Delving into Universal Lesion Segmentation: Method, Dataset, and Benchmark

Yu Qiu[0000−0001−6722−3039] and Jing Xu[0000−0001−8532−2241]

College of Artificial Intelligence, Nankai University, Tianjin 300350, China
yqiu@mail.nankai.edu.cn, xujing@nankai.edu.cn

Abstract. Most efforts on lesion segmentation from CT slices focus on one specific lesion type. However, universal and multi-category lesion segmentation is more important because the diagnoses of different body parts are usually correlated and carried out simultaneously. The existing universal lesion segmentation methods are weakly-supervised due to the lack of pixel-level annotation data. To bring this field into the fully-supervised era, we establish a large-scale universal lesion segmentation dataset, SegLesion. We also propose a baseline method for this task. Considering that it is easy to encode CT slices owing to the limited CT scenarios, we propose a Knowledge Embedding Module (KEM) to adapt the concept of dictionary learning for this task. Specifically, KEM first learns the knowledge encoding of CT slices and then embeds the learned knowledge encoding into the deep features of a CT slice to increase the distinguishability. With KEM incorporated, a Knowledge Embedding Network (KEN) is designed for universal lesion segmentation. To extensively compare KEN to previous segmentation methods, we build a large benchmark for SegLesion. KEN achieves state-of-the-art performance and can thus serve as a strong baseline for future research. The data and code have released at https://github.com/yuqiuyuqiu/KEN.

Keywords: Universal lesion segmentation · Lesion segmentation · Dictionary learning · Knowledge embedding

1 Introduction

When reading medical images such as computed tomography (CT), radiologists first need to search across the image to find lesions for further characterization and measurement [45]. To reduce radiologists' burden and improve accuracy, much effort has been paid to develop automatic lesion segmentation techniques [7,41,2,59,31,43,53,26]. Moreover, automatic lesion segmentation from CT slices also plays a crucial role in many computer-aided diagnosis (CAD) tasks such as pathology detection [34], tumor growth monitoring [20], and quantitative disease progression [28]. Recently, great progress has been brought to this field with the fast development of convolutional neural networks (CNNs), especially fully convolutional networks (FCNs) [38].

It is widely accepted for semantic segmentation models [38,55,10,48,42] that the performance gains mainly benefit from large amounts of accurately labeled

training data such as Cityscapes [12] and ADE20K [58] datasets. However, unlike natural images, medical images are difficult to obtain due to their high privacy and secrecy. What's worse, medical image annotation is not only time-consuming and expensive but also requires extensive clinical expertise, resulting in the lack of publicly available medical segmentation data [40,16,29]. Therefore, the biggest challenge of training an accurate lesion segmentation model is the lack of a large-scale dataset.

Besides the limited data scale, another problem is that the existing medical datasets only contain a particular lesion type such as the liver dataset [40], kidney tumor dataset [16], breast mass datasets [29,23], and lung nodules dataset [3]. Based on this, the existing lesion segmentation models usually focus on segmenting one specific lesion type from the corresponding body part [11,25,8]. However, in fact, many types of lesions are correlated with each other. For example, metastases can spread to new areas of the body through the lymph system or bloodstream. In practice, a patient should have radiological examinations on different body parts at the same time so that the radiologists are able to make a more accurate diagnosis by observing relevant clinical findings. Therefore, it is necessary to develop a universal and multi-category CAD framework, capable of segmenting multiple lesion types.

To address the above problems, we first establish a new large-scale multi-category radiological image dataset for lesion segmentation, namely **SegLesion**. SegLesion consists of 9623 lesions in 9456 CT slices with corresponding pixel-level annotations. These CT slices are collected from 4321 series of 3178 studies for 1356 unique patients. Different from existing datasets [16,3,40,29,23], SegLesion contains a variety of lesions, including lung nodules, liver lesions, enlarged lymph nodes, kidney lesions, bone lesions, and so on. SegLesion is based on the DeepLesion dataset [46] that only has bounding box annotations for universal lesion detection. With DeepLesion, many weakly-supervised lesion segmentation algorithms have been presented owing to the importance of lesion segmentation [7,41,2]. To bring universal lesion segmentation into the fully-supervised era, we establish SegLesion by carefully labeling the pixels of each lesion according to the bidimensional RECIST (Response Evaluation Criteria in Solid Tumours) diameters [13] (two short crossing lines on the lesion in Fig. 1) and the bounding boxes provided by DeepLesion.

Lesions of different types usually exhibit a wide variety of sizes, shapes, and appearances. For instance, as shown in Fig. 1, some lesions have a normal size, while other lesions only occupy a few pixels; and meanwhile, different shapes and appearances of lesions are also observed. As a result, it is suboptimal to directly apply the existing segmentation methods [36,38,57,5,9,10,60,32,51,48,56,55,42,14] to universal lesion segmentation. To design a better universal lesion segmenter, we observe that the scenarios of CT slices are very limited (*i.e.*, just human organs) and thus easy to be encoded. Motivated by this, we consider adopting the concept of *dictionary learning* for this task. To this end, we propose a Knowledge Embedding Module (KEM). KEM first encodes the knowledge of CT scenarios by learning a *dictionary* as in dictionary learning. Then, the learned dictionary

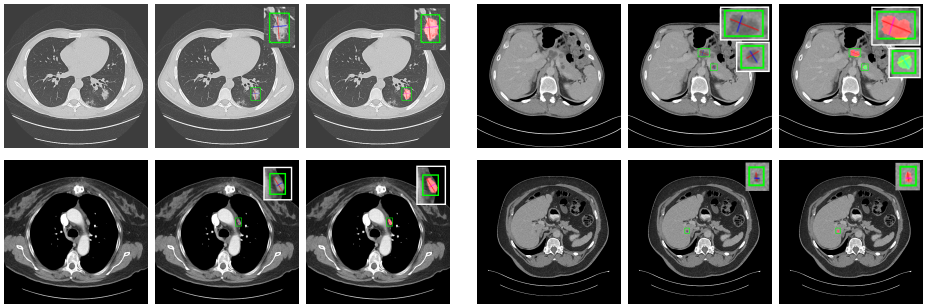


Fig. 1: Samples of CT slices without annotation, with RECIST-diameters and bounding-boxes, and with pixel-wise masks from left to right. Note that for visual clarity, we keep the bounding box in the right image.

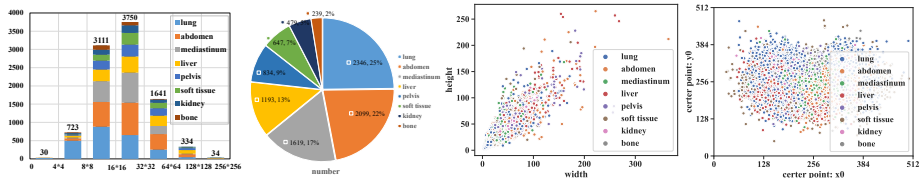
is embedded into the deep features of a CT slice to improve feature distinguishability. With KEM incorporated, we design an elegant network, *i.e.*, Knowledge Embedding Network (KEN). Despite the wide variety, KEN can accurately segment lesions through dataset-level knowledge encoding.

For extensively comparing our method with previous methods, we build a comprehensive benchmark, including well-known methods for both medical image segmentation and semantic image segmentation. We adopt four popular evaluation metrics in medical image segmentation for evaluation. Experimental results demonstrate that our method performs favorably against previous state-of-the-art methods and thus can be served as a strong baseline for future research on this topic. This comprehensive benchmark would also be useful for future research. We summarize our contributions as follows:

- We establish a large-scale multi-category lesion segmentation dataset, SegLesion, with high-quality annotations, for universal lesion segmentation.
- We propose a universal lesion segmentation method, KEN, by embedding the learned data knowledge of CT scenarios into the deep features of a CT slice to increase the distinguishability.
- We build a comprehensive segmentation benchmark for our new SegLesion dataset to promote future research on this topic.

2 Related Work

Lesion Segmentation Dataset. There have existed some lesion segmentation datasets for a specific lesion. For example, the 3D-IRCADb dataset [40] and the Liver Tumor Segmentation (LiTS) challenge organized in MICCAI 2017 [6] are two competitive and widely-used datasets for liver lesion segmentation. The 2019 Kidney Tumor Segmentation Challenge (KiTS19) [16] is for kidney tumor segmentation by collecting arterial phase abdominal CT scans from 300 patients who underwent partial or radical nephrectomy. INbreast [29] and DDSM-CBIS



(a) (b) (c) (d)

Fig. 2: Data statistics of the new SegLesion dataset. (a) Lesion size distribution; (b) Numbers of lesions of different types; (c) Lesions' height *vs.* width; (d) Coordinates of lesion center points.

[23] are two popular datasets for breast mass segmentation. There are also other datasets targeting on pelvic mass segmentation, thyroid nodule segmentation, axillary lymph node segmentation, and so on [4,3]. However, the scales of these datasets are limited, and each dataset only contains a specific lesion type, making it difficult to train a universal lesion segmentation framework.

Lesion Segmentation. Recently, CNNs, especially FCNs, have been widely applied for lesion segmentation, such as U-Net [36], UNet++ [60], and Attention U-Net [32]. For example, Cao *et al.* [8] proposed a dual-branch residual network for lung nodule segmentation. Christ *et al.* [11] designed a cascaded FCN for liver tumor segmentation. Besides, segmentation models designed for natural images [38,55,10,48,42] can also be used for medical image segmentation seamlessly. However, our SegLesion dataset is more challenging than existing datasets due to its lower contrast, more complicated distribution, and larger anatomical variability in size and shape, making previous segmentation methods unable to achieve satisfactory results. In this paper, we propose a dictionary-learning-based method according to the characteristics of medical images, which can serve as a strong baseline for future research on universal lesion segmentation.

3 SegLesion Dataset

3.1 Data Collection and Annotation

Yan *et al.* [46] introduced a large-scale medical image dataset, *i.e.*, DeepLesion, by releasing CT slices that have been collected for two decades in their institute. DeepLesion collects 32735 lesions in 32120 CT slices from 10594 studies of 4427 unique patients. Lesions of DeepLesion are annotated by bidimensional RECIST diameters (Response Evaluation Criteria in Solid Tumours) diameters [13] (two short crossing lines on the lesion in Fig. 1) that can tell us the location of each lesion. The authors automatically generated bounding boxes for these lesions by adding 5-pixel padding to each direction (left, top, right, and bottom) of the bounding box of RECIST diameters. The authors then selected 9816 lesions in 9624 CT slices and manually labeled them into eight types (*i.e.*, lung, abdomen,

mediastinum, liver, pelvis, soft tissue, kidney, and bone) to form a universal and multi-category lesion detection dataset. With DeepLesion, many studies have emerged for lesion detection [46], weakly-supervised lesion segmentation [7,41,2], body part recognition [44], lesions relationship learning [47], and so on. Note that DeepLesion only provides bounding box annotations, so only weakly-supervised lesion segmentation can be explored on this dataset. Our goal of this paper is to bring universal lesion segmentation to the fully-supervised era.

To this end, we carefully label the selected 9816 lesions in 9624 CT slices with pixel-level masks. Unlike DeepLesion that automatically generates the bounding box for each lesion, we manually label the mask of each lesion using the online annotation tool of Polygon-RNN++ [1] [1]. To ensure the accuracy and reliability, we conduct a triple-check annotation process, and the third annotator is an experienced doctor. In detail, the first annotator labels lesions with masks according to the RECIST diameters that have indicated the location of each lesion. After finishing this job, the first annotator checks the annotations carefully and re-annotates unsatisfactory ones (1^{st} check). Then, the annotations are re-checked by the second annotator. If he has different opinions on some annotations, the first and second annotators will discuss and re-annotate these CT slices together (2^{nd} check). At last, the annotations are further re-checked by the third annotator who is an experienced doctor. If he disagrees with some annotations, these three annotators will discuss and make re-annotations together (3^{rd} check). Moreover, we abandon some CT slices with lesions whose boundaries are fuzzy for accurate recognition or whose masks are too small to label manually. Finally, SegLesion is composed of 9623 lesions in 9456 CT slices from 4321 series of 3178 studies of 1356 unique patients. We show eight examples with different annotation types in Fig. 1.

3.2 Data Statistics

All CT slices in SegLesion are in a resolution of 512×512. The distribution of the lesion sizes is shown in Fig. 2a, from which we can see that most lesions only occupy a small part of the whole image. In detail, about 71.3% of lesions have a size ranging from 64 pixels to 1024 pixels. The number of lesions whose sizes are over 4096 (64×64) pixels is 368, only accounting for 3.8% of all lesions in SegLesion. Among all lesions, the smallest one only has 8 pixels, and the largest one has 57161 pixels, occupying 21.8% of the entire image. We also plot a height *vs.* width figure for all lesions, as shown in Fig. 2c. It is easy to see that the lesions in SegLesion are very small in general. In Fig. 2d, we plot the locations of center points of lesions. We can observe that the lesions are randomly distributed on the CT slices without bias, indicating the universal property of SegLesion.

Following DeepLesion, we coarsely divide the 9456 CT slices of SegLesion into eight types, including lung (2346), abdomen (2099), mediastinum (1619), liver (1193), pelvis (834), soft tissue (647), kidney (479), and bone (239), as depicted in Fig. 2b. Mediastinum lesions are mainly lymph nodes in the chest.

[1] http://www.cs.toronto.edu/~amlan/demo/

Table 1: Number of CT slices of different lesion types in each split. Note that data splitting is conducted at the patient level. ME: Mediastinum; ST: Soft tissue.

Splits	Lung	Abdomen	ME	Liver	Pelvis	ST	Kidney	Bone	Total
Training	1575	1435	1149	837	573	451	328	175	6523
Validation	354	333	242	167	139	114	84	35	1468
Test	417	331	228	189	122	82	67	29	1465
Total	2346	2099	1619	1193	834	647	479	239	9456

The abdomen type consists of miscellaneous lesions that are not in the liver or kidney. The soft tissue type refers to lesions in the muscle, skin, and fat. To facilitate and standardize the future use, SegLesion is randomly split into training, validation, and test sets *at the patient level*, accounting for about 70%, 15%, and 15% of lesions of each type, respectively. A summary of dataset splits and lesion types can be found in Tab. 1. Since SegLesion contains multi-category CT slices with a wide variety of sizes, shapes, and appearances, it is possible to use SegLesion to train universal lesion segmentation frameworks.

3.3 Potential Applications

The potential applications of SegLesion include:

- Lesion segmentation: This is a direct application of the SegLesion dataset. Unlike previous lesion segmentation for only one specific lesion type, SegLesion is the first public large-scale dataset for universal lesion segmentation. More future researches are expected to push this field to clinical applications.
- 3D lesion segmentation: By combining 2D masks with two-dimensional diameter measurements in DeepLesion [46], we can develop weakly-supervised 3D segmentation algorithms to analyze lesions in a 3D view.
- Lesion retrieval: With lesion masks, it is convenient to conduct region- and context-based lesion retrieval. This would benefit the clinical diagnosis by finding the most similar lesion cases given a query CT slice.
- Lesion growth analysis: Lesion masks in our SegLesion can provide better information than the bounding boxes in DeepLesion [46] for analyzing lesion changes based on their sizes, shapes, and appearances.

As discussed above, this paper explores universal lesion segmentation on the SegLesion dataset by proposing an effective baseline method and building a comprehensive segmentation benchmark.

3.4 Data Naming

We follow the similar naming pattern of DeepLesion [46], *i.e.*, the real patient IDs, accession numbers, and series numbers are replaced by self-defined indices of patient, study, and series (starting from 1) for anonymization. Therefore, each CT slice in SegLesion is named with the format "{patient index}_{study index}_{series index}_{slice index}".

4 Methodology

4.1 Knowledge Embedding Module

Our technical motivation comes from the traditional concept of *sparse dictionary learning*, also called *sparse coding*. Sparse coding aims at representing the input data with a linear combination of basic elements. These basic elements are called *atoms* or *codewords*, which compose a *dictionary*. We observe that the scenarios of CT slices are very limited, *i.e.*, just medical imaging of human organs, unlike natural images that would have countless unforeseen new scenarios, so it is easy to encode CT scenarios. Instead of using traditional optimization algorithms for dictionary learning, this paper tries to adapt the idea of dictionary learning to deep learning. Specifically, we leverage CNNs to learn codewords for all CT training data. Then, we embed the learned knowledge, *i.e.*, codewords, into CT features to increase the distinguishability of abnormal and normal areas for easing subsequent pixel-wise classification, which is dubbed *knowledge embedding*.

In detail, we aim at learning K codewords $\mathbf{V}_k \in \mathbb{R}^C$ $(k \in \{1, 2, \cdots, K\})$, *i.e.*, $\mathbf{V} \in \mathbb{R}^{K \times C}$, which encodes the essential knowledge of *all CT slices*. We also learn a scale matrix $\mathbf{S} \in \mathbb{R}^{K \times C}$ for knowledge embedding. Hence, $\mathbf{V}$ and $\mathbf{S}$ are learnable encoding variables. Suppose $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ is the deep feature map extracted from a CT slice $\mathbf{I} \in \mathbb{R}^{H' \times W'}$ using a deep FCN. Here, C, H, and W are the number of channels, height, and width of the feature map, respectively. Similarly, H' and W' are height and width of the input CT slice, respectively. In this paper, we use the convolutional part of VGG16 [39] or ResNet50 [15] with a stride of 8 for feature extraction, so we have $H = H'/8$ and $W = W'/8$. Our specific idea is to first learn codewords $\mathbf{V}$ and scale $\mathbf{S}$ for all data and then embed the learned information into each pixel of the feature map $\mathbf{X}$ to get a new feature map $\mathcal{T}(\mathbf{X}) \in \mathbb{R}^{C \times H \times W}$, where $\mathcal{T}(\cdot)$ can be viewed as a transformation function for this process. Benefiting from the universal knowledge, the new feature map $\mathcal{T}(\mathbf{X})$ is expected to be easier for pixel-wise classification, and better segmentation results can thus be achieved.

With the above motivation and definitions, we continue by proposing KEM for our goal. As illustrated in Fig. 3, the feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ is first reshaped to $\mathbb{R}^{N \times C}$ $(N = H \times W)$ and then replicated to $\mathbb{R}^{N \times K \times C}$. We subtract $\mathbf{V}$ from $\mathbf{X}$ in the spirit of residual learning, which can be formulated as

$$\mathbf{D} = \mathbf{X} - \mathbf{V}, \quad \mathbf{D} \in \mathbb{R}^{N \times K \times C}, \tag{1}$$

where $\mathbf{V}$ is reshaped and replicated to the same size as $\mathbf{X}$ before subtraction. In this way, we obtain the residual values $\mathbf{D}$ between the feature of each point in $\mathbf{X}$ and each codeword in $\mathbf{V}$. Next, *local scale coefficients* are computed for aggregating $\mathbf{D}$ as

$$\begin{aligned} \mathbf{A}_{:,k,:} &= \frac{\exp(\mathbf{S}_{k,:} \otimes \mathbf{D}^2_{:,k,:})}{\sum_{k' \in \{1,2,\cdots,K\}} \exp(\mathbf{S}_{k',:} \otimes \mathbf{D}^2_{:,k',:})}, \\ k \in \{1, 2, \cdots, K\}, &\qquad \mathbf{A} \in \mathbb{R}^{N \times K \times C}, \end{aligned} \tag{2}$$

in which $\otimes$ means element-wise multiplication, before which two matrices are replicated to the same size. Inspired by the *softmax* function, Eq. (2) ensures $\sum_{k\in\{1,2,\cdots,K\}} \mathbf{A}_{n,k,c} = 1$ for $\forall n \in \{1, 2, \cdots, N\}$ and $\forall c \in \{1, 2, \cdots, C\}$. In this way, the local scale coefficient matrix $\mathbf{A}$ is not only derived from the learnable variable $\mathbf{S}$ but also the residual values $\mathbf{D}$, containing both universal and input-related properties. Here, "local" means that the local weight $\mathbf{A}$ at each position is determined by $\mathbf{D}$ at the same position.

With $\mathbf{D}$ and $\mathbf{A}$ prepared, we can calculate the reweighted residual values as

$$\mathbf{D}' = \mathbf{D} \otimes \mathbf{A}, \quad \mathbf{D}' \in \mathbb{R}^{N\times K\times C}. \tag{3}$$

$\mathbf{D}'$ is aggregated along the dimension of K:

$$\mathbf{D}'' = \sum_{k\in\{1,2,\cdots,K\}} \mathbf{D}'_{:,k,:}, \tag{4}$$

after which $\mathbf{D}''$ is reshaped to $\mathbf{D}'' \in \mathbb{R}^{C\times H\times W}$, the same size as $\mathbf{X}$. Now, *global scale coefficients* are further computed for $\mathbf{D}''$ to enhance its intra-channel representation. This operation is called **Embedding Re-scaling (ER)**. ER first aggregates $\mathbf{D}'$ along the dimension of K to summarize the residual values in terms of different codewords, and then aggregates the result along the dimension of N to compute the global information of a CT slice, which can be expressed as

$$\mathbf{E} = \sum_{n\in\{1,2,\cdots,N\}} \sum_{k\in\{1,2,\cdots,K\}} \mathbf{D}'_{n,k,:}, \tag{5}$$

where we have $\mathbf{E} \in \mathbb{R}^C$ representing the overall encoder of the input CT slice $\mathbf{I}$. Next, $\mathbf{E}$ is transformed to an attention vector by

$$\mathbf{E}' = \sigma(\mathrm{FC}(\mathbf{E})), \quad \mathbf{E}' \in \mathbb{R}^C, \tag{6}$$

in which $\mathrm{FC}(\cdot)$ is a fully-connected layer and $\sigma(\cdot)$ is the standard *sigmoid* function. Then, $\mathbf{E}'$ is replicated to $\mathbf{E}' \in \mathbb{R}^{C\times H\times W}$. The output of KEM is easy to write as

$$\mathbf{Y} = \mathbf{D}'' \otimes \mathbf{E}' + \mathbf{D}'' + \mathbf{X}, \quad \mathbf{Y} \in \mathbb{R}^{C\times H\times W}. \tag{7}$$

This intra-channel representation enhancement is different from traditional channel attention [17] that only relies on the feature map (*i.e.*, self-attention), because $\mathbf{E}'$ is based on both the feature map and the learned universal data knowledge.

The proposed KEM can be viewed as an extension of traditional *dictionary learning*. Specifically, KEM defines learnable variables ($\mathbf{V}$ and $\mathbf{S}$) to encode the scenarios of CT slices, achieved by making the inherent dictionary differentiable. $\mathbf{V} \in \mathbb{R}^{K\times C}$ can be viewed as K *codewords* of the dictionary. The learned dictionary is embedded into the feature map $\mathbf{X}$ in a pixel-wise manner to construct a new feature map $\mathbf{Y}$. The universal knowledge in the dictionary would increase the feature distinguishability, so it is easier to discriminate each pixel in $\mathbf{Y}$ to be normal or abnormal. The input $\mathbf{X}$ and output $\mathbf{Y}$ of KEM have the same size so that KEM is flexible to be plugged into any CNNs.

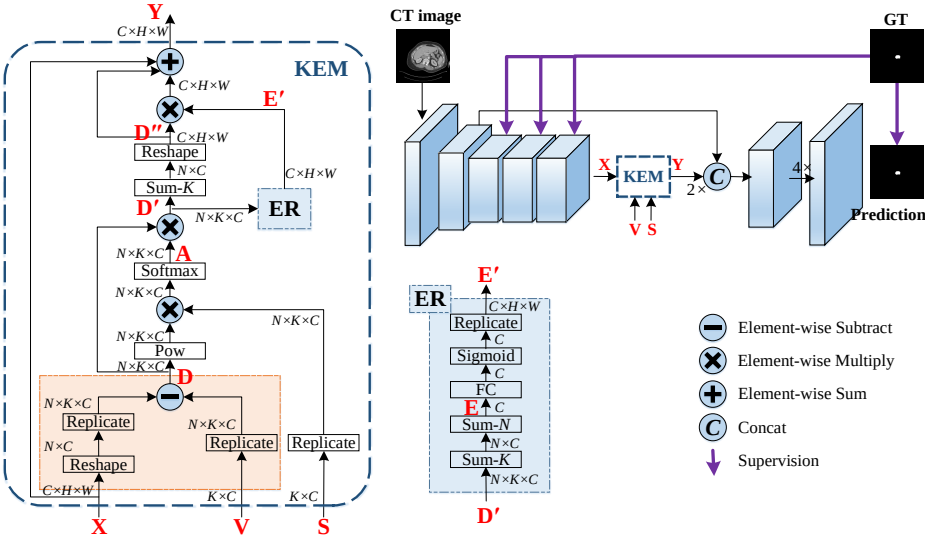


Fig. 3: Network architecture of the proposed KEN.

4.2 Knowledge Embedding Network

In this part, we elaborate on KEN by incorporating KEM. Let us take VGG16 [39] as an example, and the ResNet50 [15] version can be similarly defined. As shown in Fig. 3, we remove all fully-connected layers in VGG16 to obtain an FCN, so that we can obtain five convolutional feature maps, *i.e.*, $\mathbf{X}^{(i)}$ ($i \in 1, 2, \cdots, 5$), corresponding to five convolutional stages of VGG16. Besides the existing convolution layers, we add two more convolutions to deepen VGG16, *i.e.*,

$$\begin{aligned} \mathbf{X}_1^{(6)} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}^{3\times3}(\mathbf{X}^{(5)}))), \\ \mathbf{X}_2^{(6)} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}^{1\times1}(\mathbf{X}_1^{(6)}))). \end{aligned} \tag{8}$$

Here, $\mathrm{Conv}^{3\times3}(\cdot)$ and $\mathrm{Conv}^{1\times1}(\cdot)$ are a 3×3 convolution with 512 output channels and a 1×1 convolution with 1024 output channels, respectively. $\mathrm{BN}(\cdot)$ and $\mathrm{ReLU}(\cdot)$ are batch normalization [19] and ReLU [30] layers, respectively. Note that ResNet50 does not need these two more convolutions because ResNet50 is deep enough for lesion segmentation. Following previous segmentation methods [57,48,52,56,55,18,61], we change the stride of the last two downsampling layers from 2 to 1, leading to the smallest scale of 1/8. Dilated convolutions are used to keep the receptive field.

A KEM is put on top of $\mathbf{X}_2^{(6)}$ to embed the universal knowledge into it, which can be written as

$$\begin{aligned} \dot{\mathbf{X}}^{(6)} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}^{1\times1}(\mathbf{X}_2^{(6)}))), \\ \hat{\mathbf{X}}^{(6)} &= \mathrm{ReLU}(\mathrm{BN}(\mathcal{T}(\dot{\mathbf{X}}^{(6)}))), \end{aligned} \tag{9}$$

Table 2: Effects of the main components of KEN. The symbol ✔ indicates that a design choice is used. The 0^{th} column is the results of the standard FCN, and the 4^{th} column is the results of our final model in this paper.

Components		0	1	2	3	4	5	6	7	8
$\mathbf{X}^{(6)}$			✔	✔	✔	✔	✔	✔	✔	✔
KEM				✔	✔	✔			✔	✔
Decoder					✔	✔	✔	✔		
Deep supervision						✔	✔	✔	✔	✔
KEM w/o ER							✔			
KEM w/o $\mathbf{D}''$								✔		
Decoder w/ $\mathbf{X}^{(2)}$									✔	
Decoder w/ $\mathbf{X}^{(4)}$										✔
Metrics (%)	mIoU ↑	62.56	63.05	64.35	64.80	**65.78**	65.15	64.45	65.63	65.15
	SEN ↑	30.98	33.24	46.30	50.76	51.09	42.92	42.67	46.74	**52.11**
	SPE ↑	67.48	67.72	92.43	95.31	97.48	81.30	96.02	96.51	**97.94**
	DSC ↑	27.32	28.61	35.68	39.19	**41.01**	34.73	32.91	39.45	37.79

where $\text{Conv}^{1\times1}$ is a 1×1 convolution with 256 output channels. Then, we upsample $\hat{\mathbf{X}}^{(6)}$ from 1/8 scale to 1/4 scale and fuse it with $\mathbf{X}^{(3)}$ that is in 1/4 scale, like

$$\begin{aligned} \hat{\mathbf{X}}^{(3)} &= \text{ReLU}(\text{BN}(\text{Conv}^{1\times1}(\mathbf{X}^{(3)}))), \\ \mathbf{X}^{\text{concat}} &= \text{Concat}(\hat{\mathbf{X}}^{(3)}, \text{Upsample}(\hat{\mathbf{X}}^{(6)}, 2)), \\ \mathbf{X}^{\text{fuse}} &= \text{ReLU}(\text{BN}(\text{Conv}^{3\times3}(\mathbf{X}^{\text{concat}}))), \end{aligned} \tag{10}$$

in which $\text{Upsample}(\cdot, 2)$ upsamples a feature map by two times, $\text{Conv}^{1\times1}$ is a 1×1 convolution with 64 output channels, and $\text{Conv}^{3\times3}$ is a 3×3 convolution with 256 output channels. $\mathbf{X}^{\text{fuse}}$ is used to predict the final lesion masks using a 1×1 convolution and upsampling by four times. During training, we also apply deep supervision [22] to $\mathbf{X}^{(4)}$, $\mathbf{X}^{(5)}$, and $\mathbf{X}^{(6)}$ for better optimization, as shown in Fig. 3.

5 Experiments

5.1 Experimental Setup

Implementation Details. The proposed method is implemented using the PyTorch framework [33]. We use ImageNet-pretrained VGG16 [39] or ResNet50 [15] as the backbone. We use $K = 24$ as the default setting. We initialize other convolution layers using the default setting of PyTorch. The Adam optimizer [21] is used for training. The learning rate policy is *poly*, in which the current learning rate equals the base one multiplying $(1 - curr\_iter/max\_iter)^{power}$, where the initial learning rate is set to 1e-4 and *power* is set to 0.9. The weight decay is 1e-4. We train our model for 50 epochs with a batch size of 16. In our experiments, we found that more than 50 epochs do not bring improvement for all models due to the large scale of our new SegLesion dataset. All experiments are conducted using a TITAN Xp GPU.

Table 3: Ablation studies for the hyper-parameter settings of KEN. "#Channels of decoder" means the numbers of channels for $\hat{\mathbf{X}}^{(3)}$ and $\hat{\mathbf{X}}^{(6)}$ in Eq. (10). The default settings are $K = 24$, $C = 512$, and #Channels $= (64, 256)$, respectively.

Configurations		Metrics (%)			
		mIoU ↑	SEN ↑	SPE ↑	DSC ↑
Default Configuration		65.78	51.09	97.48	**41.01**
K of KEM	16	65.47	51.31	97.81	39.61
	32	65.73	52.07	97.31	39.84
	48	65.63	46.50	96.97	38.40
C of KEM	128	65.90	48.78	98.54	39.36
	256	65.27	49.50	97.38	38.80
	1024	**66.31**	52.15	97.90	40.68
#Channels of decoder	(64, 128)	65.35	49.76	**98.63**	38.33
	(64, 512)	65.75	48.15	93.12	39.30
	(32, 128)	64.98	51.05	94.09	39.11
	(32, 256)	65.91	48.14	96.85	39.69
	(128, 256)	65.28	**52.83**	94.77	39.39

Table 4: Quantitative comparison between our KEN and 22 state-of-the-art segmentation methods.

Methods	Publication	Backbone	ImageNet	#Params	FLOPs	Speed	Metrics (%)			
							mIoU ↑	SEN ↑	SPE ↑	DSC ↑
U-Net	MICCAI'2015	-	No	33.72M	261.92G	27.86fps	61.62	39.92	82.72	27.76
FCN-8s	TPAMI'2017	VGG16	Yes	15.53M	105.97G	51.65fps	62.56	30.98	67.48	27.32
SegNet	TPAMI'2017	-	No	28.75M	160.44G	42.41fps	57.21	22.92	90.75	15.35
FRRN	CVPR'2017	-	No	17.30M	237.70G	17.41fps	62.74	37.80	81.19	28.41
PSPNet	CVPR'2017	ResNet50	Yes	64.03M	257.79G	26.52fps	64.67	31.59	72.72	27.93
DeepLabv3	CVPR'2017	ResNet50	Yes	38.71M	163.83G	20.59fps	66.06	42.11	86.34	36.36
DenseASPP	CVPR'2018	-	Yes	27.93M	122.28G	15.95fps	61.73	22.50	71.09	20.92
DFN	CVPR'2018	ResNet50	Yes	43.53M	81.88G	89.39fps	62.27	33.39	76.74	26.46
EncNet	CVPR'2018	ResNet50	Yes	51.25M	217.46G	23.21fps	64.45	33.73	68.28	29.19
DeepLabv3+	ECCV'2018	Xception	Yes	53.33M	82.87G	32.26fps	59.88	23.06	77.03	20.14
BiSeNet	ECCV'2018	ResNet18	Yes	12.50M	13.01G	335.50fps	60.52	20.14	68.51	17.15
UNet++	DLMIA'2018	-	No	35.77M	552.16G	10.56fps	62.52	34.10	74.34	26.61
Attention U-Net	arXiv'2018	-	No	34.06M	266.31G	24.25fps	61.69	35.47	81.11	25.95
OCNet	arXiv'2018	ResNet50	Yes	51.60M	220.69G	28.00fps	65.82	42.77	93.28	35.71
DUpsampling	CVPR'2019	ResNet50	Yes	28.46M	123.01G	31.26fps	65.99	34.82	66.69	30.28
DANet	CVPR'2019	ResNet50	Yes	64.87M	275.72G	24.53fps	65.18	32.94	64.93	28.21
CCNet	CVPR'2019	ResNet50	Yes	46.32M	197.92G	27.56fps	65.18	35.92	71.47	30.86
ANNNet	ICCV'2019	ResNet50	Yes	47.42M	203.07G	21.50fps	65.77	34.49	66.07	29.75
GFF	AAAI'2020	ResNet50	Yes	90.57M	374.03G	17.52fps	64.54	28.14	59.54	20.80
CPNet	CVPR'2020	ResNet50	Yes	48.59M	207.43G	22.58fps	64.59	30.24	62.90	27.55
OCRNet	ECCV'2020	ResNet50	Yes	37.94M	161.44G	28.07fps	65.70	37.31	82.61	32.89
DNL	ECCV'2020	ResNet50	Yes	46.51M	197.52G	26.07fps	64.36	35.38	69.83	30.13
KEN	-	VGG16	Yes	20.11M	148.99G	31.79fps	65.78	51.09	97.48	41.01
KEN	-	ResNet50	Yes	26.74M	139.36G	26.56fps	**66.64**	**58.48**	**97.82**	**42.06**

Evaluation Criteria. This paper adopts four popular segmentation metrics in medical image analysis, including mean intersection over union (mIoU), sensitivity (SEN), specificity (SPE), and dice similarity coefficient (DSC). The higher these metrics, the better the performance.

5.2 Ablation Studies

Here, we conduct ablation studies to evaluate the effect of various designs using the VGG16 backbone. We train the models on the SegLesion training set and

evaluate on the validation set. We start with the standard FCN (the 0^{th} column of Tab. 2), which is viewed as the baseline.

First, to prove the effect of $\mathbf{X}^{(6)}$, we add $\mathbf{X}^{(6)}$ to the baseline. The results are shown in the 1^{st} column of Tab. 2. Note that we do not use $\mathbf{X}^{(6)}$ for ResNet50 owing to its enough depth. Second, we add KEM which is the key component of our model to the above 1^{th} model to prove the effect of KEM and its designs. The results are put in the 2^{nd} column of Tab. 2. Third, as shown in the 3^{rd} column of Tab. 2, we prove the effect of the simple decoder of KEN, which fuses the features at 1/4 ($\mathbf{X}^{(3)}$) and 1/8 ($\mathbf{X}^{(6)}$) scales. Forth, we also show the effect of training with deep supervision in the 4^{th} column of Tab. 2. The experimental results of above ablation studies indicate the significant improvement brought by each component of KEN, especially by KEM. Fifth, KEM uses ER to enhance the intra-channel representation, as in Eq. (5) - Eq. (7). We provide the results of removing ER from KEM in the 5^{th} column of Tab. 2. Sixth, another question would be what will happen if we only use ER for knowledge embedding, *i.e.*replacing $\mathbf{D}''$ in Eq. (7) with the input $\mathbf{X}$. The results of this study are given in the 6^{th} column of Tab. 2. Finally, we try different decoders by replacing $\mathbf{X}^{(3)}$ with $\mathbf{X}^{(2)}$ (1/2 scale) or $\mathbf{X}^{(4)}$ (1/8 scale). The results of these two ablation studies are shown in 7^{th} and 8^{th} columns of Tab. 2, respectively. We can see that the default design of our model shows significant superiority over all above variants.

In Tab. 3, we evaluate the impact of the number of knowledge vectors K and the number of channels of the feature map C, which are the most important hyper-parameters of KEM. We also evaluate the impact of different numbers of channels of $\hat{\mathbf{X}}^{(3)}$ and $\hat{\mathbf{X}}^{(6)}$ in the decoder. We can see that KEN is robust to different parameter settings, and the default setting performs slightly better.

5.3 Performance Comparison

Quantitative Evaluation. On our SegLesion test set, we build a comprehensive benchmark for extensively comparing the proposed KEN with 22 state-of-the-art methods, including U-Net [36], FCN-8s [38], SegNet [5], FRRN [35], PSPNet [57], DeepLabv3 [9], DenseASPP [48], DFN [52], EncNet [56], DeepLabv3+ [10], BiSeNet [51], UNet++ [60], Attention U-Net [32], OCNet [55], DUpsampling [42], DANet [14], CCNet [18], ANNNet [61], GFF [24], CPNet [50], OCRNet [54], and DNL [49]. For a fair comparison, we use the code released by the authors. Besides the accuracy evaluation in terms of mIoU, SEN, SPC, and DSC, we also report the number of parameters, the number of FLOPs, and speed, where the default SegLesion resolution of 512×512 is adopted for testing.

The numeric comparison is summarized in Tab. 4. From this table, we can find that the proposed KEN significantly outperforms all competitors in terms of all metrics. The main reason may be that the SegLesion dataset exhibits a wide variety of sizes, shapes, and appearances, which are beyond the consideration of previous segmentation methods, resulting in the unsatisfactory performance of these methods for universal lesion segmentation. This also suggests that universal lesion segmentation is a new research field *worthy of study*. The ResNet50 version of KEN further boosts the performance compared to the VGG16 version.

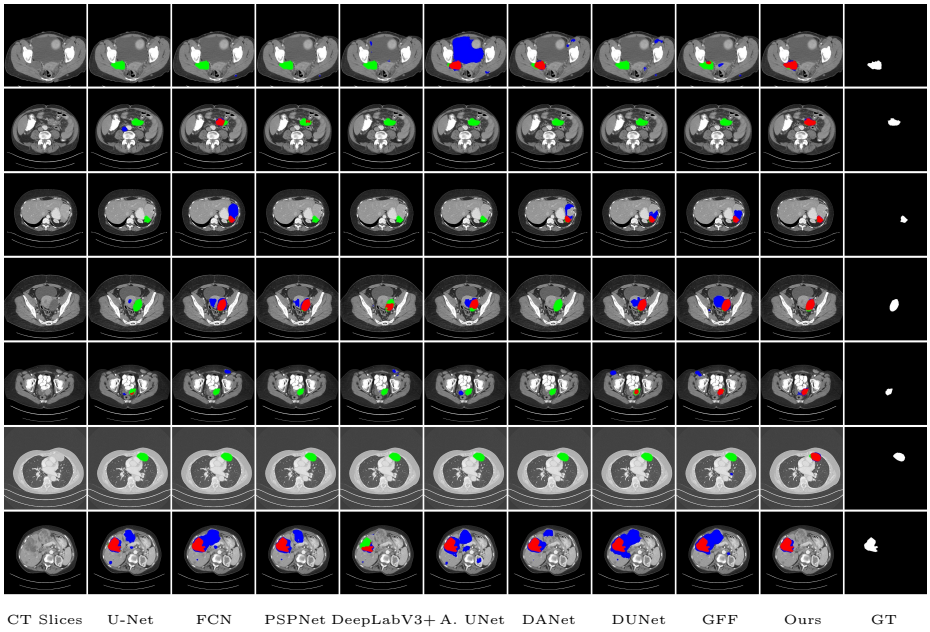


Fig. 4: Qualitative comparison between KEN and eight state-of-the-art competitors. GT: ground-truth lesion mask, A. UNet: Attention UNet. **Red**: true positive; **Green**: false negative; **Blue**: false positive.

Table 5: Lesion segmentation accuracy of KEN for different lesion types on the SegLesion test set. This paper focuses on universal lesion segmentation, and this is table is just shown for clarification. ME: Mediastinum; ST: Soft tissue.

Lesion Types	Metrics (%)			
	mIoU↑	SEN ↑	SPE ↑	DSC ↑
All Types	65.78	51.09	97.48	41.01
Lung	73.91	66.26	98.01	53.60
Abdomen	61.72	37.01	97.03	30.33
ME	66.88	59.20	98.95	45.93
Liver	67.74	57.81	98.59	46.11
Pelvis	64.39	35.85	96.40	28.76
ST	58.52	30.64	91.26	24.67
Kidney	58.12	36.42	98.29	29.15
Bone	63.92	41.84	96.47	35.19

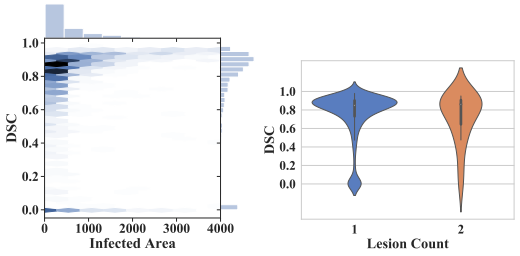


Fig. 5: Statistical analysis for KEN on the SegLesion test set. (a) The DSC score *vs.* the infected area; (b) The probability distribution of the DSC score *vs.* the lesion count in the corresponding CT slice.

The number of parameters of KEN is also favorably small, implying that its improvement mainly comes from our idea of knowledge embedding learning. KEN

is also one of the fastest methods because the proposed KEM is computationally flexible, and the computational load mainly comes from the backbone networks.

In Tab. 5, we provide the segmentation accuracy of KEN for different lesion types. KEN consistently achieves good performance for all lesion types. Specifically, KEN achieves the best performance for lung lesions and the worst performance for soft-tissue lesions. Note that previous lesion segmentation can only handle a single lesion type, which is the problem that our SegLesion resolves.

Note that this paper does not aim to solve/end a problem. Instead, our main contribution is to start a new task of universal lesion segmentation by proposing a large-scale dataset, a comprehensive benchmark, and a strong baseline method. Although the performance of our baseline is not enough for clinical applications, we believe lots of studies will appear to push this field to clinical deployment.

Qualitative Evaluation. To explicitly show the effectiveness of the proposed KEN, we select some representative CT slices and display the qualitative comparison between KEN and eight state-of-the-art methods in Fig. 4. We can see that CT slices are with lower contrast compared with nature images, and different lesion types exhibit a wide variety of sizes, shapes, and appearances, making universal lesion segmentation very challenging. Generally, KEN can successfully segment the lesions with fine details, leading to better results.

Statistical Analysis. To further study the stability of the proposed KEN, we perform statistical analysis on our SegLesion dataset. Fig. 5a shows the relationship between the most popular metric of DSC and the infected area. We can see that most CT images have DSC in the range of [0.6, 1.0]. Fig. 5b displays the relationship between DSC and the lesion count in a CT slice. The medium DSC is above 0.8, regardless of lesion counts. These analyses demonstrate the effectiveness of the proposed KEN in universal lesion segmentation.

6 Conclusion

Universal lesion segmentation is of vital importance but has not been well explored due to the lack of labeled data. Currently, there only exist some weakly-supervised methods. To bring this field to the fully-supervised era, we establish a large-scale universal lesion segmentation dataset. Motivated by traditional dictionary learning, we propose a knowledge embedding approach, *i.e.*, KEN, for universal lesion segmentation. We also build a large benchmark to compare KEN to previous methods extensively. KEN consistently outperforms other competitors and can thus serve as a strong baseline for future research. The limitation of this paper would be that the proposed dataset is still not large enough like natural image datasets [37,27], although it is the largest dataset for universal lesion segmentation now. In the future, we will continue our work by extending this dataset as large as possible.

Acknowledgement. This work is supported by Science and Technology Planning Project of Tianjin, China (Grant No. 20YDTPJC01810), Tianjin Natural Science Foundation, China (Grant No. 21JCYBJC00110 and 19JCQNJC00300).

References

1. Acuna, D., Ling, H., Kar, A., Fidler, S.: Efficient interactive annotation of segmentation datasets with Polygon-RNN++. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 859–868 (2018) 5
2. Agarwal, V., Tang, Y., Xiao, J., Summers, R.M.: Weakly-supervised lesion segmentation on CT scans using co-segmentation. In: Medical Imaging 2020: Computer-Aided Diagnosis. vol. 11314, p. 113141J (2020) 1, 2, 5
3. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. Medical Physics **38**(2), 915–931 (2011) 2, 4
4. Armato III, S.G., McLennan, G., McNitt-Gray, M.F., Meyer, C.R., Yankelevitz, D., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., et al.: Lung image database consortium: Developing a resource for the medical imaging research community. Radiology **232**(3), 739–748 (2004) 4
5. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **39**(12), 2481–2495 (2017) 2, 12
6. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al.: The liver tumor segmentation benchmark (LiTS). arXiv preprint arXiv:1901.04056 (2019) 3
7. Cai, J., Tang, Y., Lu, L., et al.: Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3D mask generation from 2D RECIST. In: Int. Conf. Med. Image Comp. Comput.-Assist. Interv. (MICCAI). pp. 396–404. Springer (2018) 1, 2, 5
8. Cao, H., Liu, H., Song, E., Hung, C.C., Ma, G., Xu, X., Jin, R., Lu, J.: Dual-branch residual network for lung nodule segmentation. Applied Soft Computing **86**, 105934 (2020) 2, 4
9. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 2, 12
10. Chen, L.C., Zhu, Y., Papandreou, G., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Eur. Conf. Comput. Vis. (ECCV). pp. 801–818 (2018) 1, 2, 4, 12
11. Christ, P.F., Ettlinger, F., Grün, F., Elshaera, M.E.A., Lipkova, J., Schlecht, S., Ahmaddy, F., Tatavarty, S., Bickel, M., Bilic, P., et al.: Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. arXiv preprint arXiv:1702.05970 (2017) 2, 4
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 3213–3223 (2016) 2
13. Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al.: New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer **45**(2), 228–247 (2009) 2, 4
14. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 3146–3154 (2019) 2, 12

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 770–778 (2016) 7, 9, 10
16. Heller, N., Sathianathen, N., et al.: The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019) 2, 3
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 7132–7141 (2018) 8
18. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: Criss-cross attention for semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 603–612 (2019) 9, 12
19. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Int. Conf. Mach. Learn. (ICML). pp. 448–456 (2015) 9
20. Katzmann, A., Muehlberg, A., Suehling, M., Noerenberg, D., Holch, J.W., Heinemann, V., Gross, H.M.: Predicting lesion growth and patient survival in colorectal cancer patients using deep neural networks. In: Microsoft Interface Definition Language (2018) 1
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Int. Conf. Learn. Represent. (ICLR) (2015) 10
22. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artif. Intell. Stat. (AISTATS). pp. 562–570 (2015) 10
23. Lee, R.S., Gimenez, F., Hoogi, A., et al.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data **4**, 170177 (2017) 2, 4
24. Li, X., Zhao, H., Han, L., Tong, Y., Tan, S., Yang, K.: Gated fully fusion for semantic segmentation. In: AAAI Conf. Artif. Intell. (AAAI). pp. 11418–11425 (2020) 12
25. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging (TMI) **37**(12), 2663–2674 (2018) 2
26. Lian, C., Ruan, S., Denœux, T., Li, H., Vera, P.: Joint tumor segmentation in PET-CT images using co-clustering and fusion based on belief functions. IEEE Trans. Image Process. (TIP) **28**(2), 755–766 (2019) 1
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Eur. Conf. Comput. Vis. (ECCV). pp. 740–755 (2014) 14
28. Lu, K., Bascom, R., Mahraj, R.P., Higgins, W.E.: Quantitative analysis of the central-chest lymph nodes based on 3D MDCT image data. In: Medical Imaging 2009: Computer-Aided Diagnosis. vol. 7260, p. 72600U (2009) 1
29. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: INbreast: Toward a full-field digital mammographic database. Academic Radiology **19**(2), 236–248 (2012) 2, 3
30. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Int. Conf. Mach. Learn. (ICML). pp. 807–814 (2010) 9
31. Nikan, S., Van Osch, K., Bartling, M., Allen, D.G., Rohani, S.A., Connors, B., Agrawal, S.K., Ladak, H.M.: PWD-3DNet: A deep learning-based fully-automated segmentation of multiple structures on temporal bone CT scans. IEEE Trans. Image Process. (TIP) **30**, 739–753 (2020) 1
32. Oktay, O., Schlemper, J., Folgoc, L.L., et al.: Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018) 2, 4, 12

33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library. In: Annu. Conf. Neur. Inform. Process. Syst. (NeurIPS). pp. 8026–8037 (2019) 10
34. Peng, Y., Yan, K., Sandfort, V., Summers, R.M., Lu, Z.: A self-attention based deep learning method for lesion attribute detection from CT reports. In: IEEE International Conference on Healthcare Informatics. pp. 1–5 (2019) 1
35. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 4151–4160 (2017) 12
36. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Int. Conf. Med. Image Comp. Comput.-Assist. Interv. (MICCAI). pp. 234–241. Springer (2015) 2, 4, 12
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015) 14
38. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **39**(4), 640–651 (2017) 1, 2, 4, 12
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Int. Conf. Learn. Represent. (ICLR). pp. 1–14 (2015) 7, 9, 10
40. Soler, L., Hostettler, A., Agnus, V., et al.: 3D image reconstruction for comparison of algorithm database: A patient-specific anatomical and medical image database. IRCAD, Strasbourg, France, Tech. Rep (2010) 2, 3
41. Tang, Y., Cai, J., Lu, L., et al.: CT image enhancement using stacked generative adversarial networks and transfer learning for lesion segmentation improvement. In: International Workshop on Machine Learning in Medical Imaging. pp. 46–54 (2018) 1, 2, 5
42. Tian, Z., He, T., Shen, C., Yan, Y.: Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 3126–3135 (2019) 1, 2, 4, 12
43. Wang, Z., Wei, L., Wang, L., Gao, Y., Chen, W., Shen, D.: Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning. IEEE Trans. Image Process. (TIP) **27**(2), 923–937 (2018) 1
44. Yan, K., Lu, L., Summers, R.M.: Unsupervised body part regression via spatially self-ordering convolutional neural networks. In: IEEE Int. Symp. Biomed. Imaging (ISBI). pp. 1022–1025 (2018) 5
45. Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M.: MULAN: Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In: Int. Conf. Med. Image Comp. Comput.-Assist. Interv. (MICCAI). pp. 194–202. Springer (2019) 1
46. Yan, K., Wang, X., Lu, L., Summers, R.M.: DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging **5**(3), 036501 (2018) 2, 4, 5, 6
47. Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A.P., Bagheri, M., Summers, R.M.: Deep lesion graphs in the wild: Relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 9261–9270 (2018) 5

48. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: DenseASPP for semantic segmentation in street scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 3684–3692 (2018) 1, 2, 4, 9, 12
49. Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., Hu, H.: Disentangled non-local neural networks. In: Eur. Conf. Comput. Vis. (ECCV). pp. 191–207 (2020) 12
50. Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., Sang, N.: Context prior for scene segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 12416–12425 (2020) 12
51. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In: Eur. Conf. Comput. Vis. (ECCV). pp. 325–341 (2018) 2, 12
52. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 1857–1866 (2018) 9, 12
53. Yu, Q., Shi, Y., Sun, J., Gao, Y., Zhu, J., Dai, Y.: Crossbar-Net: A novel convolutional neural network for kidney tumor segmentation in CT images. IEEE Trans. Image Process. (TIP) **28**(8), 4060–4074 (2019) 1
54. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Eur. Conf. Comput. Vis. (ECCV). pp. 173–190 (2020) 12
55. Yuan, Y., Wang, J.: OCNet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916 (2018) 1, 2, 4, 9, 12
56. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 7151–7160 (2018) 2, 9, 12
57. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 2881–2890 (2017) 2, 9, 12
58. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. Int. J. Comput. Vis. (IJCV) **127**(3), 302–321 (2019) 2
59. Zhou, S., Nie, D., Adeli, E., Yin, J., Lian, J., Shen, D.: High-resolution encoder–decoder networks for low-contrast medical image segmentation. IEEE Trans. Image Process. (TIP) **29**, 461–475 (2020) 1
60. Zhou, Z., Siddiquee, M.M.R., et al.: UNet++: A nested U-Net architecture for medical image segmentation. In: DLMIA and ML-CDS, pp. 3–11. Springer (2018) 2, 4, 12
61. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: Int. Conf. Comput. Vis. (ICCV). pp. 593–602 (2019) 9, 12