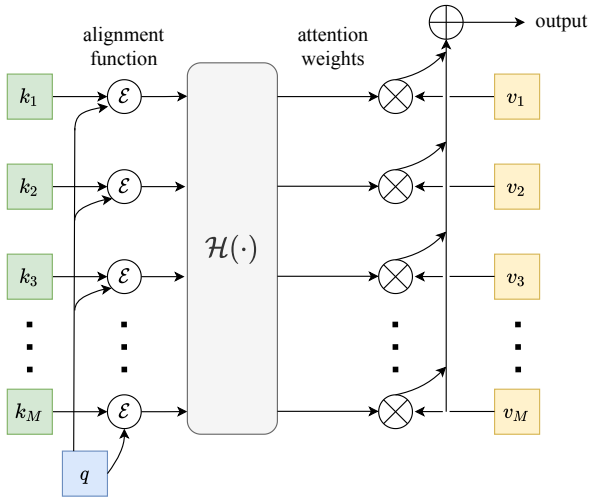

Fig. 4. Flow-chart diagram of a general attention mechanism function. Refer to Sect.IV-E for further details.

we have a query $q \in \mathbb{R}^{d_q}$ and M pairs of key $k \in \mathbb{R}^{d_k}$ and value $v \in \mathbb{R}^{d_v}$ vectors $\{(k_1, v_1), \dots (k_M, v_M)\}$, where all of them can be obtained from intermediate CNN features or embedding patches of an input image $\boldsymbol{x}$. The attention is computed step-by-step following three operations: alignment, weighting and contextualization. In the alignment step $\mathcal{E}(\cdot)$, each query is matched against the M keys to compute a score value. Several commonly used alignment functions are further described in Tab.II where additive [92] and dot-product [93, 94] functions are the most widely used. In the next step, the alignment scores are passed through a function $\mathcal{H}(\cdot)$ (e.g. softmax, sigmoid) to generate the final attention weights by normalizing all the scores to a probability distribution (Eq.10). A contextualization vector (Eq.11) is then instantiated for each q as a weighted sum of the M values v_i by the set of weights

TABLE II
SUMMARY OF SEVERAL POPULAR ALIGNMENT FUNCTIONS USED TO COMPUTE THE MATCHING SCORE BETWEEN A QUERY AND KEYS. $\mathbf{W}$ AND $\mathbf{V}$ ARE TRAINABLE WEIGHT MATRICES, MEANWHILE d_k STANDS FOR THE DIMENSION OF A VECTOR k. $[.;.]$ STANDS FOR CONCATENATION.

Name	Alignment function
Additive or concatenation	$\mathcal{E}(q,k) = \boldsymbol{V}^T \tanh(\boldsymbol{W}[q;k])$ [92]
Content-based attention	$\mathcal{E}(q,k) = \cos(q,k)$ [95]
Location-based attention	$\mathcal{E}(q) = \boldsymbol{W}q$ [93]
General	$\mathcal{E}(q,k) = q\boldsymbol{W}k$ [93]
Dot-product	$\mathcal{E}(q,k) = q^T k$ [93]
Scaled dot-product	$\mathcal{E}(q,k) = \frac{q^T k}{\sqrt{d_k}}$ [94]

$\mathcal{A}$. This computation enables to define how much attention should be paid to each feature v_i. Fig.4 depicts the procedure followed to get the context vector for a particular query q.

$$\mathcal{A}(q,k_i) = \mathcal{H}(\mathcal{E}(q,k_i)) \tag{10}$$

$$\mathcal{C}(q,\{(k_1,v_1),\dots(k_M,v_M)\}) = \sum_{i=1}^{M} \mathcal{A}(q,k_i) \times v_i \tag{11}$$

Four main categories of attention techniques can be found in the literature dedicated to medical image segmentation: channel attention (what to attend), spatial attention (where to attend), branch channel (which to attend) as well as hybrid (e.g. channel and spatial attention) methods.

Channel attention is based on the idea that, in deep CNN models, each channel represents a different feature map that typically denotes distinct objects. As a result, the role of channel attention is to adaptively calibrate the weight of each channel, serving as an object selector of the entities that should deserve more attention. Channel attention, particularly, squeeze-and-excitation (SE) block [96], has proved to be extremely effective in tasks such as head and neck primary tumor segmentation [97], prostate zonal segmentation [98], brain structure segmentation [99] and micro-vessel segmentation [100]. The SE block, depicted in Fig.5, passes an intermediate features map through a squeeze operation (i.e. global average pooling) that captures global spatial information. Then, an excitation module (alignment function) captures channel-wise relationships and outputs an attention vector using fully-connected and non-linear layers (e.g. ReLU), followed by a sigmoid function (weighting). Lastly, each channel of the input features map is scaled by multiplying the corresponding element in the attention vector (contextualization). Others works including [26] extended skip-connections between both encoder and decoder branches through channel attention gates. By scaling the encoder features by importance, the network may concentrate on a specific aspect of the input to generate the segmentation. A similar approach was presented in [101] for lesion segmentation. This model also used the extracted feature maps from the encoder path for the computation of attention weights which are afterward merged with the up-sampled feature maps in the decoder branch.

Spatial attention and channel attention have relatively similar functioning. Spatial attention consists in adaptively calibrating the weight of each part of the image. This mechanism

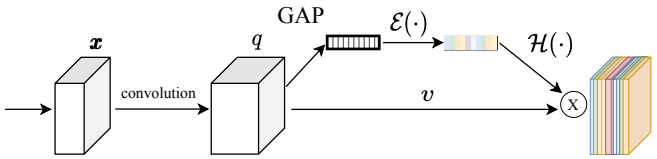

Fig. 5. Flow-chart diagram of a squeeze-and-excitation block [96]. GAP stands for global average pooling.

chooses where to focus attention through an adaptive spatial area selection procedure. One may integrate an attention block into the U-Net architecture to learn semantic representations that prioritize spatial regions with high saliency levels for the task of tumor segmentation [68]. In the same direction, the addition of a spatial-channel squeeze-and-excitation block can improve the performance of various convolutional architectures dedicated to medical image segmentation [102].

Branch attention mechanisms, on its side, separate the attention problem into several sub-modules (branches), each of which focuses on a certain aspect (e.g. channel, spatial, resolution, degradation) while exchanging effective information. Then, the role of the attention mechanism is to adaptively calibrate the weight of each branch, acting as a dynamic branch selection procedure, choosing which to focus on. To segment scleral blood vessels, Yao et al. [103] developed a U-Net-inspired architecture with deep feature concatenation and an attention mechanism branching into numerous attention gates. This enables the network to focus on the border segmentation of tiny blood veins. Other works such as [56] and [104] exploited the multi-scale nature of CNN models. They generated series of multi-scale attention modules at different resolutions and integrated local deep attention features with a global context. Lastly, channel and spatial attention incorporated the benefits of both attention mechanisms. This system acts as a dynamic spatial area and object selection mechanism, deciding what and where to focus attention [105, 106].

F. Multi-structure analysis

The multi-structure analysis aims at incorporating inter-structure relations into a given DL-based segmentation model, thus leading to a more accurate representation of complex anatomies. A typical application in this direction deals with multi-organ segmentation from CT [23, 40, 107, 108], PET [109] or MR [60, 107, 109] images. Although computationally effective, global approaches which consider one single model for all structures of interest are computationally effective, extracting high-level relationship patterns between multiple structures but cannot fully exploit the local characteristics of each different component to delineate. Conversely, modeling multiple structures by a set of structure-specific models enables generating local structure-specific features but sacrifices the ability to take advantage of inter-structure relationships.

In their survey, Cerrolaza et al. reported that multi-level (e.g. nested, multi-resolution) or sequential models are robust alternatives to global or individual models as they combine the robustness and specificity of global approaches with the flexibility of structure-specific models [27]. Multi-level nested models decompose the data into different levels of detail

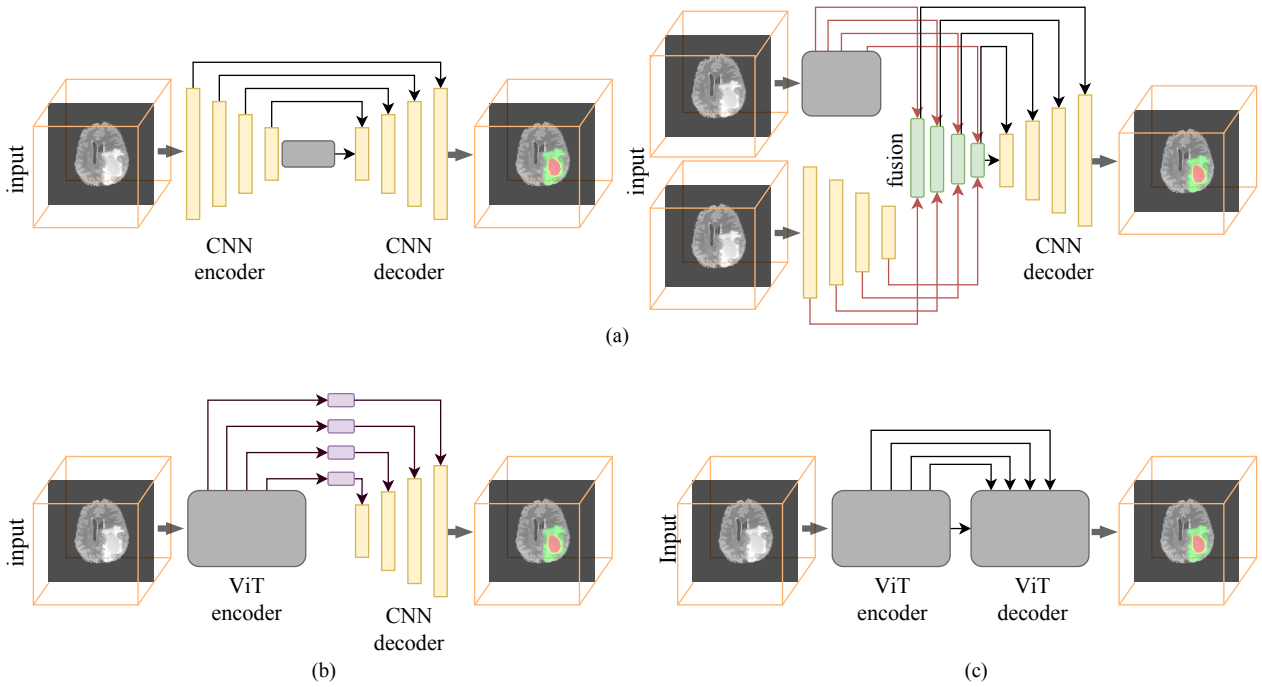

Fig. 6. Typical Transformers-based model architectures for medical image segmentation: (a) hybrid Transformers-CNN encoder, (b) pure Transformers-based encoder, (c) full Transformers-based network. Refer to Sect.V-A for further details.

according to coarse-to-fine analysis rules. Conversely, in multi-level multi-resolution models, global inter-organ relations are modeled at coarser resolutions while local organ-specific variations are extracted from higher resolutions. On its side, sequential modeling deals with the analysis of multiple structures following a pre-defined order of increasing complexity. In the same vein, it is also relevant to mention the application related to lesion segmentation for which delineating the organ of interest as a first stage before localizing its inner lesions is usually followed to promote the extraction of organ-specific features and to narrow the search area (Sect.III). As an example, such two-step sequential approach has been successfully applied in [57] for liver lesion segmentation from CT scans.

G. Learning frameworks

The success of DL in medical image segmentation not only originates from the development of novel learning paradigms but also from the network architecture design itself and the focus given to data management and optimization processes. Especially, a trend can be noticed towards the development of unified frameworks such as NiftyNet [110] or nnU-Net [28]. Since the design choices towards an optimal framework are usually dedicated to a specific segmentation task (i.e. a given tissue type for a given modality) and cannot easily be transferred to another application, pipelines that can configure their sub-components in an automated fashion are highly requested. In particular, nnU-Net [28] which has been designed to deal with the dataset diversity found in the domain has proven its efficiency by winning many challenges. It condenses and automates the key decisions for designing a successful pipeline for any given dataset. Thus, nnU-Net has become one of the reference frameworks when targeting a new segmentation task.

In the same scope, neural architecture search (NAS) is another direction under investigation with the goal to automate the iterative network design process usually handled manually by researchers. Among the existing research works in this area, NAS-UNet [111] is based on the design of three types of primitive operation set on search space to automatically find two cell architectures (DownSC, UpSC) for semantic segmentation purposes. Promising results were reported for various imaging modalities including CT, MR and ultrasound.

V. Emerging trends

A. Medical Transformers

Transformers, as attention-based structures [94], have first demonstrated their tremendous force in natural language processing (NLP) [112, 113] and have gradually gain attraction on different computer vision tasks such as image classification, detection, segmentation and video analysis. Their popularity is now also rapidly growing in medical image analysis [114], especially for medical image segmentation with an exponential growth of related publications in the last year [15, 115]. The pioneering work of vision Transformers [116] was an interesting and meaningful attempt to replace convolutional backbones with convolution-free models. In contrast to CNN, vision Transformers (ViT) offer parallel processing and a complete field of view in a single layer.

ViT has a columnar structure where the 3D input volume $\boldsymbol{x} \in \mathbb{R}^{H \times W \times D \times C}$ is split into n_p 3D non-overlapping patches $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_{n_p}\}$ with $\boldsymbol{x}_i \in \mathbb{R}^{P \times P \times P \times C}$, C represents the number of modalities, (P, P, P) is the resolution of each patch and $n_p = HWD/P^3$ is the resulting number of patches, which is also the effective input length of the Transformer. Since Transformer layers operate over fixed size

1D set of vectors, the n_p patches are flattened and mapped to a d-dimensional embedding space through a trainable linear projection matrix $\boldsymbol{E} \in \mathbb{R}^{(P^3 \cdot C) \times d}$. To preserve spatial information, a 1D positional embedding $\boldsymbol{E}_{pos} \in \mathbb{R}^{n_p \times d}$ is added to each of the n_p patches, and the resulting sequence of embedding is used as input to the Transformer encoder:

$$\boldsymbol{z}_0 = [\boldsymbol{x}_1\boldsymbol{E}; \boldsymbol{x}_2\boldsymbol{E}; \boldsymbol{x}_3\boldsymbol{E}; \ldots; \boldsymbol{x}_{n_p}\boldsymbol{E}] + \boldsymbol{E}_{pos} \tag{12}$$

The Transformers-based encoder consists of alternating L layers of multi-head self-attention (MSA) and multilayer perceptron (MLP) blocks. A layer normalization (LN) is applied before each block and a residual connection after each block. One layer of a Transformer block can be formulated as:

$$\begin{aligned} \boldsymbol{z}'_l &= \text{MSA}(\text{LN}(\boldsymbol{z}_{l-1})) + \boldsymbol{z}_{l-1} \\ \boldsymbol{z}_l &= \text{MLP}(\text{LN}(\boldsymbol{z}'_l)) + \boldsymbol{z}'_l \end{aligned} \tag{13}$$

with $l = \{1, \ldots, L\}$. The MSA block consists of h parallel self-attention (SA) heads, where each SA head attends to bring information from different representation sub-spaces at different positions through a scoring function $\mathcal{A}$. To achieve this goal, an input sequence $\boldsymbol{z} \in \mathbb{R}^{n_p \times d}$ is mapped into query ($Q \in \mathbb{R}^{n_p \times d_k}$), key ($K \in \mathbb{R}^{n_p \times d_k}$) and value ($V \in \mathbb{R}^{n_p \times d_v}$) matrices using three learnable parameters: $\boldsymbol{W}^q \in \mathbb{R}^{d \times d_k}$, $\boldsymbol{W}^k \in \mathbb{R}^{d \times d_k}$ and $\boldsymbol{W}^v \in \mathbb{R}^{d \times d_v}$.

$$\begin{aligned} Q &= \boldsymbol{z}\boldsymbol{W}^q \\ K &= \boldsymbol{z}\boldsymbol{W}^k \\ V &= \boldsymbol{z}\boldsymbol{W}^v \end{aligned} \tag{14}$$

Then, the attention distribution function is computed following Eq.15 and the resulting attention weights are applied to the V matrix obtaining the SA maps, as described in Eq.16.

$$\mathcal{A}(Q, K) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \tag{15}$$

$$\text{SA}(Q, K, V) = \mathcal{A}(Q, K) \times V \tag{16}$$

For MSA, Q, K, and V are computed once for each head using h different learnable parameters ($\boldsymbol{W}_i^{q,k,v}$), and the final attention map results from the concatenation of the h heads multiplied by a learnable aggregation matrix $\boldsymbol{W}^o \in \mathbb{R}^{hd_v \times d}$. The computational cost of single head attention with full d-dimension is maintained by setting d_k and d_v equal to d/h.

$$\text{MSA}(Q, K, V) = [\text{head}_1; \ldots; \text{head}_h]\boldsymbol{W}^o \tag{17}$$

with $\text{head}_i = \text{SA}(Q_i, K_i, V_i) = \text{SA}(\boldsymbol{z}\boldsymbol{W}_i^q, \boldsymbol{z}\boldsymbol{W}_i^k, \boldsymbol{z}\boldsymbol{W}_i^v)$. In the context of medical Transformers, U-Net-shaped architecture remains the preferred choice to build the Transformer segmentation models. From them and as illustrated in Fig.6, three categories can be identified [117]: pure Transformers-based encoder, hybrid Transformers-CNN encoder as well as full Transformers-based network.

1) Pure Transformers-based encoder: The first category exploits the global context modeling capability of Transformers to effectively encode the relationships between spatially distant voxels. A convolution-free encoder is introduced by forwarding flattened image representations to Transformers, whose outputs are then reorganized into 3D tensors followed by CNN up-sampling blocks with multi-level feature aggregation. For instance, [118, 119] employed a 3D ViT as an encoder and connected it to the CNN decoder via skip-connections. At the bottleneck of the encoder, the feature map was reshaped and up-sampled by a factor of 2. Then, the previous Transformer layer was used as a skip connection and concatenated with the resized feature to be later up-sampled through convolution, normalization and linear activation. This process was repeated until the initial resolution was reached. However, as anatomical structures can substantially vary in scale, they cannot be properly modelled using a set of fixed sub-regions of the image. Recently, hierarchical ViT such as Swin [120] or PVT [121] Transformers have been introduced to overcome these challenges by extracting features at different resolutions. This improves the performance of Transformers in dense prediction tasks while saving the linear computational complexity with respect to the image size. Hierarchical ViT architectures introduce CNN-like properties into the Transformers as they compute local attention with shifted windows, starting from small-sized patches and gradually merging neighboring patches in the subsequent layers. To reduce the design complexity of traditional hierarchical ViT, a 3D U-shape model inspired by nested hierarchical Transformers [122] exploited the idea of global self-attention within smaller non-overlapping 3D blocks [123]. Cross-block self-attention communication was achieved by hierarchically nesting these Transformers and connecting them with a specific aggregation function. Valanarasu et al. proposed in [124] a gated axial-attention model that extended previous designs by incorporating a new control mechanism in the self-attention module. Furthermore, the model operated on the whole image and patches to simultaneously learn both global and local features.

2) Hybrid CNN-Transformers encoder: This second category integrates the global context modeling ability of Transformers with the CNN inductive bias [29, 125, 126]. CNN layers capture the multi-scale context feature maps by stacking convolution blocks. Meanwhile, Transformers capture the long-term dependencies among the features that would be potentially lost with purely convolutional models. Lastly, a CNN-based decoder gradually up-samples the Transformers output into a 4D feature map to recover the full segmentation mask (Fig.7). Others approaches modified the traditional SA blocks using deformable Transformers [127] or squeeze-and-expansion Transformers layers [128]. In [129], Chen et al. proposed a 2D hybrid network that combines two independent self-attention blocks to model the long-range interactions and global spatial relationships. In addition, a multi-scale skip-connection scheme aggregated multiple features in the decoder at a different scale to generate more discriminative representations. PC-SwinMorph [130]

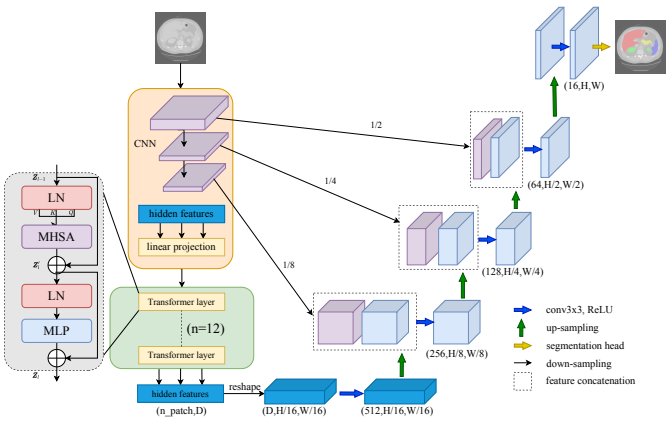

Fig. 7. Overview of the TransUNet architecture proposed in [29] including a schematic illustration of a Transformer layer.

performed registration-based segmentation using both CNN and patch-based contrastive strategies followed by a Swin Transformer which enforced the capture both global and local anatomical representations. Another type of hybrid Transformers-CNN encoders deal with multi-branch fusion schemes like TransFuse [131], HybridCTrM [132] and CrossTeaching [133]. Commonly, the two parallel branches, one for CNN and the other for Transformers, are fused to benefit from the two learning paradigms. The CNN branch provides the ability to focus on local information while Transformers learn long-range voxel dependencies.

3) Full Transformers network: Full Transformers architectures are built in an end-to-end Transformer-based fashion [134, 135]. Cao et al. proposed in [136] a Swin 2D U-Net network which includes a patch expanding layer to up-sample the feature maps of the decoder. This architecture showed superior performance in capturing fine details compared to decoders based on bi-linear up-sampling. Lin et al. went a step further in [137] by first adopting a dual-scale encoder based on 2D Swin-Transformer to extract both coarse and fine-grained feature representations at different scales. They also included an interactive fusion module to effectively establish global dependencies between features of different scales through the self-attention mechanism. To better leverage multi-scale feature hierarchies, Huang et al. proposed in [138] a 2D hierarchical encoder-decoder architecture whose main contribution was the inclusion of an enhanced Transformer context bridge to capture the correlation and local context of multi-scale features generated by the hierarchical Transformer encoder. Peiris et al. developed in [139] VTU-Net that works directly in 3D using Swin Transformers. The Swin decoder introduces parallel cross-attention and self-attention, which creates a bridge between queries from the decoder, keys, and values from the encoder. Such parallelization enables to preserve the full global context during the decoding process, which is key towards a robust delineation of medical images.

B. Multi-task learning

Another way to improve image segmentation consists in exploiting the ability of deep models to simultaneously deal with multiple tasks. Multi-task learning aims at taking advantage of the information shared among two or more connected or auxiliary tasks while better handling each task individually [52]. As investigated in [140], multi-task learning can be seen as a form of inductive transfer where the introduced inductive bias allows to prefer some hypotheses over others (i.e. the ones that explain more than one task), towards solutions that generalize better than their individual counterparts. One of the first multi-task learning frameworks dedicated to medical image segmentation was developed in [141] which investigated whether a single CNN can be trained to perform different segmentation tasks, i.e. in a multi-domain fashion (Sect. VI-B). The combined training procedure was therefore suited for identifying anatomical structures, tissue classes as well as imaging modalities at once. More globally, multi-task learning can leverage various and heterogeneous forms of annotations, from global images to finer-grained and pixel-level labels. Based on the finding that several segmentation, classification, regression or detection issues can be effectively solved at once using a single network, multi-tasking strategies have emerged through the design of a cascade of task-specific sub-networks [30] or the development of networks with shared encoder and task-specific decoders [142–145] to benefit from partial parameters sharing between tasks or sub-tasks.

In particular, the encoder-decoder architecture designed in [142] featured a single encoding path and multiple decoding branches for concurrent segmentation tasks. The encoding module used a generic set of parameters shared by multiple tasks whereas the decoding branches were task-specific. An auxiliary cost was also added at the end of the encoding module to predict the presence or absence of lesions. Supplementary sub-tasks including contour detection and distance map estimation were incorporated in [143] to refine coarse and discontinuous segmentation predictions from convolutional models. In [144], a single multi-task network was proposed to simultaneously address gastric tumor segmentation and lymph node classification from CT scans. An attention-based reconstruction task was integrated into the segmentation pipeline of [145] to leverage unlabeled medical images in a semi-supervised segmentation framework (Sect.VI-D). In [73], Keshwani et al. improved vessel CT segmentation not only by considering a single decoding branch dedicated to vessel segmentation but also by involving two additional decoders: a centerness decoder whose task was formulated as a regression problem and a topology distance decoder aiming at enhancing the vessel connectivity which is key regarding clinical needs.

By enabling fruitful cooperation between related tasks, these multi-task approaches have been shown to outperform traditional independent or segmentation-only models. As for deep supervision (Sect.IV-D), the relative weighting between each task's loss can be tuned by hand or automatically [146].

C. Segmentation uncertainty

In DL-based image segmentation, the forward pass is a deterministic process that maps a voxel to a unique label. This apparent determinism however fails to take into account the various sources of uncertainty that affect a neural network pre-

diction. Such awareness may be crucial to detect potential segmentation errors. There is a clear need to understand the limitations of segmentation models via the assessment of voxel-wise confidence measures, which is the purpose of uncertainty quantification applied to segmentation [147, 148]. Uncertainty modelling may also importantly be used to directly improve segmentation performance, as in [149] where the distinction between quality control and model improvement techniques is highlighted. This is typically done by averaging out voxel-level errors using multiple stochastic forward passes in models that can sample across the uncertainty space [31, 147].

According to Bayesian terminology, uncertainty can be divided into epistemic and aleatoric uncertainty [146, 150]. Epistemic (or model) uncertainty relates to the lack of accuracy in the model parameters due to insufficient training. It is a kind of uncertainty that can be reduced by providing more training time and/or data. Aleatoric uncertainty, on the other hand, relates to the inherent uncertainty of the data itself, which can further be divided into homoscedastic uncertainty (constant for all inputs) and heteroscedastic uncertainty (variable between inputs). In image segmentation, both image and label spaces are affected by aleatoric uncertainty. Example causes of homoscedastic uncertainty in radiation imaging are physical processes such as Compton scattering or positron range. Image-space heteroscedastic uncertainty can be, for instance, due to dataset shifts in multi-center studies, while label-space heteroscedastic uncertainty may be due to heterogeneity in annotation quality [151]. Ideally, uncertainty assessment should be calibrated. Calibration means that prediction confidence c should equal its likelihood of being correct, i.e. a value of $c \in [0, 100]$ should indeed translate to a model being accurate $c\%$ of the time over multiple instances, which is an open research subject in medical imaging [147, 148, 152]. Several techniques may be employed to produce voxel-level confidence maps for both epistemic and aleatoric uncertainty quantification of segmentation predictions. The most popular epistemic uncertainty measurement is Monte-Carlo (MC) dropout, also known as test-time dropout (TTD) [150], where many (e.g. from 10 to 50) stochastic dropout forward passes of a model equipped with dropout weights are performed during training [153–155]. The dissimilarity in the predictions, assessed through variance, mutual information or entropy [154] can then be assimilated to a voxel-wise epistemic uncertainty map. An obvious drawback of MC dropout is the requirement for dropout during training. Dropout may indeed be detrimental to segmentation performance, and a number of state-of-the-art segmentation solutions including nnU-Net [28] do not include dropout. Alternatives to MC dropout for epistemic uncertainty quantification include performing forward passes at various training checkpoints of the optimization [156], following the empirical observation that less certain predictions are less stably predicted along training. A more computationally demanding method is deep ensembling, whereby independently trained networks are averaged together to get uncertainty maps [157]. Albeit more demanding, deep ensembling is a common practice due to its consistency in improving segmentation results [28]. Thus, uncertainty maps may be derived freely as a by-product of this main objective.

Data or aleatoric uncertainty, on the other hand, can be assessed through test-time data augmentation (TTA), in which multiple forward passes are performed to inputs altered through basic data augmentations (e.g. flips, rotations, scaling) [31]. The resulting outputs are then aggregated with similar methods to TTD (i.e. averaging, entropy). TTA is easier to implement than TTD as it does not require any modification to the network architecture and can readily be achieved through off-the-shelf data augmentation and segmentation frameworks. Qualitative results seem to suggest that aleatoric uncertainty estimates provide more expressive qualitative maps for medical image segmentation uncertainty assessment [31].

Regarding the improvement of performance through uncertainty sampling, using epistemic uncertainty modelling with TTD or MC dropout generally yields moderate but consistent improvement of segmentation results [31, 156]. For instance, epistemic uncertainty-aware networks achieved state-of-the-art performance on the medical image segmentation decathlon challenge [153, 158]. On the other hand, TTA seems to be more effective than TTD for improving medical image segmentation results, with performance enhanced by up to several Dice points [31]. TTA is therefore a generally recommended step if inference cost is a secondary concern. The question as to what is the optimal way to pool TTA predictions and what augmentation to select is an open research subject [159]. Applications of deep uncertainty modelling to PET-CT are less popular than in MR modality, with few related works in segmentation [65, 160]. Sudarshan et al. leverage physics-based heteroscedastic uncertainty modelling for low-dose PET-MR image denoising [161]. This relative lack is arguably due to the novelty of the topic in medical imaging. Uncertainty quantification being an emerging trend, more contributions are expected in the future, especially in radiation imaging.

D. Contrastive learning

Whatever the image analysis task involving representation learning, extracting robust features means reaching distinct clusters reflecting the different classes involved. In this direction, contrastive learning tends to enforce the model to learn an efficient and disentangled feature representation by comparing the input image with comparing images (i.e. anchors). The comparison is performed between positive pairs of similar inputs (e.g. generated through data augmentation) and negative pairs of dissimilar inputs (e.g. other image samples used for training). The original contrastive loss was initially defined for classification purposes in computer vision [162] and its adoption in the medical image processing community has been relatively late [163]. The underlying idea was to pull together data points from the same class while pushing apart negative samples in embedded space [162], thus imposing intra-class cohesion and inter-class separation.

1) Global contrastive learning: Most existing contrastive learning methods deal with a global contrastive loss and target image classification (Fig.8*a*). For image segmentation, a global contrastive loss can still be used by projecting the data through the encoder path to the latent space [32, 164]. Let us describe

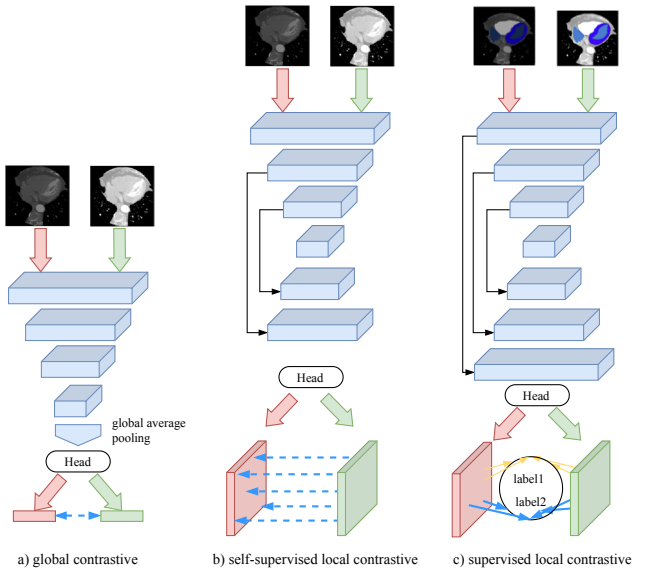

Fig. 8. Self-supervised global (a), self-supervised (b) and supervised (c) local contrastive learning [164]. Refer to Sect.V-D for further details.

how a global contrastive loss can be defined in this context. We consider an input batch $\mathcal{B} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_b\}$ where each $\boldsymbol{x}_i$ corresponds to a 3D medical image or one of its 2D slice. Using data augmentation, one can transform twice each $\boldsymbol{x}_i$ to form a pair of augmented images. These new images form a set of augmented images $\mathcal{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{2\times b}\}$. For a given image $\boldsymbol{x}_i$, $\mathcal{A}$ contains two related augmented versions, referred to $\boldsymbol{a}_j$ and $\boldsymbol{a}_k$. A global contrastive loss can therefore be used:

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{A}|} \sum_{j\in\{1,\ldots,2\times b\}} \log \frac{\exp(\boldsymbol{z}_j.\boldsymbol{z}_k/\tau)}{\sum_{l\neq j} \exp(\boldsymbol{z}_j.\boldsymbol{z}_l/\tau)} \tag{18}$$

where $\boldsymbol{z}_j$ and $\boldsymbol{z}_k$ are the normalized features obtained after applying a header function $h(.)$ (e.g. multi-layer perceptron) at the output of the encoder E such that $\boldsymbol{z}_j = |h(E(\boldsymbol{a}_j))|$. τ is a temperature scaling parameter. As opposed to standard approaches that operate on image classes, a dataset label information was leveraged in [35] to enhance intra-domain similarity and impose inter-domain margins, in a multi-task multi-domain segmentation scenario.

2) Local contrastive learning: More suited for medical image segmentation, a local contrastive learning approach can be followed by designing a local version of the contrastive loss (Eq.18) able to learn distinctive representations of local regions instead of relying on global representations. Only teaching the encoder to extract image-level disentangled features may not be sufficient since segmentation requires a class prediction for all voxels [164]. Two main techniques can be followed to train the decoder at extracting distinctive local representations through contrastive learning: self-supervised (Fig.8*b*) and supervised (Fig.8*c*) local contrastive learning.

For the first category, Chaitanya et al. introduced in [32] a local extension of the contrastive loss that is helpful for per-pixel segmentation as it learns distinctive representations of local regions. Thus, it compares the local features of the image to be equivalent underneath various transformations and also focuses on the dissimilarity with other regions from the same image. Validated on three MR datasets, this method lead to a substantial increment in delineation accuracy. Since generating pairs of data for the use of contrastive learning is challenging in medical image segmentation due to the potential presence of similar tissue or anatomical structure across the dataset, Zeng et al. explored in [165] a novel method called positional contrastive learning. The method dealed with generating contrastive data pairs based on the position of a slice within 3D volumes. Based on the slice distance, more closest slices were referred to as positive pairs, and far slices were considered negative. This enabled the reduction of false-negative image pairs and improved the segmentation results against state-of-the-art. Self-supervised local contrastive learning is only one of the strategies followed in self-supervised learning whose perimeter is further explained in Sect.VI-C.

Conversely, a supervised local contrastive loss (Fig.8*c*) that leverages limited pixel-wise annotation to force pixels with the same label to gather around in the embedding space was proposed in [164]. Some papers combine global and self-supervised [32] or supervised [164] local contrastive learning.

Contrastive learning was also adopted in a federated learning context (Sect.VI-E). In [38], Wu et al. explored contrastive learning for volumetric medical image segmentation in the presence of limited labeled data. Following a similar approach to the above-discussed ones, clients first learn a shared encoder on unlabeled data from various sites. Then, a network is fine-tuned on a labeled dataset. The mixed contrastive data are supplied to each medical location, enabling the use of data variousness for contrastive learning. This enables performing global structural matching to learn an encoder with suitable representations among clients.

E. Knowledge distillation

Compared to the computationally expensive scenario consisting of training many different models on the same data and then averaging their predictions, compressing the knowledge into a single model through knowledge distillation is much easier to deploy [166]. The knowledge distillation (KD) mechanism tends to distill (i.e. transfer) information from a well-trained cumbersome teacher network to a lightweight and compact student network with the final goal of improving the performance of this student model. In a standard KD setting, the teacher model generates soft predictions which are used to supervise the student model by calculating the difference of their final layer with some measurement functions such as cross-entropy or Kullback-Leibler divergence (KL). The soft predictions are obtained after the last convolution layer by mean of a parameterized softmax function following:

$$\boldsymbol{q}_c = \frac{\exp(\frac{\boldsymbol{z}_c}{\tau})}{\exp\left(\sum_{c\in\mathscr{C}} \frac{\boldsymbol{z}_c}{\tau}\right)} \tag{19}$$

where the logits $\boldsymbol{z}_c$ associated to each class c are converted into probabilities $\boldsymbol{q}_c$. τ is a parameter called temperature which

controls the softness of the output probabilities. Since tuning τ is a difficult process and considering τ as a constant can lead to sub-optimal results, feature normalized knowledge distillation (FNKD) was proposed in [167] to exploit reliable soft predictions irregardless of the feature scale by rather considering the L2 norm instead of a scalar value:

$$\boldsymbol{q}_c = \frac{\exp\left(\frac{\boldsymbol{z}_c}{\|\boldsymbol{z}\|}\right)}{\exp\left(\sum_{c\in\mathscr{C}} \frac{\boldsymbol{z}_c}{\|\boldsymbol{z}\|}\right)} \tag{20}$$

with $\|.\|$ corresponding to the L2 norm. Instead of focusing on the final layer only from both teacher and student models, guiding the compact network to mimic intermediate features from the teacher network was adopted in [168] by means of attention maps. Initially proposed in the context of image classification, KD mechanisms were rapidly extended for semantic segmentation purposes in computer vision [169, 170] and medical image analysis [33] fields. Thus, a prediction map distillation module was used in [33] to enable the student network to learn predictive capability from the output segmentation maps provided by the teacher. Let ϕ^s be the student model, $\boldsymbol{q}_n^s$ and $\boldsymbol{q}_n^t$ the soft probability maps respectively from the student and teacher networks. The loss function for training the student ϕ^s with standard KD can be defined as:

$$\mathcal{L}_{\phi^s} = \frac{1}{N}\sum_{n=1}^{N} \lambda \ \texttt{KL}(\boldsymbol{q}_n^s||\boldsymbol{q}_n^t) + (1-\lambda)\ \ell_{CE}(\phi^s(\boldsymbol{x}_n), \boldsymbol{y}_n) \tag{21}$$

with $\lambda \in [0, 1]$ a scalar value adjusting the contributions of both terms. Note that cross-entropy can easily replace KL for the knowledge distillation sub-loss. The standard supervised loss (cross-entropy in Eq.21) can be any dedicated loss function (e.g. DiceCE), as mentioned in Sect.II-A. Such loss function making the student mimicking the ability of the teacher to generate soft prediction maps is however not enough to really boost the performance of the student since only pixel-level information is considered. In this direction, more context and class-related information are needed. As for classification [168], constraints on intermediate multi-scale features arising from both teacher and student were therefore integrated [33] with importance maps distillation modules able to encode feature maps into a transformable form to deal with the diversity of feature sizes between teacher and student models. Other constraints were proposed in the specific context of knowledge distillation such as boundary-guided [171], region affinity [33], class-similarity [172], anatomical knowledge [173] or holistic distillation [174] in order to align high-order relations between what both teacher and student generated.

To go further, a novel KD based framework called multiple teachers single student (MTSS) was developed as a new privacy multi-organ segmentation setting learning from multiple pre-trained single-organ segmentation models [175]. Formulated into a special unsupervised ensemble distillation problem, multiple single-organ models served as teachers from different specialties and collaboratively teach one general student, i.e. the multi-organ segmentation model. Further, the integration of a co-training strategy and weight-averaged models unified multi-organ segmentation from few-organ datasets [176]. Self-distilling a Transformer-based U-Net by simultaneously learning global semantic information and local spatial-detailed features was also investigated in [177].

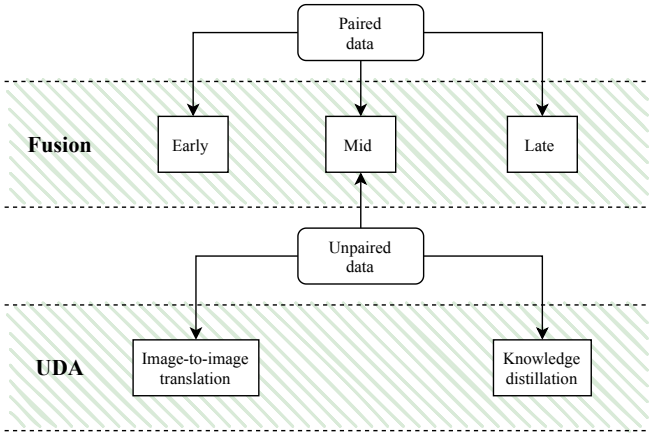

Fig. 9. Categorization of cross-modal segmentation frameworks. UDA stands for unsupervised domain adaptation. Refer to Sect.VI-A for further details.

VI. Emerging applications

A. Cross-modality segmentation

Most of the approaches presented in the literature do not consider the multi-modal nature of medical imaging data, leaving aside potentially valuable cross-modal information unused. However, exploiting complementary and redundancy information across modalities can possibly improve overall segmentation performance, making better use of the scarcity of annotated medical imaging data. The research field of multi-modal image segmentation brings different technical challenges and open questions to solve [178], including:

- Is the data available for training pairwise-aligned or comes from different patients?
- How to fuse different modalities to simultaneously reduce the heterogeneity gap and enable the transfer knowledge?
- How to map data from one modality to another?

Considering these questions, we introduce in what follows the multi-modal works depending on the type of data available during training (paired, unpaired), the followed fusion strategy (early, mid or late) [34] as well as the adopted translation approach. Fig.9 depicts a generic categorization of the cross-modal segmentation framework whereas Fig.10 illustrates a cross-modal pipeline when managing paired data.

1) Paired data and early fusion strategy: When multi-modal paired data is available, one may carry out any of the three fusion strategies. The most straightforward is the early (also known as input-level) fusion strategy (Fig.10*a*) which integrates at the input level of the deep network the different m-modalities. Therefore, the final segmentation is defined as $\mathbf{y} = \phi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ where ϕ represents the segmentation

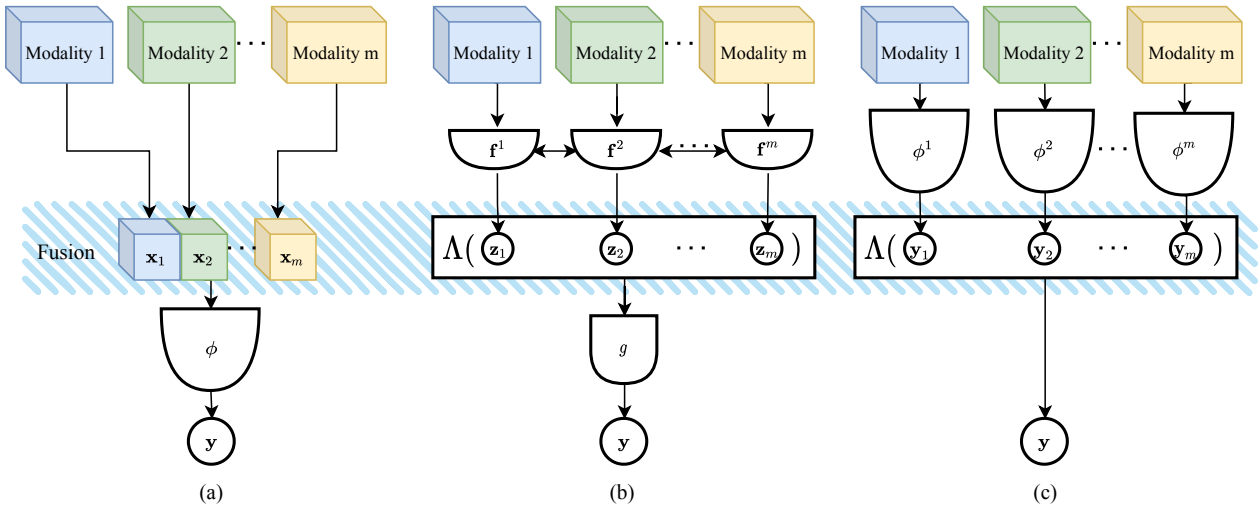

Fig. 10. Pipeline of cross-modal frameworks for paired data. (a) Early or input-level fusion, (b) mid or layer-level fusion, (c) late or decision-level fusion.

network. Early fusion has the advantage of its simplicity, allowing more complex segmentation strategies such as multi-task, multi-view, multi-scale or GAN-based approaches. One of the first works devoted to solving multi-modal image segmentation using CNN can be found in [179]. They explored two-pathway cascaded architectures using different receptive field sizes to capture both local and global context information. Qin et al. proposed in [180] an adaptive convolutional layer named autofocus to effectively change the size of the receptive field to perform multi-modal brain tumor segmentation. The autofocus layer captured the multi-scale information by parallelizing multiple convolutional layers with different dilatation rates that are later merged using a weighted soft-attention mechanism to choose the optimal scales.

The above-mentioned methods and others [181, 182] did not make dense predictions and are therefore slow in the inference stage. To promote efficiency, encoder-decoder architecture derived from U-Net [14] has been widely adopted. For instance, Shapey et al. used in [183] a 2.5D U-Net to segment the vestibular schwannoma in contrast-enhanced T1-weighted and high-resolution T2-weighted MR imaging. Spatial attention modules were added to each level of the decoder to deal with small target regions, giving more attention to them and penalizing voxels belonging to the background. In [184], the modalities were fused as multi-channel inputs and passed through an adversarial network (Sect.IV-A). The generator is a 3D residual U-Net that performs the segmentation while the discriminator distinguishes between generated segmentation and ground truth masks. An extra constraint was added via active contour modeling by measuring the dissimilarity between ground truth and prediction contours. To handle the class imbalance problem, Zhou et al. carried out in [185] a coarse-to-fine segmentation inspired on model cascades for brain tumor segmentation. The main difference with previous works lies in applying only a one-pass multi-task network (OM-Net) that performs three tasks that are gradually introduced in an order of increasing difficulty based on curriculum learning. The first task learns to differentiate between tumors and normal tissue until the loss curve tends to flatten. The second task is then added and split the complete tumor into intra-tumoral classes. This task continues until its loss curve displays a flattening trend. Lastly, the third task is introduced and trained simultaneously with the previous ones to precisely segment the enhancing tumor. In this way, the model parameters and the training data are transferred from an easier to a more difficult task. Unfortunately, early fusion makes it hard to discover highly non-linear relationships between the low-level features from different modalities, especially when the modalities have significantly different statistical properties.

2) Paired data and mid-fusion strategy: Mid or layer-level fusion separately processes the multi-modal data in different paths (Fig.10*b*). For each modality, m, the input $\mathbf{x}_m$ is encoded in each branch $\mathbf{f}^m$ to learn the modality-specific representation $\mathbf{z}_m$. Then, each representation $\mathbf{z}_m$ is mapped into a common latent space via a fusion operation Λ and use this as input of the decoding transformations $g(\Lambda(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m))$. The main goal of this strategy is to learn an optimal joint representation that emphasizes the most informative features across modalities. In mid-fusion, we can distinguish two types of multi-pathway network architectures based on the following fusion strategy: single-layer or multi-layer fusion.

Multi-modal segmentation networks based on single-layer fusion generally employ encoder-decoder architectures where each modality has its own encoder, with no interactions between them, and a single decoder. They mainly differ on the conducted fusion operation and are typically carried out via concatenation, addition, averaging or convolution. For instance, Havaei et al. used in [186] modality-specific convolutional layers to later compute for each feature map the first and second moments. Then, the moments were concatenated and processed by further convolutional layers, yielding the final segmentation. For their part, Tseng et al. took in [187] the encoded representation from each modality and performed a cross-modal convolution to combine the spatial information of each feature map, modeling the correlations among them. In-

spired by the success of the attention mechanism (Sect.IV-E), recent fusion strategies incorporated spatial and channel-wise attention to learn more informative features among modalities [188, 189]. To name a few, Zhou et al. proposed in [188] a three-stage segmentation network. In the first stage, a 3D U-Net architecture got rough mask predictions. Then, binarization and erosion operations were used to obtain the context constraints for the following stage. The second stage consisted of a multi-encoder-based framework where each encoder produces a modality-specific latent representation that is further fused with the assistance of attention mechanisms. This process was repeated for each structure to be segmented. In the third stage, a two-encoder-based 3D U-Net segmentation network was applied to combine and refine the three single prediction results. A correlation block to discover the latent correlation between modalities was introduced in [189], followed by a dual attention block that consists of a modality attention module and a spatial attention module. In this way, the network is encouraged to learn the most correlated features across modalities and more useful spatial information to boost segmentation results. Despite the great results of single-layer fusion schemes, there is no complete freedom to learn within and in-between modalities due to its single level of abstraction.

Regarding multi-layer fusion, it extends the idea of residual learning in multi-modal frameworks allowing skip-connections that by-pass spatial features between modalities [190–192]. Therefore, low-level and high-level features are fused at different levels of abstraction, increasing the learning capabilities of the network to capture complex cues across modalities. Li et al. proposed in [190] four dilated Inception blocks consisting of three dilated convolutional layers for each modality. In this way, the receptive field of the network was expanded without losing resolution, while multi-scale features were also learned. In order to obtain the final segmentation, the features at different levels were concatenated and up-sampled. On the other hand, Dolz et al. proposed in [191] HyperDenseNet, a 3D model where each modality has its own path. Dense connections not only occur between the layers within the same stream but also across modalities. Thus, the network can learn more powerful feature representations at all levels of abstraction. To encode more rich contextual information across modalities, Zhang et al. developed in [192] a cross-modal self-attention distillation network. The model extracted attention maps of intermediate layers to further perform layer-wise attention distillation among modalities. Significant spatial information can be distilled from an attention map of one modality and then used to ease attention learning of the other modalities. Fusing multi-modal contextual information at multi-layer stages represents the current trend. Moreover, semantic guiding across modalities by attention mechanisms can be used to bridge early feature extraction and late decision-making.

3) Paired data and late fusion: Similar to mid-fusion, late fusion separately processes multi-modal data $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ with the difference that the segmentation branches $(\phi_1, \phi_2, \ldots, \phi_m)$ are integrated at the decision level. More precisely, during the decoding stage, all feature maps computed by the branches are mapped into a common feature space via fusion operations, (e.g. concatenation, averaging, weighted voting), followed by series of convolutional layers [193]. The final output of late fusion can be formulated as $y = \Lambda(\phi_1(\mathbf{x}_1), \phi_2(\mathbf{x}_2), \ldots, \phi_m(\mathbf{x}_m))$ where Λ is the fusion operation. Thus, common features learned by the transformation network are considered as a further refinement of decoding and prediction. Some conventional layer-level methods as [194] are thus categorized into late fusion strategy. Many late fusion strategies have been proposed. Most of them are based on averaging or majority voting. For averaging strategy, Kamnitsas et al. trained in [72] three networks separately and then averaged the confidence of the individual networks. The final segmentation was obtained by assigning each voxel with the highest confidence. For the majority voting strategy, the final label of a voxel depends on the majority of the labels of the individual networks. The statistical properties of the different modalities are different, which makes it difficult for a single model to directly find correlations across modalities. Therefore, in a decision-level fusion scheme, the multiple segmentation networks can be trained to fully exploit multi-modal features. On the other hand, Zhang et al. proposed in [194] a modality-aware module that fused the modality-specific models at a high semantic level. Specifically, each modality was embedded by a different modality-specific FCN. Then, the outputs of FCN models were fused and passed to an attention module to generate a modality-specific attention map to adaptively measure the contributions of each modality. Moreover, they designed a mutual learning strategy to enable interactive knowledge transfer, where the modalities interact as teacher and student simultaneously. In the same line, Zhang et al. employed in [192] a transfer knowledge strategy across modalities that differs from previous works in the use of GAN. The authors applied cycleGAN [195] to capture the knowledge transition across modalities. Each generator represented a single-modality feature learning branch. Then, they were merged by extra convolution layers followed by an attention block to learn powerful fusion features. The intuition behind the use of GAN is that GAN models can learn the modality patterns of each modality and their content patterns.

Mid and late-fusion can achieve better performance because each modality is employed as an input of one network that can learn complex and complementary feature information compared to an input-level fusion network. However, they require more memory due to the use of multiple networks. Therefore, the trade-off between accuracy and execution time should be carefully considered. Despite the impressive advances reached in the field of multi-modal image segmentation, collecting large sets of paired images is often either prohibitively expensive or not possible. As a result, techniques that make use of unpaired datasets have attracted increasing attention in cross-modal segmentation.

4) Unpaired data and domain adaptation: When only unpaired datasets are available, cross-modal segmentation is commonly managed by domain adaptation (DA) techniques. Let $\mathcal{X}_m \times \mathcal{Y}_m$ represent the joint feature space and the corresponding label space for a specific modality m. A domain can

be formulated as $\mathcal{D}_m = \{\mathcal{X}_m, P(\mathbf{x}_m)\}$ where $\mathcal{X}_m$ is the feature space and $P(\mathbf{x}_m)$ the marginal probability distribution of the data $\mathbf{x}_m$. Let us assume that we have two domains (e.g. from two different modalities): a source domain $\mathcal{D}_S = \{\mathcal{X}_S, P(\mathbf{x}_S)\}$ and a target domain $\mathcal{D}_T = \{\mathcal{X}_T, P(\mathbf{x}_T)\}$. The DA problem is formulated as a sub-class of transfer learning where the label spaces in the source and target domains are the same $\mathcal{Y}_S = \mathcal{Y}_T = \mathcal{Y}$ but where the domains are different $\mathcal{D}_S \neq \mathcal{D}_T$. In cross-modal image segmentation, the feature spaces between source and target domains are identical (i.e. $\mathcal{X}_S = \mathcal{X}_T$), differing only in data distribution (i.e. $P(\mathbf{x}_S) \neq P(\mathbf{x}_T)$). Hence, the goal is to learn a segmentation function $f(\cdot)$ that performs well in both source and target domains by finding a transformation $T(\cdot)$ such that $P(T(\mathbf{x}_S)) = P(T(\mathbf{x}_T))$. The previous single-source single-target definition can be extended to multi-source single-target or single-source multi-target DA.

Three groups of DA techniques are used for cross-modal image segmentation: supervised domain adaption in which both labeled source and target data are available $\{\mathcal{X}_S, \mathcal{Y}_S, \mathcal{X}_T, \mathcal{Y}_T\}$, semi-supervised domain adaptation in which labeled source data in addition to some labeled target data are available, and unsupervised domain adaptation (UDA) in which both labeled source data and unlabeled target data are available $\{\mathcal{X}_S, \mathcal{Y}_S, \mathcal{X}_T\}$. Most related works in cross-modal image segmentation are based on single-source single-target UDA [196] using reconstruction-based methods, domain-invariant feature learning [197] and more largely GAN models [198].

The main drawbacks of existing approaches deal with their limited scalability and robustness in handling more than two domains since they rely on pairwise alignment using GAN (cycleGAN or similar [198]). Hence, different models should be built independently for every pair of image domains which demands high computational resources. Lastly, either paired or unpaired methods are application-dependent which limits their transferability between different clinical settings.

B. Multi-domain segmentation

A strong assumption in the way deep segmentation pipelines are usually designed and evaluated is that both training and test data arise from the same probability distribution. Their accuracy usually degrades when applied to new (i.e. unseen) data that differ from the training data [62]. Instead of designing pipelines specific to a given intensity domain, an emerging application consists in training a deep segmentation model over multiple intensity domains [35, 107, 199, 200]. The underlying assumption is that exploiting the redundancy between multiple intensity domains can enable the extraction of robust domain-invariant feature representations to finally achieve better performance than domain-specific (i.e. marginal) computational models. Managing various domains can partially solve the issue of dealing with the scarcity of imaging resources [199]. The improved generalization abilities of the resulting models are a further step to facilitate their integration into routine.

In practice, intensity domains can be very different in nature: multi-center, multi-scanner, multi-modal (Sect.VI-A), or multi-protocol. Reasons explaining the acquisition shift include differences in imaging systems, reconstruction settings

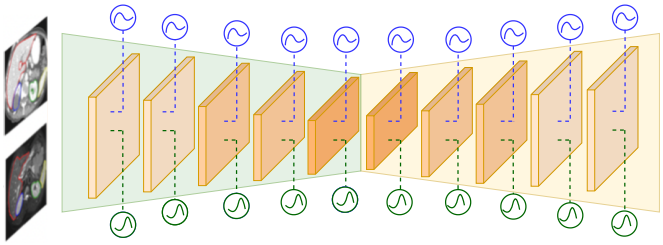

Fig. 11. Multi-domain segmentation with shared convolutional kernels and domain-specific feature normalization, as employed in [35, 200].

or acquisition protocols. It is worth mentioning that variations can even happen in a single center since both clinical practices and imaging systems may significantly evolve over time. To deal with this diversity, segmentation pipelines can integrate an adversarial network to learn domain-invariant features [62], exploit transfer learning and fine-tuning between domains [199], share a common decoder (resp. encoder) while using domain-specific encoders (resp. decoders) [107] or share their latent space only [107, 200]. In particular, the very different statistical distributions between unpaired multi-modal images making the task of learning shared representations challenging has motivated the design of a single encoder-decoder segmentation network in [200] through shared convolutional kernels but domain-specific feature normalization (Fig.11). Indeed, modality-agnostic kernels can extract expressive universal representations across domains only if the features are well-calibrated upstream. Let us consider D domains $\{d_1, d_2, \ldots, d_D\}$. Let $\boldsymbol{v}^j_{i,l,m}$ be the m^{th} feature map from the l^{th} layer produced by the i^{th} image arising from the intensity domain d_j. The calibration can be performed through domain-specific batch normalization (DSBN) [35, 200], following:

$$\texttt{DSBN}_{\alpha^j_{l,m},\beta^j_{l,m}}(\boldsymbol{v}^j_{i,l,m}) = \alpha^j_{l,m}\frac{\boldsymbol{v}^j_{i,l,m} - \mu^j_{l,m}}{\sqrt{(\sigma^j_{l,m})^2 + \epsilon}} + \beta^j_{l,m} \quad (22)$$

where $\mu^j_{l,m}$ and $\sigma^j_{l,m}$ are respectively the domain-specific mean and standard deviation computed for images from domain d_j belonging to a given batch. ϵ is a scalar value that is used to reach numerical stability. DSBN layers are therefore defined by trainable domain-specific shift and scale weights $\{\alpha^j_{l,m}, \beta^j_{l,m}\}_{l,m}$ set for each feature map of each layer.

Rather than focusing on a given anatomical target across various intensity domains, developing a single multi-task multi-domain network can enable simultaneously segmenting multiple anatomies while leveraging shared features between various domains and datasets [141]. In this direction, shared convolutional kernels and domain-specific feature normalization from [200] were combined in [35] with both contrastive (Sect.V-D) and shape regularizations (Sect.IV-C) to segment bone structures from multiple scarce pediatric datasets.

C. Self-supervised learning

The need for a large amount of annotated training data is a strong constraint given the complexity of reaching a significantly well-annotated medical imaging dataset. Generative

models were therefore used to increase the number of training samples through image synthesis [201]. As an alternative, the availability of a reduced dataset only motivated researchers to exploit the power of unlabeled images which may be easier to collect. Unsupervised learning approaches assuming that a related but unlabeled large dataset is available aim at learning transferable feature representations from unlabeled images.

In particular, self-supervised learning can consist of pre-training the model (or any of its constituents) by means of pretext tasks to finally be more able to efficiently delineate the targeted structures [36]. With the hypothesis of good generalization ability of self-learned features, Taleb et al. investigated the effectiveness of several pretext tasks for 3D medical image segmentation purposes: rotation prediction, jigsaw puzzles, relative patch location... More complex pretext tasks such as semantic inpainting revealed their effectiveness for better solving downstream segmentation tasks [202].

Leaving aside self-supervised pre-text tasks, self-supervised contrastive learning (Sect.V-D) is another manner adopted in the medical image analysis area to learn expressive feature representations from unlabeled images [203]. In this direction, Chaitanya et al. proposed in [32] a local contrastive loss able to capture local features to provide complementary information and therefore boost the segmentation accuracy.

D. Semi-supervised learning

Given the usual scarcity of many existing annotated medical dataset and apart from transfer learning [204] whose goal is to learn from related learning problems, researchers have also explored semi-supervised learning approaches to exploit the availability of unlabeled datasets. Among the existing semi-supervised strategies [37], semi-supervised consistency regularization is commonly employed through the use of a mean teacher model [205]. In [206], a novel uncertainty-aware semi-supervised learning framework was proposed and evaluated for left atrium segmentation from MR images. Teacher and student models were built in such a way that the student model learned from the teacher model by minimizing the segmentation loss on the labeled data as well as a consistency loss with respect to the targets from the teacher model on all data (i.e. labeled and unlabeled). The predicted targets from the teacher model being potentially unreliable and noisy on unlabeled data, Yu et al. designed an uncertainty-aware mean teacher framework [206], where the student model gradually learned from the meaningful and reliable targets by exploiting the uncertainty information arising from the teacher model (Sect.V-C). To better deal with noisy labels for COVID-19 pneumonia lesion segmentation, the main novelty in [207] was to propose two mechanisms: an adaptive teacher that suppresses the contribution of the student when the latter has a large training loss and an adaptive student that learns from the teacher only when the teacher outperforms the student.

As followed in [208], semi-supervised pseudo labeling is another strategy to deal with both annotated and unlabeled data. Thus, Fan et al. generated pseudo-labels by relying on a first training with 50 labeled images only. The newly pseudo-labeled examples were then included in the original labeled training dataset to re-train the model. This updated model was used to generate pseudo-labels for another batch of unlabeled images and so on. This process was repeated up to obtain efficient performance in COVID-19 lung infection CT segmentation. However, the created pseudo-labels usually do not have the same quality as ground truth labels, which may limit their potential for improvements from unlabeled data.

As a powerful alternative, one may adopt an auxiliary task on unlabeled data to facilitate performing image segmentation with limited labeled data. In this direction, Chen et al. proposed in [145] a semi-supervised image segmentation method that simultaneously optimizes both supervised segmentation and unsupervised reconstruction objectives. The reconstruction task had the particularity to exploit an attention mechanism that separated the reconstruction of image regions corresponding to different classes. Such a simple yet effective multi-task learning scheme (Sect.V-B) achieved strong improvements for brain tumor and white matter hyper-intensities segmentation.

E. Federated Learning

Collecting large medical image datasets is a difficult and time-consuming task for research needs. Accurate labeling of these images requires clinical experience and is challenging to obtain. Many imaging centers have large image datasets, but many of them are unorganized or poorly annotated in spite of their richness regarding deep model training [209, 210]. In addition, medical images are usually linked to personal health information related to the patient. Data protection to prevent sharing sensitive patient data is essential when working with multiple medical institutions, in a collaborative manner.

To solve this issue, federated learning (FL) enables distributed training of DL models without really sharing data between multiple clinical institutions. Fig.12 shows the general framework of federated learning. To work in a collaborative fashion, FL allows various clinical institutes or hospitals to work in coordination by using a central server. Each hospital keeps an individual model which focuses on the local data only. Before the training process, each institution submits a request to download the global model from a central server. The requested query is then approved by the central server and the global model weights can be downloaded. Once the training process is executed, the local client model weights are sent to the central server for updating purposes. The central server aggregates the feedback received from individual institutions and updates the global model weights based on pre-defined rules. These rules permit the model to measure the quality of the feedback obtained from the client servers.

More and more research works in medical image segmentation involve a FL scheme [211]. Recently, Xu et al. introduced in [212] a new federated cross-learning segmentation approach that handles data that are not independently and identically distributed. Unlike the conventional FL methods that combine multiple individually trained local models on a server node, the proposed method named FedCross consecutively trained the global model across multiple clients in a round-robin fashion. The authors also suggested a new federated cross-ensemble learning technique that together trains and sets up

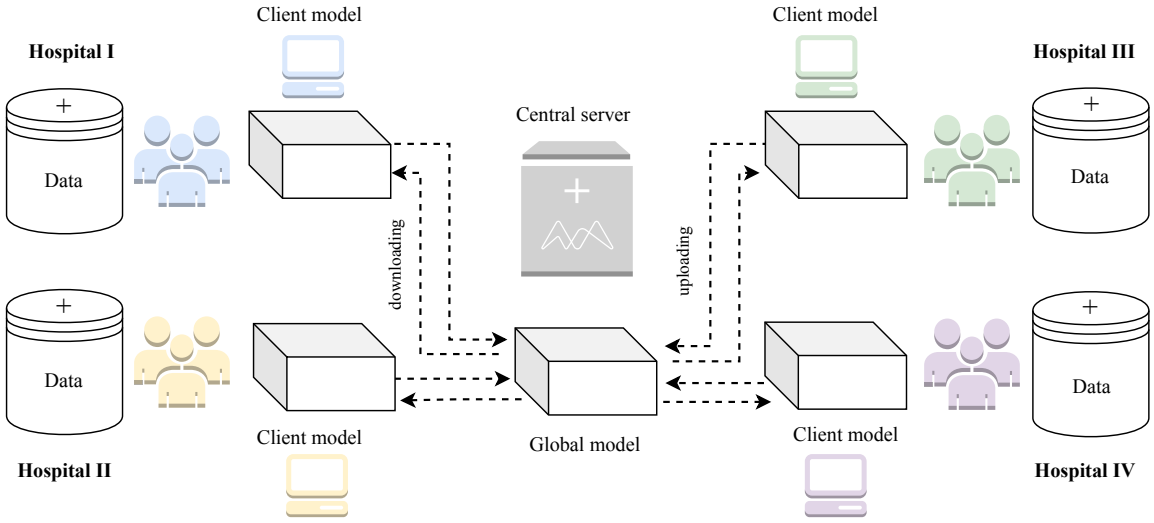

Fig. 12. General framework of federated learning. Refer to Sect.VI-E for further details.

various models. Wicaksana et al. [213] proposed FedMix, an FL strategy that employed mixed image labels specifically to segment anatomical region-of-interest from medical images. These labels incorporated substantial pixel-wise annotations, weak bounding boxes and image-wise class annotations. The authors initially created the pseudo annotations through clients and employed refinement under supervision to enrich pseudo-annotations. Later, each client contained high-quality data which was determined using active sample picking for local model updates. Based on the quantity and quality of data, the approach provided updates through dynamic aggregation techniques which allow for modification of each local client's weight. FedMix was validated on breast tumor segmentation from ultrasound and skin lesion segmentation.

Wu et al. [38] introduced a new federated contrastive learning (FCL) framework for 3D volumetric image segmentation that requires limited annotations only. The local clients first started with learning a shared encoder to spread unlabeled images. Later, annotated images were incorporated to fine-tune the model. Through feature exchange in which each client exchanges the features (i.e. low-dimensional vectors) of its local data with other clients, the approach enables better local contrastive learning while avoiding raw data sharing. A global structural matching technique was developed to learn the structural similarity of encoded features with suitable representations to be shared with other remote clients.

More globally, FL has shown potential for improving the accuracy of medical image segmentation while protecting the privacy of individual patient data. By offering scalability, flexible training scheduling and large training datasets via multi-site collaborations [210], FL combines essential conditions to consider increasing its deployment in various clinical settings.

F. Active learning

Active learning (AL) is a learning technique that involves training a model on a small, initial set of labeled data and then iteratively selecting new data to be labeled and added to the training set (Fig.13). Thus, it assists annotators in the annotation process to select the most useful samples to train a DL-based model. This is particularly useful in medical image segmentation, as manually labeling large amounts of medical images is time-consuming and costly [52]. By using AL, the model can learn to accurately segment images with less human inputs, making the process more efficient and cost-effective. Additionally, AL allows the model to focus on the most difficult and important samples, resulting in improved delineation performance. Nevertheless, choosing the best data enabling the improvement of the model learning capability remains challenging. In particular, there are multiple methods to measure informativeness which mainly involves uncertainty and representativeness criteria [39]. DL-based segmentation methods are able to measure uncertainty (Sect.V-C) to some extent. Computing for each voxel the sum of the lowest class probability is one of the simplest manners. If the prediction is uncertain, an increased number of annotated data is required to exploit richer feature information. On the contrary, representativeness deals more with choosing the samples from distinct regions of the data distribution such that the variability among the whole dataset is taken into account. In this context, a good balance between exploration and exploitation among the distribution is highly desirable.

To name a few, a cascaded 3D U-Net with CNN-correction label curation was employed in [214] for kidney segmentation from abdominal CT images in order to save the annotation efforts and improve the segmentation outcomes. AL was concluded to be able to reduce labeling efforts through CNN-corrected segmentation and increase training efficiency by iterative learning with limited data. Shen et al. presented in [215] an AL approach able to alleviate the image annotation issues towards brain tumor segmentation. The authors combined both uncertainty and representativeness information to ensure that AL selects enough informative and diverse data. Contrary to existing studies based on uncertainty or representativeness estimated at the scale of a single image, Yan et al. scored in [216] dual-view mammograms according to their prediction consistency, towards better breast mass

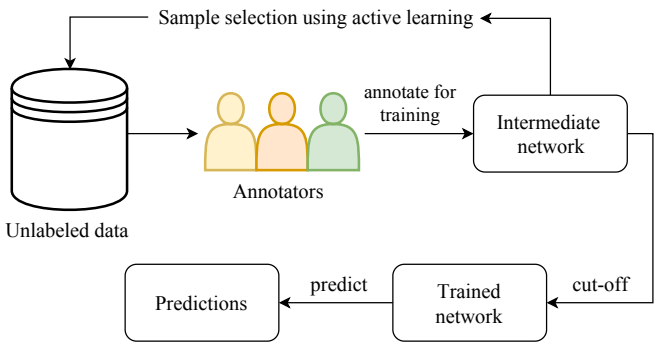

Fig. 13. General framework of active learning. See Sect.VI-F for more details.

delineation. A future possible extension is to integrate multiple single and dual-view criteria to reach a unified AL system. More globally, combining the strengths of AL and human-in-the-loop computing [39] into end-to-end systems should play an increasingly important role in the upcoming years.

G. Lightweight networks

Despite high performance, applying CNN or Transformers-based networks for 3D medical image segmentation is computationally expensive. For instance, 3D convolution layers made of many filters involve a large set of parameters to train as well as a huge amount of floating point operations (FLOP) [40]. Medical images themselves obviously require a large computational storage cost. However, the device on which a given 3D model is deployed may have limited computational power, making the deployment process hard in clinical routine. In this context, the development of lightweight deep segmentation models with smaller model sizes, lower computational cost and inference time has recently attracted increasing attention. First attempts investigated depth-separable convolutions [66] consisting in replacing 3D convolution kernel $3 \times 3 \times 3$ with $1 \times 3 \times 3$ intra-slice and $3 \times 1 \times 1$ inter-slice convolutions, combinations between point-wise, group-wise and dilated convolutions [63] or the reduction of channels [217]. However, obtaining the same performance as computationally expensive heavy-weight models is tedious and resolving the trade-off between trainable parameters and performance remains a challenge. Towards a better trade-off, Zhao et al. managed in [218] the limited feature learning ability of spatio-temporal separable convolutions via an attention-based feature calibration mechanism providing more contextual information with a larger receptive field.

The knowledge distillation mechanism fully described in Sect.V-E which tends to distill information from a well-trained cumbersome teacher network to a lightweight and compact student network [33] could be also seen as a powerful alternative towards lightweight segmentation networks.

VII. Discussion

Deep learning (DL) has proven to be a powerful tool for medical image segmentation. Its ability to automatically learn complex and hierarchical representations from data enabled to achieve a high level of robustness in segmentation tasks tackling various diseases, anatomies and imaging modalities. The availability of large datasets and open-source DL frameworks has facilitated the development and deployment of DL-based segmentation algorithms, making them more accessible to researchers and clinicians. The success of DL in medical image segmentation has been thus demonstrated in a variety of studies including whole or sub-structure organ segmentation, abnormality extraction or vascular system delineation. Especially, lesion segmentation is increasingly benefiting from the availability of combined anatomical and nuclear imaging. More globally, medical image segmentation with DL stars to have a concrete impact and to play a key role in diagnosis, surgery or therapeutic planning, follow-up, prognostic, dosimetric or radiomics applications at large.

Since the introduction of U-Net and U-shaped convolutional encoder-decoder derivatives with data augmentation and encoder pre-training, various developments and methodological breakthrough have emerged in the medical image analysis community. Among current trends, the relevancy of conditional generative adversarial networks, cascaded networks, deep supervision and attention mechanisms have been proven to enable the improvement of segmentation accuracy for both large and small anatomical or pathological structures. Regularization techniques embedding prior knowledge such as shape, topological or adjacency constraints tend to be democratized towards greater robustness and generalizability of deep segmentation models. Additionally to novel architecture designs and learning paradigms, a significantly strong focus has been recently devoted to both data management and optimization processes through the development of unified frameworks such as nnU-Net whose popularity is steadily growing. Given the substantial progress made in recent years, considering a standard U-Net as the sole baseline no longer seems relevant.

Despite these successes, there are still challenges to the use of DL for medical image segmentation. These challenges include the need for large amounts of labeled data for training, the sensitivity of DL models to noise, non-uniform contrast and artifacts in medical images, the needed incorporation of local and global context to benefit from both short- and long-range spatial dependencies, the management of small structures and weak boundaries, the robustness to inter-subject variability and various multi-center, multi-scanner, multi-modal or multi-protocol intensity domains as well as the risks for biased or non-generalizable results.

Given these challenges, the use of DL has shown great promise in line with the emergence of vision Transformers whose ability to model long-range dependencies from 3D medical images appears better than standard convolutional only architecutres. Either hybrid when used in conjunction with convolutional layers or purely Transformers-based, these approaches are still at an early stage. More works in this direction are expected, especially in the context of low-data regimes and cross-modal analysis. The great potential of multi-task learning, constrastive learning and knowledge distillation which are likely to emerge can be also emphasize as powerful trends to follow. Multi-task learning enables to share information and exploit fruitful cooperation between connected or auxiliary tasks while contrastive learning tends

to strengthen the extraction of distinctive and disentangled representations. Against computationally expensive scenarios, knowledge distillation techniques have proven great skills to distill information from a cumbersome teacher network to a lightweight student network. Uncertainty modelling is another important path to study as it may improve the learning process and provide clinicians with locally-estimated confidence information. Further research and development beyond the application of off-the-shelf DL solutions are needed to address the above-mentioned challenges and enable a wider adoption of image segmentation with DL into clinical routine.

To bridge the gap between DL paradigms and clinical needs, recent investigations have struggle with novel and concrete emerging applications. Among these applications, cross-modality segmentation has gained in popularity in order to fully exploit both complementary and redundancy across modalities when managing paired or unpaired multi-modal datasets. More globally, a special attention has been paid in recent years to multi-domain segmentation strategies which are far more relevant than focusing on multiple intensity domains separately. In this direction, multi-task and multi-domain techniques with multiple anatomies as targets should deserve further investigation in the near future. Given the complexity of collecting and annotating a large amount of medical images, self-supervised, semi-supervised and active learning are subfields of clear progress. However, more research efforts are needed to maximize or avoid the time-consuming and costly manual efforts made by clinical experts. Since medical data is often sensitive and subject to strict regulations on sharing, federated learning now offers the possibility for multiple hospitals and research institutions to collaborate by training a shared model on their own local data while keeping the data private and secure. Thus, federated learning enables the use of larger and more diverse datasets, resulting in improved segmentation performance. Finally, the development of lightweight models with few memory and computational resource requirements will for sure be beneficial to ease the deployment of DL-based solutions on computationally-limited platforms.

Overall, the potential for bias in DL approaches is a common concern across medical image analysis tasks including segmentation. In this context, encouraging the collection of large and diverse datasets through collective work with various experts is highly recommended. The design of novel evaluation metrics reflecting the clinical applicability is also an area for improvement. Finally, demonstrating a better reproducibility when designing DL pipelines could increase the trust and confidence of researchers and clinicians and make them more suitable for large-scale clinical applications.

ACKNOWLEDGMENTS

This work benefited from state aid managed by the National Research Agency under the Future Investment Program bearing the reference ANR-17-RHUS-0005 (FollowKnee project). This work was also partially funded by France Life Imaging (grant ANR-11-INBS-0006). All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this paper.

REFERENCES

[1] A. Saporta, T.-H. Vu, M. Cord, and P. Pérez, "Multi-target adversarial frameworks for domain adaptation in semantic segmentation," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9072–9081.

[2] B. E. Nelms, W. A. Tomé, G. Robinson, and J. Wheeler, "Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer," *International Journal of Radiation Oncology, Biology, Physics*, vol. 82, no. 1, pp. 368–378, 2012.

[3] X. Zhuang, K. S. Rhode, R. S. Razavi, D. J. Hawkes, and S. Ourselin, "A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI," *IEEE Transactions on Medical Imaging*, vol. 29, no. 9, pp. 1612–1625, 2010.

[4] N. Decaux, P.-H. Conze, J. Ropars, X. He, F. T. Sheehan, C. Pons, D. B. Salem, S. Brochard, and F. Rousseau, "Semi-automatic muscle segmentation in MR images using deep registration-based label propagation," *Pattern Recognition*, p. 109529, 2023.

[5] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Medical Image Analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[6] S. Kim, D. Lee, S. Park, K.-S. Oh, S. W. Chung, and Y. Kim, "Automatic segmentation of supraspinatus from MRI by internal shape fitting and autocorrection," *Computer Methods and Programs in Biomedicine*, vol. 140, pp. 165–174, 2017.

[7] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Medical Image Analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[8] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 897–908, 1999.

[9] V. Barra and J.-Y. Boire, "Segmentation of fat and muscle from MR images of the thigh by a possibilistic clustering algorithm," *Computer Methods and Programs in Biomedicine*, vol. 68, no. 3, pp. 185–193, 2002.

[10] S. Purushwalkam, B. Li, Q. Meng, and J. McPhee, "Automatic segmentation of adipose tissue from thigh magnetic resonance images," in *International Conference Image Analysis and Recognition*, 2013, pp. 451–458.

[11] S. Zhou, J. Wang, S. Zhang, Y. Liang, and Y. Gong, "Active contour model based on local and global intensity information for medical image segmentation," *Neurocomputing*, vol. 186, pp. 107–118, 2016.

[12] P.-H. Conze, V. Noblet, F. Rousseau, F. Heitz, V. de Blasi, R. Memeo, and P. Pessaux, "Scale-adaptive supervoxel-based random forests for liver tumor segmentation in dynamic contrast-enhanced CT scans," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 2, pp. 223–233, 2017.

[13] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[15] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873*, 2022.

[16] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[17] L. Liu, J. M. Wolterink, C. Brune, and R. N. Veldhuis, "Anatomy-aided deep learning for medical image segmentation: a review," *Physics in Medicine & Biology*, vol. 66, no. 11, 2021.

[18] H. Kaur, N. Kaur, and N. Neeru, "Evolution of multiorgan segmentation techniques from traditional to deep learning in abdominal CT images: A systematic review," *Displays*, 2022.

[19] S. Niyas, S. Pawan, M. A. Kumar, and J. Rajan, "Medical image segmentation with 3D convolutional neural networks: A survey," *Neurocomputing*, vol. 493, pp. 397–413, 2022.

[20] 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.

[21] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical*

Image Analysis, 2022.

[22] V. K. Singh, H. A. Rashwan, S. Romani, F. Akram, N. Pandey, M. M. K. Sarker, A. Saleh, M. Arenas, M. Arquez, D. Puig *et al.*, "Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network," *Expert Systems with Applications*, vol. 139, p. 112855, 2020.

[23] H. R. Roth, C. Shen, H. Oda, T. Sugino, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 417–425.

[24] A. Boutillon, B. Borotikar, V. Burdin, and P.-H. Conze, "Multi-structure bone segmentation in pediatric mr images with combined regularization from shape priors and adversarial network," *Artificial Intelligence in Medicine*, vol. 132, p. 102364, 2022.

[25] H. Dou, D. Karimi, C. K. Rollins, C. M. Ortinau, L. Vasung, C. Velasco-Annis, A. Ouaalam, X. Yang, D. Ni, and A. Gholipour, "A deep attentive convolutional neural network for automatic cortical plate segmentation in fetal MRI," *IEEE Transactions on Medical Imaging*, vol. 40, no. 4, pp. 1123–1133, 2020.

[26] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[27] J. J. Cerrolaza, M. L. Picazo, L. Humbert, Y. Sato, D. Rueckert, M. Á. G. Ballester, and M. G. Linguraru, "Computational anatomy for multi-organ analysis in medical imaging: A review," *Medical Image Analysis*, vol. 56, pp. 44–67, 2019.

[28] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[29] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[30] L. Song, J. Lin, Z. J. Wang, and H. Wang, "An end-to-end multi-task deep learning framework for skin lesion analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2912–2921, 2020.

[31] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[32] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 546–12 558, 2020.

[33] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3820–3831, 2021.

[34] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, p. 100004, 2019.

[35] A. Boutillon, P.-H. Conze, C. Pons, V. Burdin, and B. Borotikar, "Generalizable multi-task, multi-domain deep segmentation of sparse pediatric imaging datasets via multi-scale contrastive regularization and multi-joint anatomical priors," *Medical Image Analysis*, p. 102556, 2022.

[36] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3D self-supervised methods for medical imaging," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 158–18 172, 2020.

[37] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[38] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu, "Federated contrastive learning for volumetric medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 367–377.

[39] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.

[40] J. Ma, Y. Zhang, S. Gu, X. An, Z. Wang, C. Ge, C. Wang, F. Zhang, Y. Wang, Y. Xu *et al.*, "Fast and low-GPU-memory abdomen CT organ segmentation: The FLARE challenge," *Medical Image Analysis*, p. 102616, 2022.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021.

[43] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[44] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[45] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision*, 2016, pp. 565–571.

[46] P.-H. Conze, S. Brochard, V. Burdin, F. T. Sheehan, and C. Pons, "Healthy versus pathological learning transferability in shoulder muscle MRI segmentation using deep convolutional encoder-decoders," *Computerized Medical Imaging and Graphics*, vol. 83, p. 101733, 2020.

[47] Z. Zhou, Z. He, and Y. Jia, "AFPNet: A 3D fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images," *Neurocomputing*, vol. 402, pp. 235–244, 2020.

[48] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[49] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[50] E. Panfilov, A. Tiulpin, S. Klein, M. T. Nieminen, and S. Saarakkala, "Improving robustness of deep learning based knee MRI segmentation: MixUp and adversarial domain adaptation," in *IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[51] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold MixUp: Better representations by interpolating hidden states," in *International Conference on Machine Learning*, 2019, pp. 6438–6447.

[52] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[53] S. Amiri and B. Ibragimov, "Improved automated lesion segmentation in whole-body FDG/PET-CT via test-time augmentation," *arXiv preprint arXiv:2210.07761*, 2022.

[54] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[55] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (Auto-Net) for brain extraction in magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2319–2330, 2017.

[56] Y. Wang, Z. Deng, X. Hu, L. Zhu, X. Yang, X. Xu, P.-A. Heng, and D. Ni, "Deep attentional features for prostate segmentation in ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 523–530.

[57] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi *et al.*, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 415–423.

[58] T. L. Kline, P. Korfiatis, M. E. Edwards, J. D. Blais, F. S. Czerwiec, P. C. Harris, B. F. King, V. E. Torres, and B. J. Erickson, "Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys," *Journal of Digital Imaging*, vol. 30, no. 4, pp. 442–448, 2017.

[59] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[60] P.-H. Conze, A. E. Kavur, E. Cornec-Le Gall, N. S. Gezer, Y. Le Meur, M. A. Selver, and F. Rousseau, "Abdominal multi-organ segmentation

with cascaded convolutional and adversarial deep networks," *Artificial Intelligence in Medicine*, vol. 117, p. 102109, 2021.

[61] B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, J. Qin, S. Chen, T. Wang, and S. Wang, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Medical Image Analysis*, vol. 64, p. 101716, 2020.

[62] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International Conference on Information Processing in Medical Imaging*, 2017, pp. 597–609.

[63] C. Chen, X. Liu, M. Ding, J. Zheng, and J. Li, "3D dilated multi-fiber network for real-time brain tumor segmentation in MRI," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 184–192.

[64] D. Jin, D. Guo, T.-Y. Ho, A. P. Harrison, J. Xiao, C.-k. Tseng, and L. Lu, "Deep esophageal clinical target volume delineation using encoded 3D spatial context of tumors, lymph nodes, and organs at risk," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 603–612.

[65] J. Ma and X. Yang, "Combining CNN and hybrid active contours for head and neck tumor segmentation in CT and PET images," in *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, 2020, pp. 59–64.

[66] W. Lei, H. Mei, Z. Sun, S. Ye, R. Gu, H. Wang, R. Huang, S. Zhang, S. Zhang, and G. Wang, "Automatic segmentation of organs-at-risk from head-and-neck CT using separable convolutional neural network with hard-region-weighted loss," *Neurocomputing*, vol. 442, pp. 184–199, 2021.

[67] Y. Yan, P.-H. Conze, G. Quellec, M. Lamard, B. Cochener, and G. Coatrieux, "Two-stage multi-scale breast mass segmentation for full mammogram analysis without user intervention," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 746–757, 2021.

[68] A. Vakanski, M. Xian, and P. E. Freer, "Attention-enriched deep learning model for breast tumor segmentation in ultrasound images," *Ultrasound in Medicine & Biology*, vol. 46, no. 10, pp. 2819–2833, 2020.

[69] T. L. Kline, M. E. Edwards, J. Fetzer, A. V. Gregory, D. Anaam, A. J. Metzger, and B. J. Erickson, "Automatic semantic segmentation of kidney cysts in MR images of patients affected by autosomal-dominant polycystic kidney disease," *Abdominal Radiology*, vol. 46, no. 3, pp. 1053–1061, 2021.

[70] V. Oreiller, V. Andrearczyk, M. Jreige, S. Boughdad, H. Elhalawani, J. Castelli, M. Vallières, S. Zhu, J. Xie, Y. Peng *et al.*, "Head and neck tumor segmentation in PET/CT: the HECKTOR challenge," *Medical Image Analysis*, vol. 77, p. 102336, 2022.

[71] F. Ouhmich, V. Agnus, V. Noblet, F. Heitz, and P. Pessaux, "Liver tissue segmentation in multiphase ct scans using cascaded convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 8, pp. 1275–1284, 2019.

[72] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert *et al.*, "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *International MICCAI BrainLesion Workshop*, 2017, pp. 450–462.

[73] D. Keshwani, Y. Kitamura, S. Ihara, S. Iizuka, and E. Simo-Serra, "TopNet: Topology preserving metric learning for vessel tree reconstruction and labelling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 14–23.

[74] A. Sadikine, B. Badic, J.-P. Tasu, V. Noblet, D. Visvikis, and P.-H. Conze, "Semi-overcomplete convolutional auto-encoder embedding as shape priors for deep vessel segmentation," in *IEEE International Conference on Image Processing*, 2022.

[75] M. Hatt, C. Parmar, J. Qi, and I. El Naqa, "Machine (deep) learning methods for image processing and radiomics," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 104–108, 2019.

[76] M. M. Islam, B. Badic, T. Aparicio, D. Tougeron, J.-P. Tasu, D. Visvikis, and P.-H. Conze, "Deep treatment response assessment and prediction of colorectal cancer liver metastases," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 482–491.

[77] C. Pons, F. T. Sheehan, H. S. Im, S. Brochard, and K. E. Alter, "Shoulder muscle atrophy and its relation to strength loss in obstetrical brachial plexus palsy," *Clinical Biomechanics*, vol. 48, pp. 80–87, 2017.

[78] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[79] Y. Huo, Z. Xu, S. Bao, C. Bermudez, H. Moon, P. Parvathaneni, T. K. Moyo, M. R. Savona, A. Assad, R. G. Abramson *et al.*, "Splenomegaly segmentation on multi-modal MRI using deep convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1185–1196, 2018.

[80] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 203–211.

[81] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[82] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[83] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1744–1757, 2010.

[84] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640–3649.

[85] J. Wei, Z. Wu, L. Wang, T. D. Bui, L. Qu, P.-T. Yap, Y. Xia, G. Li, and D. Shen, "A cascaded nested network for 3T brain MR image segmentation guided by 7T labeling," *Pattern Recognition*, vol. 124, p. 108420, 2022.

[86] M. S. Nosrati and G. Hamarneh, "Incorporating prior knowledge in medical image segmentation: a survey," *arXiv preprint arXiv:1607.01092*, 2016.

[87] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "DCAN: Deep contour-aware networks for object instance segmentation from histology images," *Medical Image Analysis*, vol. 36, pp. 135–146, 2017.

[88] P.-A. Ganaye, M. Sdika, B. Triggs, and H. Benoit-Cattin, "Removing segmentation inconsistencies with semi-supervised non-adjacency constraint," *Medical Image Analysis*, vol. 58, p. 101551, 2019.

[89] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.

[90] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[91] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[92] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[93] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[95] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[96] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[97] A. Iantsen, D. Visvikis, and M. Hatt, "Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images," in *Head and Neck Tumor Segmentation Challenge*, vol. 12603, 2020, pp. 37–43.

[98] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi *et al.*, "USE-Net: Incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets," *Neurocomputing*, vol. 365, pp. 31–43, 2019.

[99] X. Li, Y. Wei, L. Wang, S. Fu, and C. Wang, "MSGSE-Net: Multi-scale guided squeeze-and-excitation network for subcortical brain structure segmentation," *Neurocomputing*, vol. 461, pp. 228–243, 2021.

[100] X. Shen, J. Xu, H. Jia, P. Fan, F. Dong, B. Yu, and S. Ren, "Self-attentional microvessel segmentation via squeeze-excitation transformer unet," *Computerized Medical Imaging and Graphics*, vol. 97, p. 102055, 2022.

[101] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention U-Net for lesion segmentation," in *IEEE International Symposium on Biomedical Imaging*, 2019, pp. 683–687.

[102] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel squeeze and excitation blocks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 540–549, 2018.

[103] C. Yao, J. Tang, M. Hu, Y. Wu, W. Guo, Q. Li, and X.-P. Zhang, "Claw U-Net: A U-Net variant network with deep feature concatenation for scleral blood vessel segmentation," in *International Conference on Artificial Intelligence*, 2021, pp. 67–78.

[104] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation." *IEEE Journal of Biomedical and Health Informatics*, 2020.

[105] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze & excitation in fully convolutional networks," *ArXiv*, 2018.

[106] W. Fang and X.-h. Han, "Spatial and channel attention modulated network for medical image segmentation," in *Asian Conference on Computer Vision*, 2020.

[107] V. V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker, "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 547–556.

[108] X. Zhou, "Automatic segmentation of multiple organs on 3D CT images by using deep learning approaches," *Deep Learning in Medical Image Analysis*, pp. 135–147, 2020.

[109] S. A. Taghanaki, Y. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, 2019.

[110] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu *et al.*, "NiftyNet: a deep-learning platform for medical imaging," *Computer Methods and Programs in Biomedicine*, vol. 158, pp. 113–122, 2018.

[111] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-UNet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44 247–44 257, 2019.

[112] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[113] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[114] E. Jun, S. Jeong, D.-W. Heo, and H.-I. Suk, "Medical Transformer: Universal brain encoder for 3D MRI analysis," *arXiv preprint arXiv:2104.13633*, 2021.

[115] J. Li, J. Chen, Y. Tang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with Transformers? a comparative review of key properties, current progresses, and future perspectives," *arXiv preprint arXiv:2206.01136*, 2022.

[116] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[117] G. Andrade-Miranda, V. Jaouen, V. Bourbonne, F. Lucia, D. Visvikis, and P.-H. Conze, "Pure versus hybrid Transformers for multi-modal brain tumor segmentation: a comparative study," in *IEEE International Conference on Image Processing*, 2022.

[118] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *IEEE/CVF Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[119] D. Karimi, S. D. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using Transformers," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 78–88.

[120] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MR images," *arXiv preprint arXiv:2201.01266*, 2022.

[121] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," *arXiv preprint arXiv:2108.06932*, 2021.

[122] Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. Ö. Arik, and T. Pfister, "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding," in *Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3417–3425.

[123] X. Yu, Y. Tang, Y. Zhou, R. Gao, Q. Yang, H. H. Lee, T. Li, S. Bao, Y. Huo, Z. Xu *et al.*, "Characterizing renal structures with 3D block aggregate Transformers," *arXiv preprint arXiv:2203.02430*, 2022.

[124] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical Transformer: Gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 36–46.

[125] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using Transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 109–119.

[126] J. Li, W. Wang, C. Chen, T. Zhang, S. Zha, H. Yu, and J. Wang, "TransBTSv2: Wider instead of deeper transformer for medical image segmentation," *arXiv preprint arXiv:2201.12785*, 2022.

[127] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and Transformer for 3D medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 171–180.

[128] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical image segmentation using squeeze-and-expansion Transformers," *arXiv preprint arXiv:2105.09511*, 2021.

[129] B. Chen, Y. Liu, Z. Zhang, G. Lu, and D. Zhang, "TransAttUNet: Multi-level attention-guided U-Net with Transformer for medical image segmentation," *arXiv preprint arXiv:2107.05274*, 2021.

[130] L. Liu, Z. Huang, P. Liò, C.-B. Schönlieb, and A. I. Aviles-Rivero, "PC-SwinMorph: Patch representation for unsupervised medical image registration and segmentation," *arXiv preprint arXiv:2203.05684*, 2022.

[131] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing Transformers and CNNs for medical image segmentation," *arXiv preprint arXiv:2102.08005*, 2021.

[132] Q. Sun, N. Fang, Z. Liu, L. Zhao, Y. Wen, and H. Lin, "HybridCTrm: Bridging CNN and Transformer for multimodal brain image segmentation," *Journal of Healthcare Engineering*, vol. 2021, 2021.

[133] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between CNN and Transformer," in *Medical Imaging with Deep Learning*, 2022.

[134] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnFormer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[135] Y. Wu, K. Liao, J. Chen, D. Z. Chen, J. Wang, H. Gao, and J. Wu, "D-former: A U-shaped dilated Transformer for 3D medical image segmentation," *arXiv preprint arXiv:2201.00462*, 2022.

[136] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure Transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[137] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin Transformer U-Net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, 2022.

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

[139] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A volumetric Transformer for accurate 3D tumor segmentation," *arXiv preprint arXiv:2111.13300*, 2021.

[140] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[141] P. Moeskops, J. M. Wolterink, B. H. van der Velden, K. G. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning for multi-task medical image segmentation in multiple modalities," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 478–486.

[142] C. Playout, R. Duval, and F. Cheriet, "A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 101–108.

[143] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, and M. Sivaprakasam, "Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2019, pp. 7223–7226.

[144] Y. Zhang, H. Li, J. Du, J. Qin, T. Wang, Y. Chen, B. Liu, W. Gao, G. Ma, and B. Lei, "3D multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1618–1631, 2021.

[145] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. v. Tulder, and M. d. Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 457–465.

[146] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[147] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[148] A. Jungo and M. Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 48–56.

[149] F. Galati, S. Ourselin, and M. A. Zuluaga, "From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review," *Applied Sciences*, vol. 12, no. 8, p. 3936, 2022.

[150] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2016, pp. 1050–1059.

[151] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.

[152] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, 2017, pp. 1321–1330.

[153] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Medical Image Analysis*, vol. 65, p. 101766, 2020.

[154] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical image analysis*, vol. 59, p. 101557, 2020.

[155] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *Medical Image Analysis*, vol. 60, p. 101619, 2020.

[156] Y. Zhao, C. Yang, A. Schweidtmann, and Q. Tao, "Efficient bayesian uncertainty estimation for nnu-net," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 535–544.

[157] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[158] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers *et al.*, "The medical segmentation decathlon," *Nature Communications*, vol. 13, no. 1, pp. 1–13, 2022.

[159] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1214–1223.

[160] X. Hu, R. Guo, J. Chen, H. Li, D. Waldmannstetter, Y. Zhao, B. Li, K. Shi, and B. Menze, "Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2599–2608, 2020.

[161] V. P. Sudarshan, U. Upadhyay, G. F. Egan, Z. Chen, and S. P. Awate, "Towards lower-dose PET using physics-based uncertainty-aware multimodal learning with robustness to out-of-distribution data," *Medical Image Analysis*, vol. 73, p. 102187, 2021.

[162] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 539–546.

[163] A. Jamaludin, T. Kadir, and A. Zisserman, "Self-supervised learning for spinal MRIs," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017, pp. 294–302.

[164] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 481–490.

[165] D. Zeng, Y. Wu, X. Hu, X. Xu, H. Yuan, M. Huang, J. Zhuang, J. Hu, and Y. Shi, "Positional contrastive learning for volumetric medical image segmentation," *arXiv preprint arXiv:2106.09157*, 2021.

[166] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," in *Neural Information Processing Systems*, 2014.

[167] K. Xu, L. Rui, Y. Li, and L. Gu, "Feature normalized knowledge distillation for image classification," in *European Conference on Computer Vision*, 2020, pp. 664–680.

[168] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[169] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 578–587.

[170] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.

[171] Y. Wen, L. Chen, S. Xi, Y. Deng, X. Tang, and C. Zhou, "Towards efficient medical image segmentation via boundary-guided knowledge distillation," in *IEEE International Conference on Multimedia and Expo*, 2021.

[172] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[173] E. Kats, J. Goldberger, and H. Greenspan, "Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation," in *IEEE International Symposium on Biomedical Imaging*, 2019, pp. 1563–1566.

[174] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[175] L. Zhang, S. Feng, Y. Wang, Y. Wang, Y. Zhang, X. Chen, and Q. Tian, "Unsupervised ensemble distillation for multi-organ segmentation," in *IEEE International Symposium on Biomedical Imaging*, 2022.

[176] R. Huang, Y. Zheng, Z. Hu, S. Zhang, and H. Li, "Multi-organ segmentation via co-training weight-averaged models from few-organ datasets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 146–155.

[177] N. Wang, S. Lin, X. Li, K. Li, Y. Shen, Y. Gao, and L. Ma, "MISSU: 3D medical image segmentation via self-distilling TransUNet," *arXiv preprint arXiv:2206.00902*, 2022.

[178] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[179] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[180] Y. Qin, K. Kamnitsas, S. Ancha, J. Nanavati, G. Cottrell, A. Criminisi, and A. Nori, "Autofocus layer for semantic segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 603–611.

[181] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[182] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[183] J. Shapey, G. Wang, R. Dorent, A. Dimitriadis, W. Li, I. Paddick, N. Kitchen, S. Bisdas, S. R. Saeed, S. Ourselin *et al.*, "An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI," *Journal of Neurosurgery*, vol. 134, no. 1, pp. 171–179, 2019.

[184] H.-Y. Yang, "Volumetric adversarial training for ischemic stroke lesion segmentation," in *MICCAI BrainLesion Workshop*, 2018, pp. 343–351.

[185] C. Zhou, C. Ding, Z. Lu, X. Wang, and D. Tao, "One-pass multi-task convolutional neural networks for efficient brain tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 637–645.

[186] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-modal image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 469–477.

[187] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3D biomedical segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6393–6400.

[188] T. Zhou, S. Canu, and S. Ruan, "Fusion based on attention mechanism and context constraint for multi-modal brain tumor segmentation," *Computerized Medical Imaging and Graphics*, vol. 86, p. 101811, 2020.

[189] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Latent correlation repre-

sentation learning for brain tumor segmentation with missing MRI modalities," *IEEE Transactions on Image Processing*, vol. 30, pp. 4263–4274, 2021.

[190] C. Li, H. Sun, Z. Liu, M. Wang, H. Zheng, and S. Wang, "Learning cross-modal deep representations for multi-modal MR image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 57–65.

[191] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1116–1126, 2018.

[192] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, "Cross-modality deep feature learning for brain tumor segmentation," *Pattern Recognition*, vol. 110, p. 107562, 2021.

[193] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image and Vision Computing*, vol. 105, p. 104042, 2021.

[194] Y. Zhang, J. Yang, J. Tian, Z. Shi, C. Zhong, Y. Zhang, and Z. He, "Modality-aware mutual learning for multi-modal medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021, pp. 589–599.

[195] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[196] R. Dorent, A. Kujawa, M. Ivory, S. Bakas, N. Rieke, S. Joutard, B. Glocker, J. Cardoso, M. Modat, K. Batmanghelich *et al.*, "CrossMoDA 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation," *Medical Image Analysis*, vol. 83, p. 102628, 2023.

[197] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 5, pp. 1–46, 2020.

[198] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert, "Data efficient unsupervised domain adaptation for cross-modality image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 669–677.

[199] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, "A lifelong learning approach to brain MR segmentation across scanners and protocols," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 476–484.

[200] Q. Dou, Q. Liu, P. A. Heng, and B. Glocker, "Unpaired multi-modal segmentation via knowledge distillation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2415–2425, 2020.

[201] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9242–9251.

[202] S.-Y. Hu, S. Wang, W.-H. Weng, J. Wang, X. Wang, A. Ozturk, Q. Li, V. Kumar, and A. E. Samir, "Self-supervised pretraining with DICOM metadata in ultrasound imaging," in *Machine Learning for Healthcare Conference*, 2020, pp. 732–749.

[203] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[204] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[205] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[206] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 605–613.

[207] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.

[208] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-Net: Automatic COVID-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[209] Y.-X. Zhao, Y.-M. Zhang, M. Song, and C.-L. Liu, "Multi-view semi-supervised 3D whole brain segmentation with a self-ensemble network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 256–265.

[210] D. Ng, X. Lan, M. M.-S. Yao, W. P. Chan, and M. Feng, "Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 2, p. 852, 2021.

[211] A. Chowdhury, H. Kassem, N. Padoy, R. Umeton, and A. Karargyris, "A review of medical federated learning: Applications in oncology and cancer research," in *International MICCAI Brainlesion Workshop*, 2022, pp. 3–24.

[212] X. Xu, T. Chen, H. Deng, T. Kuang, J. C. Barber, D. Kim, J. Gateno, P. Yan, and J. J. Xia, "Federated cross learning for medical image segmentation," *arXiv preprint arXiv:2204.02450*, 2022.

[213] J. Wicaksana, Z. Yan, D. Zhang, X. Huang, H. Wu, X. Yang, and K.-T. Cheng, "FedMix: Mixed supervised federated learning for medical image segmentation," *arXiv preprint arXiv:2205.01840*, 2022.

[214] T. Kim, K. Lee, S. Ham, B. Park, S. Lee, D. Hong, G. B. Kim, Y. S. Kyung, C.-S. Kim, and N. Kim, "Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: Evaluation on kidney segmentation in abdominal CT," *Scientific Reports*, vol. 10, no. 1, pp. 1–7, 2020.

[215] M. Shen, J. Y. Zhang, L. Chen, W. Yan, N. Jani, B. Sutton, and O. Koyejo, "Labeling cost sensitive batch active learning for brain tumor segmentation," in *IEEE International Symposium on Biomedical Imaging*, 2021, pp. 1269–1273.

[216] Y. Yan, P.-H. Conze, M. Lamard, H. Zhang, G. Quellec, B. Cochener, and G. Coatrieux, "Deep active learning for dual-view mammogram analysis," in *International Workshop on Machine Learning in Medical Imaging*, 2021, pp. 180–189.

[217] O. Yaniv, O. Portnoy, A. Talmon, N. Kiryati, E. Konen, and A. Mayer, "V-Net light-parameter-efficient 3D convolutional neural network for prostate MRI segmentation," in *IEEE International Symposium on Biomedical Imaging*, 2020, pp. 442–445.

[218] Q. Zhao, H. Wang, and G. Wang, "LCOV-NET: A lightweight neural network for Covid-19 pneumonia lesion segmentation from 3D CT images," in *IEEE International Symposium on Biomedical Imaging*, 2021, pp. 42–45.