IET Journals
The Institution of Engineering and Technology

Multiscale Matters for Part Segmentation of Instruments in Robotic Surgery

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Wenhao He[1], *Haitao Song*[1], *Yue Guo*[1*], *Guibin Bian*[1], *Yuejie Sun*[2], *Xiaowei Zhou*[1], *Xiaonan Wang*[1]

[1] *Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[2] *Peking University International Hospital, Beijing, China*
* *E-mail: guoyue2013@ia.ac.cn*

Abstract: A challenging aspect of instrument segmentation in robotic surgery is to distinguish different parts of the same instrument. Parts with similar textures are common in a practical instrument and are difficult to distinguish. In this work, we introduce an end-to-end recurrent model that comprises a multiscale semantic segmentation network and a refinement model. Specifically, the semantic segmentation network uniformly transforms the input images in multiple scales into a semantic mask, the refinement model is a single-scale net recurrently optimizing the above semantic mask. Through extensive experiments, we validate that the models with multiscale inputs perform better than those to fuse encoding and decoding feature maps with spatial attention. Furthermore, we verify the effectiveness of our model with state-of-the-art performances on several robotic instrument datasets derived from MICCAI Endoscopic Vision Challenges.

1 Introduction

The research community of robotic instrument segmentation has payed growing interest in current clinical practice [1–7] for its crucial importance in many tasks involved in robot-assisted minimally invasive surgery and computer-assisted surgical systems. Particularly, accurate segmentation of instruments is a fundamental step towards scene understanding in many surgical and robotic operations [8–12].

Minimally invasive surgery has been applied in a variety of surgical procedures because it causes less damage to tissues and reduces patient suffering. However, minimally invasive surgery requires high-level operation techniques and is difficult to master, which limits the widespread use of this technology. Moreover, surgeons incline to feel fatigue after long-term continuous surgical operations, and their hands tend to be unstable to ensure accurate manipulations.

Robot-assisted minimally invasive surgery addresses the above problems because locations of robotic instruments are capable of quickly reaching the target without shaking. However, potential factors such as shadows, specular reflections, partial occlusions, blood splattering, and tissue dynamics lead to the complicated surgical environment. Moreover, compact size of robotic instruments and their complex actuation mechanisms cause trajectory control and scene understanding in surgical operations more challenging.

Recent advances in semantic segmentation are driven by the success of convolutional neural networks [13–39]. Some of them classify pixels using a single semantic segmentation model with pre-trained weights [13–16, 18, 19, 21, 23, 25, 26, 29, 32, 33, 36]. Some other methods refine pixels using multiple convolutional neural networks [18, 20, 22, 24, 27, 28, 31, 34]. Still other methods post-process pixels with additional modules based on low-level boundary cues [30, 35]. In spite of their success, these methods do not refine semantic masks in a supervised way. Designed without multiscale recurrent refinement, previous methods often struggle in separating parts with similar contexts, as shown in Figure 2 (M_T is closer to the true mask than M_0).

This work introduces a multiscale recurrent refinement model for part segmentation of instruments in robotic surgery, which alleviates the above problem with a collaborative rectification mechanism. The model comprises two parts: the semantic segmentation network and the recurrent refinement network. Given an input image, the semantic segmentation network transforms the images in multiple scales to a mask with continuous semantic regions. The segmentation network is a multiscale Linknet [13], a quite efficient encoder-decoder model that can remove spotted semantic regions.

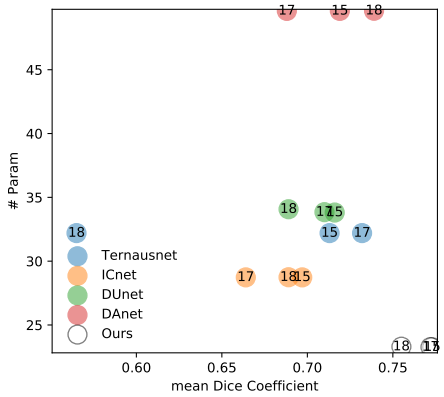

Fig. 1: Number of parameters v.s. mean Dice coefficient: Every circle represents the performance of a method, and our model outperforms others by a large margin with the fewest parameters. '15', '17', and '18' respectively denote the dataset from Endovis15, Endovis17, and Endovis18.

During inference, the semantic segmentation network simultaneously predicts masks from converted images in multiple resolutions, then all the outputs are fused into a mask. Based on Linknet [13] and Siamese network [40], the semantic segmentation model is directly trained end-to-end to capture multiscale contexts.

The recurrent refinement model predicts a more detailed mask collaboratively. Built on a basic Linknet, the recurrent refinement model effectively simplifies multiple semantic segmentation models into a single model, while achieving accurate robotic instrument segmentation. Furthermore, we merge predictions from both the semantic segmentation model and the recurrent refinement one, leveraging dependencies of masks from different paths.

Through an extensive set of experiments on several medical instrument datasets, we demonstrate superior part segmentation performances of our method on both rigid instruments and robotic ones. Furthermore, when used with other semantic segmentation models, the recurrent refinement model shows the ability to improve

the segmentation results by refining their spotted semantic regions. For the end-to-end designation, our semantic segmentation systems optimize as a whole.

In summary, the contributions of this paper are three-folded. Firstly, a multiscale Linknet is constructed to significantly reduce spotted regions when segmenting semantic parts of medical instruments. Secondly, a recurrent refinement model is introduced to optimize coarse boundaries in masks from the semantic segmentation model. Thirdly, to demonstrate the effectiveness of multiscale and recurrent refinement, widely used spatial attention modules are integrated into our framework. All the variants and state-of-the-art semantic segmentation models are evaluated and compared, which systematically demonstrates the superiority of our method.

2 Related Work

Encoder-decoder networks and fully convolutional ones have become the dominant architectures for semantic segmentation. In this work, we discuss their variants that exploit context information for semantic segmentation.

2.1 Encoders

Encoders with limited receptive fields or small kernels may not be suitable for representing semantic pixels. Therefore, convolution operations in a fully convolutional neural network are replaced with dilated convolutions [7] or deformable ones [41], and large kernels in a global convolutional network simultaneously localize and classify pixels [29, 42, 43]. However, encoders with small kernels might be faster for dense pixel classification.

Semantic features seem to generalize better using weights of pretrained models such as VGG-11 [5], ResNet-50 [34], ResNet-101 [7, 32], WideResNet [32], Xception [15], and DenseNet-169 [25]. However, these networks heavily rely on the pretrained weights, so high-level features in too deep models may not be easily transfered for segmenting targets in other tasks.

To obtain more informative representations for semantic segmentation on several occasions, features in some layers are aggregated in specified feature dimensions. Specifically, concurrent spatial and channel "squeeze and excitation" blocks are integrated into a fully convolutional network to segment brains and organs [44]. Moreover, the position-attention module and channel-attention one are designed to refine segmentation outputs [19].

To fuse multi-scale features, typically, pyramid pooling blocks are nested before decoding intermediate feature maps. For example, a pyramid parsing module followed by upsampling and concatenation layers is initially proposed [34]. After that, an atrous spatial pyramid pooling layer consists of atrous convolution layers with different rates in parallel [14]. Based on the above atrous-convolutional structure, region features are merged with boundary ones [32]. Besides, global structures of vessels are captured with a graph neural network, while local appearances on an image grid are learned with a convolutional neural network [45].

Moreover, convolutional networks can be extended and applied in three-dimensional semantic segmentation occasions including surgical tools [3, 46]. Specifically, to solve this issue, a three-dimensional fully convolutional network is constructed [3], dense paths are created inside the above model given early-fused inputs [47], and bidirectional convolutional long-short term memories are built to capture spatial-temporal correlations of the continuous slice of vertebrae and livers in three-dimensional CT scans [48].

Additionally, structure priors in images can be efficiently fused using multi-modal inputs. For example, spatial ranking maps are supervised by panoptic segmentation labels to alleviate overlapping problems among various classes [42], high-resolution feature maps are refined with inputs from multiple paths [49], streams respectively given stacked optical flows and color images models motions and appreances [50], and gated convolutional layers enforce boundary information merely processed in a shape stream [32]; images in axial, coronal, and sagittal views are independently segmented and fused into a single result with union operations [43].

2.2 Decoders

Feature maps from encoders in a fully convolutional network is directly fused to output semantic pixels, but they can be gradually decoded in different levels to capture sharper boundaries [15]. Initially, feature maps of decoders are directly concatenated with those of encoders in a U-net [51]. Recently, additional modules are linked between each encoding layer and the corresponding decoding counterpart. For example, residual attentions are built to localize the liver in every image and segment it in the volumetric space [52]; dense skip pathways are created with multiple nested convolutional blocks to bridge the semantic gap [53]; a U-net can also consist of filters with different kernel sizes and residual connections [54]; cascade branches with multiple atrous convolution blocks and residual-kernel poolings are used to connect an encoder and a decoder [55]; convolutional long-short term memories can be temporally connected in an decoder [56]. Different from the aforementioned works, Linknet [13] is constructed as the baseline network and attention gates [57] are integrated in decoders by leveraging spatial relationships between encoded features and previous decoded ones.

Feature maps from encoders can also be fused before progressive decoding. For example, multi-scale contexts are captured with spatial pyramid pooling [6], atrous spatial pyramid pooling [15] or joint pyramid upsampling [58]. However, features of encoders may be insufficient for training semantic segmentation models with small dataset, so data generation is an essential step to improve segmentation performances. For instance, motion vectors are estimated with a three-dimensional convolutional neural network, and to scale up training data, given previous frames and motion vectors, bilinear operations are applied to predict future frames [59]; part of visual cues are used to regenerate full objects which are later combined with local inputs filled with margins to predict contexts [60].

2.3 Losses

Pixel-wise losses can be directly applied for training segmentation models, and to balance binary or multiple classes, dice coefficient [46] and class-balancing cross-entropy loss [61] are used to train models. In our work, different losses are integrated into a unified loss function to supervise the segmentation task.

Perceptional losses may preserve more global structures compared to pixel-wise ones. Inspired by generative adversarial networks, generated masks are usually judged with a discriminator [62, 63]. For example, ground truth masks and predicted ones are separated with a classifier in image level [6], patch level [43], or both [60]. Moreover, their consistencies can be supervised with earth-mover distances in the Wasserstein generative adversarial networks [64, 65]. To enlarge dataset while preserving geometric information in synthetic data, depth images are auxiliary inputs of a generative adversarial network generating color images in real-world domain [17].

Except for distances between predictions and ground truths directly measured, they can also be indirectly computed. For example, the reconstruction error between original labels and compressed ones by transformations applied in data-dependent upsampling is minimized [33]; based on teacher-student networks, the dense gram matrix of feature maps in the source network should be close to that in the target model [20].

2.4 Ensembles

Global structures and local details in a high-resolution image may not be simultaneously captured well with a semantic segmentation model. Therefore, different information flows should be ensembled to address this problem.

Global information and local details can be independently processed. For instance, regions of pancreas are localized with a deep Q network and then segmented with a deformable U-net [41]; shape details of target regions are recovered with saliency transformation modules and refined recurrently with a finer-scaled network [66];

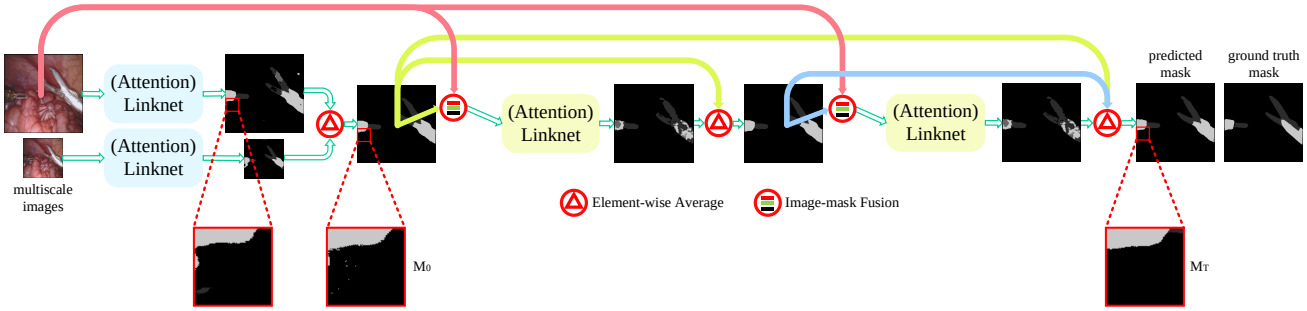

Fig. 2: An overview of the semantic segmentation framework: the proposed method consists of two modules: the multi-scale Linknet and the recurrent Linknet. Specifically, images in multiple scales are obtained by resizing the original high-resolution image at first; then every resized image is independently segmented with a multiscale Linknet, and all the outputs are averaged into a single map which is later combined with the original image; in the end, the combined features are collaboratively refined with a recurrent Linknet and updated in each iteration. It should be noted that all the relevant images are resized into the same size before the element-wise averaging.

cell-level architecture search spaces and network-level ones are used to optimize the semantic segmentation architecture [27].

Images [67], intermediate feature maps [26], or predictions at different scales [4] can be fused to refine segmentation masks. Without changing model inputs and outputs, joint learning might be another direction for segmentation refinement. Specifically, refinement and classification of bounding boxes combined with segmentation can be jointly and cascade learned [16, 21, 31].

2.5 *Motivation*

In this paper, we mainly focus on the multi-stage semantic segmentation, and masks can be progressively refined using cascade structures. For example, feature maps belonging to support images and query ones are refined with an iterative optimization module [68]; intermediate feature maps in different scales are fused with cascade feature fusion modules [67].

To reduce the number of parameters in a cascade architecture, inspired by recurrent U-net [69], the current image are recurrently combined with previous segmentation masks, and weights among basic networks are shared in our method, even though weights are better not shared in stacked hourglass networks [70].

3 Method

3.1 *Linknet*

Linknet [13] is used as the baseline of our approach due to the following two reasons: on the one hand, since the accuracy and efficiency of pixel-wise semantic segmentation are leveraged in Linknet, it is quite suitable for real-time robotic instrument segmentation; on the other hand, test speeds tend to decrease after refining mask recurrently, so a fast segmentation model must be chosen to ensure the overall efficiency of our framework.

Linknet is based on the encoder-decoder structure, relatively higher-level feature maps from the decoder layer and lower-level ones from the corresponding encoder layer are channel-wisely summed to output the feature maps for the decoder layer in a next level. A pretrained 18-layer residual network is the backbone of encoding layers, and its first convolutional block consists of a convolutional layer with 64 filters with 7×7 kernel sizes and 2×2 strides, a batch normalization layer, and a max-pooling layer with stride 2. Therefore, areas of feature maps from this block becomes $\frac{1}{16}$ as those of inputs, which accelerates feature maps computed with higher-level convolutional layers. Transposed convolutions enlarge the sizes of feature maps in the decoder module instead of bilinear or nearest interpolation. Batch normalizations and ReLUs are integrated after convolutional blocks and transposed convolutional blocks.

3.2 *Attention Linknet*

The main topic in this paper is to what extent multiscale influences semantic segmentation of parts, while spatial attention [57] is a popular module to improve semantic segmentation performances. Therefore, spatial attention is integrated into the basic Linknet for comparisons. Specifically, channel-wise sums are used to connect encoding layers and decoding ones in the basic Linknet and attention gates are integrated into the attention counterpart. Feature maps containing unrelated background regions can be progressively suppressed using attention gates.

Given feature maps from the encoding layer and those from the decoding layer, the attention module connects them and output decoding feature maps attened by the encoding features.

$$\begin{aligned} F_o &= F_d \times \alpha \\ &= F_d \times h_o\Big(\sigma\big(h_d(F_d) + h_e(F_e)\big)\Big) \end{aligned} \tag{1}$$

where F_e, F_d, F_o are respectively feature maps from the encoding layer, those from the corresponding decoding layer, and outputs of the attention gate. α has only one channel, and its size is the same as that of F_d. $h_{d/e}(\cdot)$ is the convolutional block including a convolutional layer and a batch normalization layer, but h_o has an additional sigmoid layer except for the above two layers. σ is an activation function, and ReLU is used in this model.

3.3 *Multiscale Linknet*

Image inputs in multiple resolutions are adopted to concurrently capture multiscale contexts. Specifically, images in each resolution are indpendently passed through a weight-shared basic Linknet, and the output probability maps are averaged because both local textures and global structures are important for the final decisions:

$$M_0 = \frac{1}{p}\sum_{k=1}^{p} u_n\Big(f_I\big(u_b(I, \frac{1}{s_k})\big), s_k\Big) \tag{2}$$

where I and M_0 respectively represents an image and probability maps, f_I is the multiscale Linknet, u_n and u_b are upsampling operations with nearest interpolation and those with bilinear interpolation, and s_k is the k-th upsampling scale.

3.4 *Recurrent Refinement*

Probability maps from the multiscale Linknet and images are concatenated as the input of a recurrent Linknet. Specifically, at each iteration, inputs of the recurrent model are the original image and

probability maps from previous iterations, and its outputs are used to update probability maps at the next iteration.

$$M_t = \begin{cases} M_0, & t = 0 \\ \frac{1}{t+1}\Big(\sum_{i=0}^{t-1} M_i + f_M\big(g(I, M_{t-1})\big)\Big), & t > 0 \end{cases} \quad (3)$$

where M_t represents probability maps at t-th iteration, f_M is the recurrent Linknet, and g is the concatenation operator. All the probability maps are averaged channel-wisely.

3.5 Training

To train the end-to-end recurrent semantic segmentation model, focal loss [71] and dice one are applied. Specifically, the overall loss for a single input in each class is defined as L, and summerization of L in all the classes is a complete loss used for training.

$$L = L_{fcl} + L_{dce} \quad (4)$$

where L_{fcl}, L_{ce}, and f_{dce} respectively represents focal loss, cross-entropy loss, and dice loss, and focal loss is used to balance the weights of easily-classified examples and hard-classified ones.

$$L_{ce} = -\sum_{j=1}^{m}\sum_{k=1}^{m} M_{gt}(j,k) log(M_T(j,k)) \quad (5)$$

where m is the side length of the square probability map M_T at the last iteration T.

$$L_{dce} = 1 - \frac{\sum_{j=1}^{m}\sum_{k=1}^{m} M_{gt}(j,k) M_{pd}(j,k) + \epsilon}{\sum_{j=1}^{m}\sum_{k=1}^{m} M_{gt}(j,k) + M_{pd}(j,k) + \epsilon} \quad (6)$$

where M_{gt} and M_{pd} are respectively the groundtruth probability map and the predicted one after a softmax function: $Softmax(M_{Ti}) = \frac{e^{M_{Ti}}}{\sum_{j=0}^{nc-1} e^{M_{Tj}}}$, and the numerical issue of the loss function divided by 0 is solved by using ϵ.

4 Experiments

4.1 Datasets

Datasets in this paper are from endovis-robotic instrument segmentation subchallenges including Endovis15 ("Instrument Segmentation and Tracking" in MICCAI 2015), Endovis17 ("Robotic Instrument Segmentation" in MICCAI 2017) [1], and Endovis18 ("Robotic Scene Segmentation" in MICCAI 2018) [2]. It should be noted that only the part segmentation results are used for evaluations.

4.1.1 Endovis15: Rigid instruments and robotic counterparts are respectively recorded with images and videos in this dataset. Training data and test ones are already separated, but gaps between the distribution of training data containing robotic instruments and that of test one are relatively small. Therefore, rigid instrument segmentation is chosen and experimented. Specifically, the selected dataset contains 160 images from four surgeries, and the size of every image is 640×480. Parts of rigid instruments only include shaft and manipulator.

4.1.2 Endovis17: This dataset consists of eight videos captured with stereo cameras from a da Vinci Xi robot, and there are 225 frames in every video, and they are recorded at 30 Hz in the original video and are sampled at two Hz to remove video redundancy. To obtain 1280×1024 frames, original 1960×1280 frames should be cropped starting from the pixel $(320, 28)$. Components of instruments include rigid shafts, articulated wrists, and claspers. The same as dataset splitting in prior works [6], the first six videos and the last two ones are respectively used as training data and test one.

4.1.3 Endovis18: This dataset includes 16 videos from stereo cameras using da Vinci Xi systems, and there are 149 frames in each sequence, the original videos are sampled at 60 Hz and are downsampled at two Hz to accelerate the dense pixel labeling. Different from the above two datasets, both instruments and context objects are labeled. As a result, classes in this dataset includes background-tissue, instrument-shaft, instrument-clasper, instrument-wrist, kidney-parenchyma, covered-kidney, thread, clamps, suturing-needle, suction-instrument, and small-intestine. Similar to the above dataset splitting rules, these datasets are divided into the first 12 sequences for training and the last three ones for test.

4.2 Evaluation Protocol

To quantitatively evaluate the segmentation performance of every image, several metrics are used for each class to measure the similarity between a predicted mask and a ground truth one. Specifically, to obtain the predicted mask, the segmentation model firstly outputs a multi-channel probability map, then it is transferred to a single-channel mask where each pixel corresponds to the category index with the maximum probability at the same location among all the channels in the previous multi-channel map, finally, the provided mask is transferred back to a one-hot map as the predicted mask.

Metrics including dice coefficient, intersection over union, precision, recall, acurracy, and F1 score are used for evaluation. The specific metrics are defined as follows:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (7)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

where TP, FP, TN, and FN respectively represent the number of true positives, the number of false positives, the number of true negatives, and the number of false negatives.

To measure the overall performance using each metric, similarities for all the images in a test set are computed with the corresponding metric and are averaged. As a result, mean dice coefficient $mDice$, mean intersection over union $mIoU$, mean precision $mPrec$, mean recall $mRec$, mean accuracy $mAcc$, and mean F1 score $mF1$ are provided.

4.3 Implementation Details

To prepare high-resolution images as inputs of the model, all the cropped frames are resized to 640×640 using bicubic interpolation over 4×4 pixel neighborhood. Sizes of the resized images later become 320×320 to construct an image pyramid. Data Augmentation on color images such as randomly converting to gray ones and changing their brightness and contrast are applied, specifically, the probability of converting is 0.1, and both the factor of brightness and that of contrast are 0.01.

RAdam [72] is used to optimize the model, and its base learning rate is 0.0005. Batch size for each training epoch is 4. All the inferences are run on an 8-core Alienware Laptop with an Intel CPU of 2.80 GHz and 15-GB RAM, and all the baseline models are executed with a GeForce GTX1070 GPU. It should be noted that the same implementation details are followed among all the baselines and our methods except their various architectures.

Table 1 Comparisons of Part Segmentation with Different Ensembles on Endovis15

Attention	Multiscale	Image	Mask	mDice ↑	mIoU ↑	mPrec ↑	mRec ↑	mF1 ↑	mAcc ↑
				0.743	0.708	0.810	0.741	0.743	0.962
		✓		0.757	0.716	0.809	0.757	0.756	0.962
			✓	0.715	0.667	0.750	0.740	0.714	0.955
		✓	✓	0.729	0.688	0.769	0.759	0.727	0.956
	✓			**0.765**$^{\circ}$	**0.728***	**0.817***	**0.767**$^{\circ}$	**0.764**$^{\circ}$	**0.964**$^{\circ}$
	✓	✓		**0.773***	**0.733***	**0.816***	**0.782***	**0.772***	**0.965***
	✓		✓	0.711	0.667	0.789	0.701	0.710	0.959
	✓	✓	✓	0.755	0.715	0.806	0.764	0.754	**0.965***
✓				0.736	0.699	0.789	0.747	0.735	0.960
✓		✓		0.741	0.702	0.792	0.751	0.740	0.961
✓			✓	0.510	0.476	0.578	0.523	0.509	0.928
✓		✓	✓	0.738	0.695	0.795	0.733	0.736	0.961
✓	✓			**0.767***	**0.723**$^{\circ}$	**0.811**$^{\circ}$	**0.771***	**0.766***	0.964
✓	✓	✓		0.747	0.705	0.786	0.766	0.746	0.960
✓	✓		✓	0.651	0.607	0.686	0.694	0.650	0.942
✓	✓	✓	✓	0.739	0.700	0.804	0.738	0.738	0.961

Table 2 Comparisons of Part Segmentation with Different Ensembles on Endovis17

Attention	Multiscale	Image	Mask	mDice ↑	mIoU ↑	mPrec ↑	mRec ↑	mF1 ↑	mAcc ↑
				0.734	0.674	0.763	0.739	0.724	0.959
		✓		0.752	**0.694***	0.781	0.747	0.743	**0.962***
			✓	0.738	0.666	0.775	0.717	0.725	0.960
		✓	✓	0.741	0.675	0.775	0.744	0.735	0.960
	✓			0.738	0.671	0.770	0.743	0.728	0.955
	✓	✓		0.746	0.680	0.769	0.754	0.735	0.959
	✓		✓	0.701	0.638	0.729	0.717	0.695	0.942
	✓	✓	✓	**0.772***	**0.705***	0.782	**0.793***	**0.766***	**0.966***
✓				0.748	0.676	**0.792**$^{\circ}$	0.731	0.738	0.954
✓		✓		0.758	**0.693**$^{\circ}$	0.778	**0.763***	0.747	0.960
✓			✓	0.719	0.676	0.742	0.747	0.716	0.957
✓		✓	✓	0.707	0.645	0.731	0.716	0.698	0.961
✓	✓			0.743	0.681	**0.807***	0.716	0.733	**0.962**$^{\circ}$
✓	✓	✓		**0.760**$^{\circ}$	0.686	0.784	**0.761**$^{\circ}$	**0.748**$^{\circ}$	0.961
✓	✓		✓	0.679	0.621	0.665	0.740	0.673	0.952
✓	✓	✓	✓	**0.769***	0.691	**0.794***	0.756	**0.753***	0.962

Table 3 Comparisons of Part Segmentation with Different Ensembles on Endovis18

Attention	Multiscale	Image	Mask	mDice ↑	mIoU ↑	mPrec ↑	mRec ↑	mF1 ↑	mAcc ↑
				0.604	0.587	0.515	0.513	0.490	0.763
		✓		0.684	**0.633***	0.588	0.591	0.566	0.789
			✓	0.687	0.589	0.569	0.572	0.545	0.773
		✓	✓	0.603	0.587	0.519	0.520	0.493	0.748
	✓			0.643	0.607	0.550	0.548	0.524	0.788
	✓	✓		**0.755***	0.602	**0.637***	**0.638***	**0.609***	0.766
	✓		✓	**0.736***	0.573	**0.622***	0.603	**0.579**$^{\circ}$	0.797
	✓	✓	✓	0.670	**0.646***	0.584	0.578	0.559	**0.813***
✓				0.628	0.582	0.536	0.530	0.504	0.765
✓		✓		**0.729**$^{\circ}$	0.561	0.604	0.597	0.570	0.744
✓			✓	0.691	0.516	0.581	0.521	0.505	0.687
✓		✓	✓	0.716	0.560	0.616	0.602	0.566	0.745
✓	✓			0.728	0.617	**0.621**$^{\circ}$	**0.619***	**0.595***	**0.801***
✓	✓	✓		0.696	**0.625**$^{\circ}$	0.597	**0.607**$^{\circ}$	0.575	**0.799**$^{\circ}$
✓	✓		✓	0.624	0.566	0.519	0.537	0.501	0.768
✓	✓	✓	✓	0.666	0.577	0.559	0.564	0.531	0.754

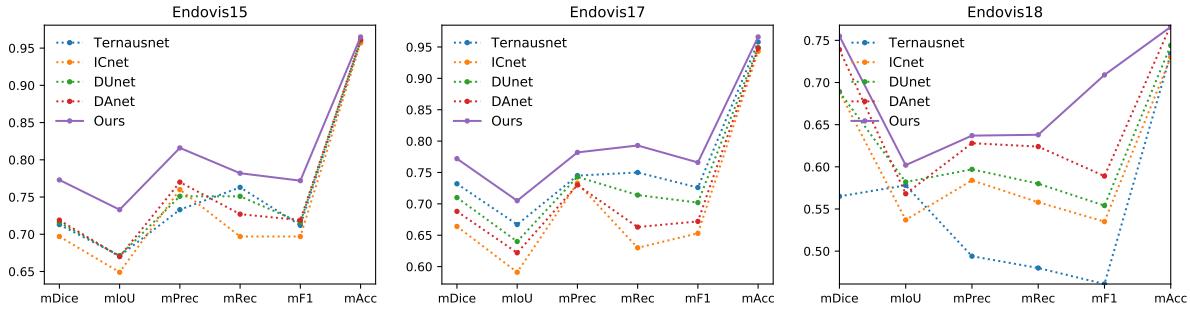

Fig. 3: Comparisons of part segmentation with state-of-the-arts on datasets from Endovis15, Endovis17, and Endovis18: we compare our method with semantic segmentation models including Ternausnet [73], ICnet [67], DUnet [33], and DAnet [19], and we evaluate all the methods on five different metrics: mDice, mIoU, mPrec, mRec, mF1, and mAcc. It can be easily seen that our method consistently outperforms others on the three datasets.

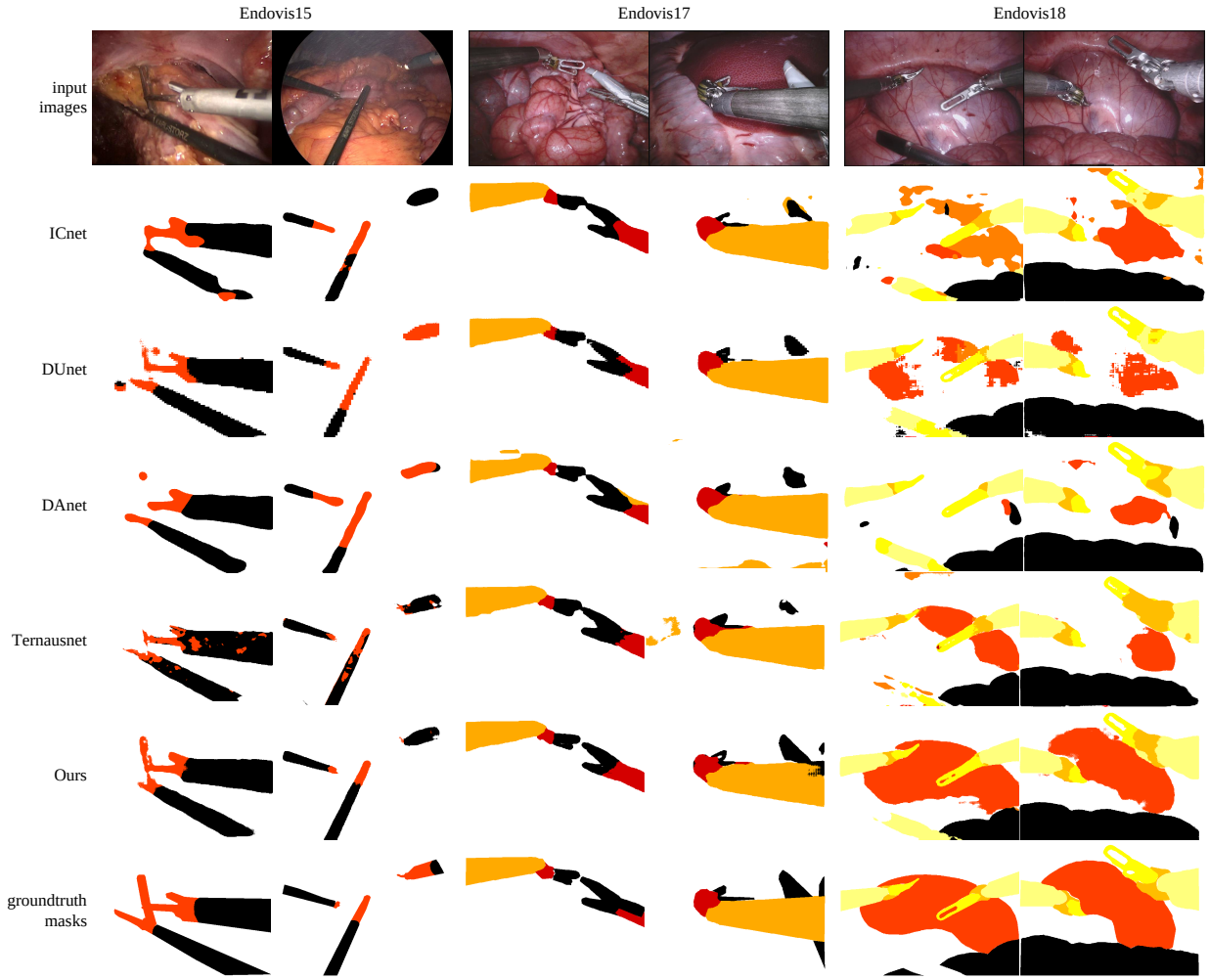

Fig. 4: Segmentation masks of baselines and our method: three columns respectively corresponds to results in Endovis15, Endovis17, and Endovis18, and every color corresponds to a specific semantic class. Most of the baselines can be used to accurately segment instruments, but sometimes our method outperforms baselines by a large margin on surgical contexts.

4.4 Ablation Studies

To search for the best architectures on test sets in Endovis15, Endovis17, and Endovis18, experimental results are respectively listed in Table 1, Table 2, and Table 3. To simplify descriptions about experimental comparisons, results are analyzed according to the dice coefficient. Besides, $[\cdot]^*$, $[\cdot]^\star$, and $[\cdot]^\circ$ correspond to models which performance ranks the first, second, and third.

4.4.1 Effectiveness of Spatial Attention: Spatial attention slightly improves the segmentation results. Its influences on the basic Linknet are mainly in both Endovis17 and Endovis18, but its positive effects on the multiscale Linknet are in Endovis15 and Endovis17. Specifically, after adding spatial attention, the mean dice coefficient of the basic Linknet is dropped from 0.743 to 0.736 in Endovis15, but it increases from 0.734 to 0.748 in Endovis17, and similar improvements also happen in Endovis18.

4.4.2 Effectiveness of Multiple Scales: Segmentation performances are consistently improved after adding images in multiple scales. It improves the basic Linknet in all the datasets, and it significantly improves the attention Linknet in Endovis15 and Endovis18. For instance, the mean dice coefficient of the basic Linknet respectively increases from 0.743, 0.734, and 0.604 to 0.765, 0.738, and 0.643 in Endovis15, Endovis17, and Endovis18. Moreover, those of attention Linknets respectively increases from 0.736 and 0.628 to 0.767 and 0.728 in Endovis15 and Endovis18, but it slightly drops by 0.005 in Endovis17.

4.4.3 Effectiveness of Recurrent Refinement: Recurrent refinement is largely affected by the input types. Images combined with probability maps partially alleviate this problem, and the mean dice coefficient of multiscale Linknet with recurrent refinement modules can reach 0.772 in Endovis17.

It should be noticed that models with only images in the recurrent refinement perform better than those with additional masks. This may due to the following reasons: all their weights for extracting visual features are inherited from the pretrained residual networks trained with large-scale datasets, but the first convolutional layer using both images and probability maps has different numbers of channels from RGB ones in pretrained networks, therefore, they have to be trained from scratch and lead to more overfitting on datasets in small scales. Concretely, categories in Endovis15 are the least among all the datasets, which the first convolutional layer has the fewest channels, so training from scratch might have less effects when images are fused with probability maps; classes in Endovis17 is also similar to those in Endovis15, but there are much more data in Endovis17, therefore, the combination of images and masks improves by a large margin; semantic types in Endovis18 is significantly larger than those in the aforementioned datasets, but data size is almost the same as that in Endovis17, so training the first convolutional layer may be more difficult, and the performance gap before and after combining masks expands.

4.4.4 Iteractions between Spatial Attention and Multiple Scales: All the aforementioned components can be used to improve overall performances to different extents, and interactions

among various components are further analyzed. Specifically, Compared to the mean dice coefficient of multiscale Linknets, these models combined with spatial attention respectively increases from 0.765 to 0.767 in Endovis15, from 0.738 to 0.743 in Endovis17, from 0.643 to 0.728 in Endovis18, which demonstrates the easy integration and effectiveness of spatial attention after combining with multiple scales. Moreover, models only combining recurrent refinement and multiple scales achieve the best performances. Specifically, their mean dice coefficients are respectively 0.773, 0.772, and 0.755 in Endovis15, Endovis17, and Endovis18. However, attention partially improves segmentation performances of multiscale recurrent Linknets in Endovis17 but deteriorates them in both Endovis15 and Endovis18. For example, the mean dice coefficient of multiscale Linknet using recurrent refinement of images increases from 0.746 to 0.760 after adding spatial attention in Endovis17, but that of multiscale Linknet using refinement of both images and masks decreases from 0.772 to 0.769 after using spatial attention in Endovis17. Similar conclusions remain true in both Endovis15 and Endovis18.

4.5 Comparison with State of the Art Methods

Among all the experimented semantic segmentation models, our method ranks the first in all the datasets, as illustrated in Figure 3. Not only instruments but also backgrounds are essential to be segmented for manipulators to interact with environments in Endovis18, so our method have more potentials than others. Specifically, compared with other baseline methods, ICNet [67] seems the most unsuitable for this task, and the mean dice coefficients in all the datasets are less than 0.7. Ternausnet [73] may be good at classifying foreground pixels with relatively small numbers of parameters (32.20 million in Endovis17), but the metric is only 0.565 in Endovis18, which might limit its further applications.

Characteristics of segmentation results using different methods are further compared, as illustrated in Figure 4. Generally, there are three columns from left to right, which respectively represent the images and segmentation results with various methods in Endovis15, Endovis17, and Endovis18. Specifically, even though ICNet performs the worst in all the datasets, but it captures the global semantic structures quite well. More details are preserved using DUnet [33] and DAnet [19], but downsampling operations might be a little more in DUnet, and some false positives exist in DAnet. Ternausnet sometimes obtains spotted semantic regions and damages the global structure. Contrastly, our method removes the spotted areas and properly classifies pixels belonging to both instruments and environments.

5 Conclusion

In this paper, a multiscale recurrent refinement model is proposed to segment parts of instruments in various surgery. Specifically, to remove the spotted semantic regions, images in multiple resolutions are segmented using a weight-shared Linknet, then the above coarse outputs are further refined recurrently with another semantic segmentation module. Unlike other models with fixed blocks and connections, flexibly connected modules are introduced and integrated into different models, and neural architectures are explicitly searched according to segmentation performances in the target dataset. Compared with other methods, our model is more flexible and easier to be searched. Extensive experimental results demonstrate state-of-art performances of our model with significantly fewer parameters for part segmentation of instruments.

It should be noted that there are still some issues in our method: even though weight sharing significantly reduces the number of parameters in a cascade refinement architecture and progressively improve the segmentation result. However, fusing the image and the predicted mask for the next iteration remains naive, and the test speed inevitably becomes slower when predicted masks are refined more times. For future work, we plan to build a recurrent refinement model for temporal semantic segmentation and establish an effective framework to segment instruments across multiple domains.

6 Acknowledgments

This work is supported by National Key R&D Program of China under Grant 2018YFB1306500, National Key R&D Program of China under Grant 2018YFB1306300, and National Natural Science Foundation (NNSF) of China under Grant 61421004.

7 References

1 Allan, M., Shvets, A., Kurmann, T., et al.: '2017 robotic instrument segmentation challenge'. *arXiv preprint arXiv:1902.06426*, 2019
2 Allan, M., Kondo, S., Bodenstedt, S., et al.: '2018 robotic scene segmentation challenge'. *arXiv preprint arXiv:2001.11190*, 2020
3 Colleoni, E., Moccia, S., Du, X., De.Momi, E. and Stoyanov, D.: 'Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers', *IEEE Robotics and Automation Letters*, 2019, **4**, (3), pp. 2714–2721
4 García.Peraza.Herrera, L.C., Li, W., Fidon, L., et al.: 'Toolnet: holistically-nested real-time segmentation of robotic surgical tools', *Proc. Int. Conf. Intelligent Robots and Systems*, 2017, pp. 5717–5722
5 Hasan, S. and Linte, C.A.: 'U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instrument'. *arXiv preprint arXiv:1902.08994*, 2019
6 Islam, M., Atputharuban, D.A., Ramesh, R. and Ren, H.: 'Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning', *IEEE Robotics and Automation Letters*, 2019, **4**, (2), pp. 2188–2195
7 Pakhomov, D., Premachandran, V., Allan, M., Azizian, M. and Navab, N.: 'Deep residual learning for instrument segmentation in robotic surgery'. *arXiv preprint arXiv:1703.08580*, 2017
8 Zhang, H., Gao, Z., Xu, L., et al.: 'A meshfree representation for cardiac medical image computing', *IEEE Journal of Translational Engineering in Health and Medicine*, 2018, **6**, pp. 1800212
9 Liu, P., Yu, H. and Cang, S.: 'Adaptive neural network tracking control for underactuated systems with matched and mismatched disturbances', *Nonlinear Dynamics*, 2019, **98**, pp. 1447–1464
10 Liu, P., Huda, M.N., Tang, Z. and Sun, K.L.: 'A self-propelled robotic system with a visco-elastic joint: dynamics and motion analysis', *Engineering With Computers*, 2020, **36**, pp. 655–669
11 Liu, P., Neumann, G., Fu, Q., et al.: 'Energy-efficient design and control of a vibro-driven robot'. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018
12 Huang, R., Bian, G., Xin, C., et al.: 'Path planning for surgery robot with bidirectional continuous tree search and neural network'. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019. pp 3302–3307
13 Chaurasia, A. and Culurciello, E.: 'Linknet: Exploiting encoder representations for efficient semantic segmentation', *IEEE Visual Communications and Image Processing*, 2017, pp. 1–4
14 Chen, L.C., Papandreou, G., Schroff, F. and Adam, H.: 'Rethinking atrous convolution for semantic image segmentation'. *arXiv preprint arXiv:1706.05587*, 2017
15 Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H.: 'Encoder-decoder with atrous separable convolution for semantic image segmentation', *Proc. Eur. Conf. Computer Vision*, 2018, pp. 801–818
16 Chen, K., Pang, J., Wang, J., et al.: 'Hybrid task cascade for instance segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 4974–4983
17 Chen, Y., Li, W., Chen, X. and Gool, L.V.: 'Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 1841–1850
18 Fu, K., Zhao, Q. and Gu, I.Y.H.: 'Refinet: A deep segmentation assisted refinement network for salient object detection', *IEEE Trans. Multimedia*, 2018, **21**, (2), pp. 457–469
19 Fu, J., Liu, J., Tian, H., et al.: 'Dual attention network for scene segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154
20 Guo, D., Pei, Y., Zheng, K., Yu, H., Lu, Y. and Wang, S.: 'Degraded image semantic segmentation with dense-gram networks', *IEEE Trans. Image Processing*, 2020, **29**, pp. 782–795
21 Hu, R., Dollár, P., He, K., Darrell, T. and Girshick, R.: 'Learning to segment every thing', *Proc. Conf. Computer Vision and Pattern Recognition*, 2018, pp. 4233–4241
22 Hung, W.C., Jampani, V., Liu, S., Molchanov, P., Yang, M.H. and Kautz, J.: 'Scops: Self-supervised co-part segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 869–878
23 Kirillov, A., He, K., Girshick, R., Rother, C. and Dollár, P.: 'Panoptic segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413
24 Kohl, S., Romera.Paredes, B., Meyer, C., et al.: 'A probabilistic u-net for segmentation of ambiguous images', *Advances in Neural Information Processing Systems*, 2018, pp. 6965–6975
25 Kreso, I., Segvic, S. and Krapac, J.: 'Ladder-style densenets for semantic segmentation of large natural images', *Proc. Int. Conf. Computer Vision*, 2017, pp. 238–245
26 Li, Y., Chen, X., Zhu, Z., et al.: 'Attention-guided unified network for panoptic segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 7026–7035
27 Liu, C., Chen, L.C., Schroff, F., et al.: 'Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 82–92
28 Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z. and Wang, J.: 'Structured knowledge distillation for semantic segmentation', *Proc. Conf. Computer Vision and Pattern*

Recognition, 2019, pp. 2604–2613
29 Peng, C., Zhang, X., Yu, G., Luo, G. and Sun, J.: 'Large kernel matters–improve semantic segmentation by global convolutional network', *Proc. Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4353–4361
30 Pont.Tuset, J., Arbelaez, P., Barron, J.T., Marques, F. and Malik, J.: 'Multiscale combinatorial grouping for image segmentation and object proposal generation', *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2016, **39**, (1), pp. 128–140
31 Rajaram, R.N., Ohn.Bar, E. and Trivedi, M.M.: 'Refinenet: Iterative refinement for accurate object localization', *Proc. Int. Conf. Intelligent Transportation Systems*, 2016, pp. 1528–1533
32 Takikawa, T., Acuna, D., Jampani, V. and Fidler, S.: 'Gated-scnn: Gated shape cnns for semantic segmentation'. *arXiv preprint arXiv:1907.05740*, 2019
33 Tian, Z., He, T., Shen, C. and Yan, Y.: 'Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 3126–3135
34 Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J.: 'Pyramid scene parsing network', *Proc. Conf. Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890
35 Krähenbühl, P. and Koltun, V.: 'Efficient inference in fully connected crfs with gaussian edge potentials', *Advances in Neural Information Processing Systems*, 2011, pp. 109–117
36 He, W., Song, H., Guo, Y., et al.: 'A gallery-guided graph architecture for sequential impurity detection', *IEEE Access*, 2019, **7**, pp. 149105–149116
37 Gao, Z., Zhang, H., Dong, S., et al.: 'Salient object detection in the distributed cloud-edge intelligent network', *IEEE Network*, 2020, **34**, (2), pp. 216–224
38 Zang, D., Bian, G., Wang, Y., et al.: 'An extremely fast and precise convolutional neural network for recognition and localization of cataract surgical tools'. *Medical Image Computing and Computer Assisted Intervention Society*, 2019. pp 655–669
39 Ni, Z., Bian, G., Wang, G., et al.: 'Pyramid attention aggregation network for semantic segmentation of surgical instruments'. *AAAI Conference on Artificial Intelligence*, 2020
40 Bromley, J., Bentz, J.W., Bottou, L., et al.: 'Signature verification using a siamese time delay neural network', *International Journal of Pattern Recognition and Artificial Intelligence*, 1993, **7**, (4), pp. 669–688
41 Man, Y., Huang, Y., Feng, J., Li, X. and Wu, F.: 'Deep q learning driven ct pancreas segmentation with geometry-aware u-net', *IEEE Trans. Medical Imaging*, 2019, **38**, (8), pp. 1971–1980
42 Liu, H., Peng, C., Yu, C., et al.: 'An end-to-end network for panoptic segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 6172–6181
43 Huo, Y., Xu, Z., Bao, S., et al.: 'Splenomegaly segmentation on multi-modal mri using deep convolutional networks', *IEEE Trans. Medical Imaging*, 2018, **38**, (5), pp. 1185–1196
44 Roy, A.G., Navab, N. and Wachinger, C.: 'Concurrent spatial and channel squeeze & excitation in fully convolutional networks', *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 421–429
45 Shin, S.Y., Lee, S., Yun, I.D. and Lee, K.M.: 'Deep vessel segmentation by learning graphical connectivity', *Medical Image Analysis*, 2019, **58**, pp. 101556
46 Milletari, F., Navab, N. and Ahmadi, S.A.: 'V-net: Fully convolutional neural networks for volumetric medical image segmentation', *Proc. Int. Conf. 3D Vision*, 2016, pp. 565–571
47 Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C. and Ayed, I.B.: 'Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation', *IEEE Trans. Medical Imaging*, 2018, **38**, (5), pp. 1116–1126
48 Novikov, A.A., Major, D., Wimmer, M., Lenis, D. and Bühler, K.: 'Deep sequential segmentation of organs in volumetric medical scans', *IEEE Trans. Medical Imaging*, 2018, **38**, (5), pp. 1207–1215
49 Lin, G., Milan, A., Shen, C. and Reid, I.: 'Refinenet: Multi-path refinement networks for high-resolution semantic segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2017, pp. 1925–1934
50 Lai, Q., Wang, W., Sun, H. and Shen, J.: 'Video saliency prediction using spatiotemporal residual attentive networks', *IEEE Trans. Image Processing*, 2020, **29**, pp. 1113–1126
51 Ronneberger, O., Fischer, P. and Brox, T.: 'U-net: Convolutional networks for biomedical image segmentation', *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241
52 Jin, Q., Meng, Z., Sun, C., Wei, L. and Su, R.: 'Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans'. *arXiv preprint arXiv:1811.01328*, 2018
53 Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J.: 'Unet++: A nested u-net architecture for medical image segmentation', *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pp. 3–11
54 Ibtehaz, N. and Rahman, M.S.: 'Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation'. *arXiv preprint arXiv:1902.04049*, 2019
55 Gu, Z., Cheng, J., Fu, H., et al.: 'Ce-net: Context encoder network for 2d medical image segmentation', *IEEE Trans. Medical Imaging*, 2019, **38**, (10), pp. 2281–2292
56 Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F. and Giro-i Nieto, X.: 'Rvos: End-to-end recurrent network for video object segmentation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 5277–5286
57 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
58 Wu, H., Zhang, J., Huang, K., Liang, K. and Yu, Y.: 'Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation'. *arXiv preprint arXiv:1903.11816*, 2019
59 Zhu, Y., Sapra, K., Reda, F.A., et al.: 'Improving semantic segmentation via video propagation and label relaxation', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 8856–8865
60 Wang, Y., Tao, X., Shen, X. and Jia, J.: 'Wide-context semantic image extrapolation'. *Proc. Conf. Computer Vision and Pattern Recognition*, 2019. pp 1399–1408
61 Maninis, K.K., Pont.Tuset, J., Arbeláez, P. and Van.Gool, L.: 'Deep retinal image understanding', *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 140–148
62 Son, J., Park, S.J. and Jung, K.H.: 'Retinal vessel segmentation in fundoscopic images with generative adversarial networks'. *arXiv preprint arXiv:1706.09318*, 2017
63 Wang, C., Dong, S., Zhao, X., et al.: 'Saliencygan: Deep learning semisupervised salient object detection in the fog of iot', *IEEE Transactions on Industrial Informatics*, 2020, **16**, (4), pp. 2667–2676
64 Enokiya, Y., Iwamoto, Y. and Chen, Y.W.: 'Automatic liver segmentation using u-net with wasserstein gans', *Journal of Image and Graphics*, 2018, **6**, (2), pp. 152–159
65 Tan, J., Jing, L., Huo, Y., Tian, Y. and Akin, O.: 'Lgan: Lung segmentation in ct scans using generative adversarial network'. *arXiv preprint arXiv:1901.03473*, 2019
66 Xie, L., Yu, Q., Wang, Y., Zhou, Y., Fishman, E.K. and Yuille, A.L.: 'Recurrent saliency transformation network for tiny target segmentation in abdominal ct scans', *IEEE Trans. Medical Imaging*, 2020, **39**, (2), pp. 514–525
67 Zhao, H., Qi, X., Shen, X., Shi, J. and Jia, J.: 'Icnet for real-time semantic segmentation on high-resolution images', *Proc. Eur. Conf. Computer Vision*, 2018, pp. 405–420
68 Zhang, C., Lin, G., Liu, F., Yao, R. and Shen, C.: 'Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning', *Proc. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226
69 Wang, W., Yu, K., Hugonot, J., Fua, P. and Salzmann, M.: 'Recurrent u-net for resource-constrained segmentation'. *arXiv preprint arXiv:1906.04913*, 2019
70 Newell, A., Yang, K. and Deng, J.: 'Stacked hourglass networks for human pose estimation', *Proc. Eur. Conf. Computer Vision*, 2016, pp. 483–499
71 Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P.: 'Focal loss for dense object detection', *Proc. Int. Conf. Computer Vision*, 2017, pp. 2980–2988
72 Liu, L., Jiang, H., He, P., et al.: 'On the variance of the adaptive learning rate and beyond'. *arXiv preprint arXiv:1908.03265*, 2019
73 Iglovikov, V. and Shvets, A.: 'Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation'. *arXiv preprint arXiv:1801.05746*, 2018