Kid-Net: Convolution Networks for Kidney Vessels Segmentation from CT-Volumes

Ahmed Taha[1], Pechin Lo[2], Junning Li[2], and Tao Zhao[2]

[1] University of Maryland, College Park , ahmdtaha@cs.umd.edu
[2] Intuitive Surgical, Inc, {firstname.lastname}@intusurg.com

Abstract. Semantic image segmentation plays an important role in modeling patient-specific anatomy. We propose a convolution neural network, called Kid-Net, along with a training schema to segment kidney vessels: artery, vein and collecting system. Such segmentation is vital during the surgical planning phase in which medical decisions are made before surgical incision. Our main contribution is developing a training schema that handles unbalanced data, reduces false positives and enables high-resolution segmentation with a limited memory budget. These objectives are attained using dynamic weighting, random sampling and $3D$ patch segmentation.

Manual medical image annotation is both time-consuming and expensive. Kid-Net reduces kidney vessels segmentation time from matter of hours to minutes. It is trained end-to-end using $3D$ patches from volumetric CT-images. A complete segmentation for a $512 \times 512 \times 512$ CT-volume is obtained within a few minutes (1-2 mins) by stitching the output $3D$ patches together. Feature down-sampling and up-sampling are utilized to achieve higher classification and localization accuracies. Quantitative and qualitative evaluation results on a challenging testing dataset show Kid-Net competence.

Keywords: CT-volumes, segmentation, kidney, biomedical, convolution, neural networks

1 Introduction and Related Work

After its success in classification [9] and action recognition [5], convolution neural networks (CNN) began achieving promising results in challenging semantic segmentation tasks [12,14,1]. One key pillar is its ability to learn features from raw input data– without relying on hand-crafted features. A second recent pillar is the ability to precisely localize these features when combining convolution features at different scales. Such localization approach eliminate the need for traditional hand-crafted post processing like Dense-CRF [2,8]. Thus, end-to-end CNN training for challenging segmentation problems becomes feasible. This sheds light on semantic segmentation applications in medical field.

Semantic segmentation for human anatomy using volumetric scans like MRI and CT-volumes is an important medical application. It is a fundamental step to perform or plan surgical procedures. Recent work uses automatic segmentation

to do computer assisted diagnosis [13,15], interventions [17] and segmentation from sparse annotations [4]. Recently, U-shaped networks [14] managed to train fully $2D$ convolution network for semantic segmentation in an end-to-end fashion. These architectures have two contradicting phases that carry out complementary tasks. The down-sampling phase detects features, while the up-sampling phase accurately localizes the detected features. Such combination is proven essential in recent literature [11,4,12] to acquire precise segmentation.

Inspired by U-shaped networks, we propose Kid-Net to segment $3D$ patches from volumetric CT-images. We agree with [12] that $3D$ convolutions are better than slice-wise approaches when processing volumetric scans. We build on [12] architecture by processing volumetric patches to enable high resolution segmentation and thus bypass GPU memory constraints. Despite the promising results, such vanilla model suffers due to unbalanced data. This leads to our main contribution which is balancing both intra-foreground and background-foreground classes within independent patches. This achieves the best results as presented in the experiments section. Accordingly, manual preprocessing like cropping or down-sampling workarounds are no longer needed for high resolution CT-volume segmentation.

In this work, we aim to segment kidney vessels: artery, vein and collecting system (ureter). This task is challenging for a number of reasons. First, the CT-volume is huge to fit in memory. To avoid processing the whole CT-volume, we process $3D$ patches individually. Second, foreground and background classes are unbalanced and most patches are foreground-free. Another major challenge is obtaining the groundtruth for training. Medical staff annotate our data; their prior knowledge leads to incomplete groundtruth. Vessels far from kidney are considered less relevant and thus ignored. Figure 1 shows a CT-slice with the three foreground classes. It highlights the problem difficulty even for a well-informed technician.
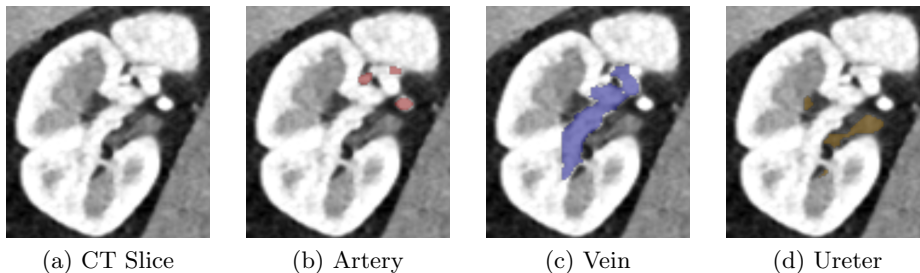

| (a) CT Slice | (b) Artery | (c) Vein | (d) Ureter |

Fig. 1. CT slice contains three foreground classes: artery (red), vein (blue) and collecting system (orange). Best seen in color and zoom

2   Method

Inspired by U-Net structure [14], Kid-Net is divided into two main phases: down-sampling and up-sampling. Figure 2 shows these phases and how Kid-Net up-

sampling phase is different from U-Net. In U-Net, down-sampled features are repeatedly up-sampled and concatenated with the corresponding feature till a single segmentation result, with the original image resolution, is obtained. Kid-Net is similar to U-Net but adds residual links extension in which each down-sampled feature is independently up-sampled 2^n times till the original resolution is restored. Thus unlike U-Net, Kid-Net generates multiple segmentation results that are averaged to obtain a final segmentation result. These residual links follow Junning et. al [10] Residual-U-Net design. Sequential non-linear functions accumulation improves deep neural networks performance in image recognition. In our paper, residual links are added in up-sampling phase only for simplicity

Kid-Net segments kidney vessels from CT-volumes. The foreground classes are artery, vein and collecting system (ureter) vessels. To avoid the large memory requirement of CT-volumes, Kid-Net is trained using $3D$ CT-volume patches from $R^{96\times96\times96}$. Without any architecture modification, wider context through bigger patches, within GPU memory constraints, are feasible. Kid-Net outputs a soft-max probability maps for artery, vein, collecting system and background. Instead of training for individual foreground classes independently, our network is trained to detect the three foreground classes. Such approach has two advantages; first, a single network decision per voxel fills in for a heuristic-based approach to merge multiple networks decisions. Second, this approach aligns with [6] recommendation that learning tasks with less data benefit largely from joint training with other tasks.
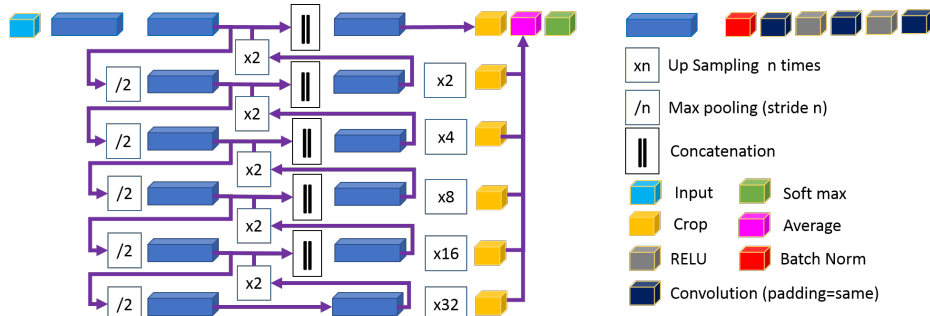

Fig. 2. KID-Net architecture. The two contradicting phases are colored in blue. The down-sampling and up-sampling phases detect and localize features respectively. The segmentation result, at different scale levels, are averaged to compute the final segmentation. Best seen in color and zoom

While training Kid-Net with $3D$ patches bypasses GPU memory constraints, a new challenge surfaces– unbalanced data. Patches majority are foreground-free. Even when a foreground class exists, it occupies a small percentage. In this paper, we leverage recent work [11,4], that addresses tiny structures precise localization,

to tackle unbalanced data. As follows, a two-fold approach is proposed to hinder bias against the tiny structured foreground classes.

The first fold assigns dynamic weights to individual voxels based on their classes and patches significance. A major challenge in medical image processing is detecting small structures. Vessels in direct contact with kidney are tiny and their resolution is limited by acquisition. Tiny vessels are challenging to annotate, more valuable to acquire. Thus, patches containing smaller vessels are more significant, have higher weight. Patch weight is inversely proportional to the vessel volume inside it. Foreground classes are also assigned higher weights to hinder the network background bias. The key idea is to measure the average foreground classes volumes per patch dynamically during training. Then, assign higher weights to classes with smaller average volumes and vice versa.

A policy that a vessel volume $\times$ weight must equal $1/n$ is imposed where n is the number of classes including the background class. Thus, all classes contribute equally from a network perspective. To account for data augmentation, vessels volumes are measured during training. Enforcing equal contribution (volume $\times$ weight) from all classes is our objective . To do so, we use the moving average procedure outlined in algorithm 1.

Algorithm 1 Our proposed moving average procedure assigns voxels dynamic weights (VW_c) based on their classes and patch weight. Patch Weight (PW) is inversely proportional to its foreground volume. Class weight (CW) is inversely proportional to its average volume per patch. Initially $V_c = \frac{1}{n}$ for every class c. Our settings $\alpha = 0.001$, $n = 4$.

Require: α : Convergence Rate
Require: P : Current $3D$ patch with axis x, y, z
Require: n : Number of classes (background included)
Require: V_c : Class (c) moving average volume
Require: PW : Current patch weight
Require: CW_c : Class (c) weight
  for all c in classes **do**
    // Measure class (c) volume in patch P
    $V_c(P) = \left(\sum_x \sum_y \sum_z P(x, y, z) == c \right) / size(P)$
    // Update class (c) moving average volume
    $V_c = V_c \times (1 - \alpha) + V_c(P) \times \alpha$
    // Update class weight based on its moving average volume
    $CW_c = 1/(n \times V_c)$
  end for
  // Set patch weight based on foreground volume
  if P contains background only **then**
    $PW = 1$
  else
    // Foreground volume $\sum_{c=1}^{n-1} V_c(P) < 1$
    $PW = 1 - log(\sum_{c=1}^{n-1} V_c(P))$
  end if
  $VW_c = PW * CW_c$ (Voxel weight is function of PW and CW_c)

Due to background relative huge volume, it's assigned tiny class weight. So the network produces a lot of false positives – it is cheap to mis-classify a background voxel. To tackle this undesired phenomena, we propose our second com-

plementary fold – Random Sampling. Random Background voxels are sampled and assigned high weights. Such method is most effective in limiting false positives because high loss is incurred for these voxels if mis-classified. Figure 3 shows our sampling schema. Two bands are constructed around kidney vessels using morphological kernel, a binary circular dilation of radii two and four. Misclassifications within the first band (< 2 voxels away from the vessel) are considered marginal errors. In a given patch, the sampled background voxels are equivalent to the foreground vessel volume, where 20% and 80% come from the red band and the volume beyond this band respectively. If a patch is foreground-free, 1% voxels are randomly sampled.

While using advanced U-shaped architectures can lead to marginal improvements, dynamic weighting and random sampling are indispensable. Both weighting and sampling are done during training phase while patches are fed into the network.
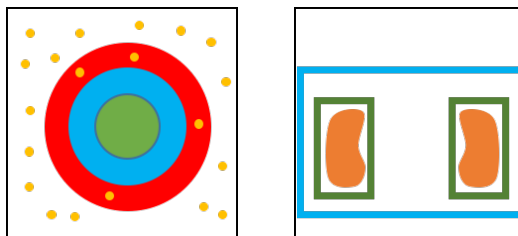

Fig. 3. (Left) Background sampling approach. Foreground vessel, in green, surrounded by two bands, blue and red, at distance 2 and 4 voxels. Equivalent foreground volume, 20% and 80%, is randomly sampled from the red band and the volume beyond this band respectively. (Right) Experiments evaluation regions. The first region is the whole region of interest defined per subject ground truth outlined in blue. The second region is the kidney bounding box outlined in green.

3 Experiments

In our experiments, we use volumetric CT-scans from 236 subjects. The average spacing is $0.757 \times 0.757 \times 0.906mm$, with standard deviation $0.075 \times 0.075 \times 0.327mm$. Kid-Net is trained using $3D$ patches from 99 cases, while 30 and 107 cases are used for validation and testing respectively. Training patches are presliced to reduce I/O bottleneck. They are uniformly sliced from CT-scans random points at background, collecting system, artery and vein centerlines. Training with foreground-free patches is mandatory. When eliminated, performance degrades because the network learns that every patch has a foreground object, and segments accordingly, which is false.

Two training schema, with and without random sampling, are evaluated. In both schema dynamic weighting following algorithm 1 is applied. Training without both dynamic weighting and random sampling leads to a degenerate

segmentation – background class-bias. Kid-Net is trained using Keras API [3], Adam optimizer [7], and a categorical cross entropy loss function. Segmentation results are evaluated using dice-coefficient – F1 score [16].

Both artery and vein have tree-structure; they are thick near aorta and vena cava, while fine at terminals near renal artery and renal vein. These fine vessels are most difficult to annotate, i.e. most valuable to acquire. Dice-coefficient is biased against fine details in such tree-structure. To overcome such limitation, we evaluate the two regions depicted in figure 3. The first region is based on the ground truth region of interest. This evaluates the whole tree-structure including the thick branches– aorta and vena cava. The second region is the kidney bounding box. It targets fine vessels in direct contact with the kidney.

The ground truth region of interest (ROI) is subject dependent. During evaluation, we clip our output in z-axis, based on the per subject ground truth ROI. It is worth-noting that the ground truth annotation is incomplete for two reasons. First, vessels in direct connect with kidney are challenging to annotate due to their tiny size, thus the ground truth is a discretized vessel islands – 12 islands on average. Second, vessels far from kidney have little value, for kidney surgery, to annotate and so are typically missing. To avoid penalizing valid segmentations, we evaluate predictions overlapping with known ground-truth vessel islands. This evaluation approach reduces the chances of falsely penalizing unannotated detections. It also aligns with the premise that neural networks assist, but not replace, human especially in medical applications.

Table 1. Quantitative evaluation for different training schema in two evaluation regions. Dynamic Weighting (DW) plus Random Sampling (RS) achieves the highest accuracies. Artery F1 score is the highest.

	Whole ROI		Kidney Bounding box	
	DW	DW+RS	DW	DW+RS
Artery	0.86	**0.88**	0.72	**0.72**
Vein	**0.59**	0.57	0.60	**0.67**
Ureter	0.32	**0.62**	0.41	**0.63**

Table 1 summarizes the quantitative results and highlights our network ability to segment both coarse and fine vessels around the kidney. Artery segmentation is the most accurate because all scans are done during arterial phase. This suggests that better vein and ureter segmentations are feasible if venous and waste-out scans are available. The thick aorta boosts artery segmentation F1 score in the whole ROI. In the kidney bounding box, tiny artery vessels become more challenging, and F1 score relatively decreases. The same argument explains vein vessels F1 score. Since arterial scans are used, concealed vena cava penalizes F1 score severely in the whole ROI region. This observation manifests in figure 4, second column. While aorta is easy to segment and boosts F1 score, vena cava is more challenging and thus F1 score degrades.
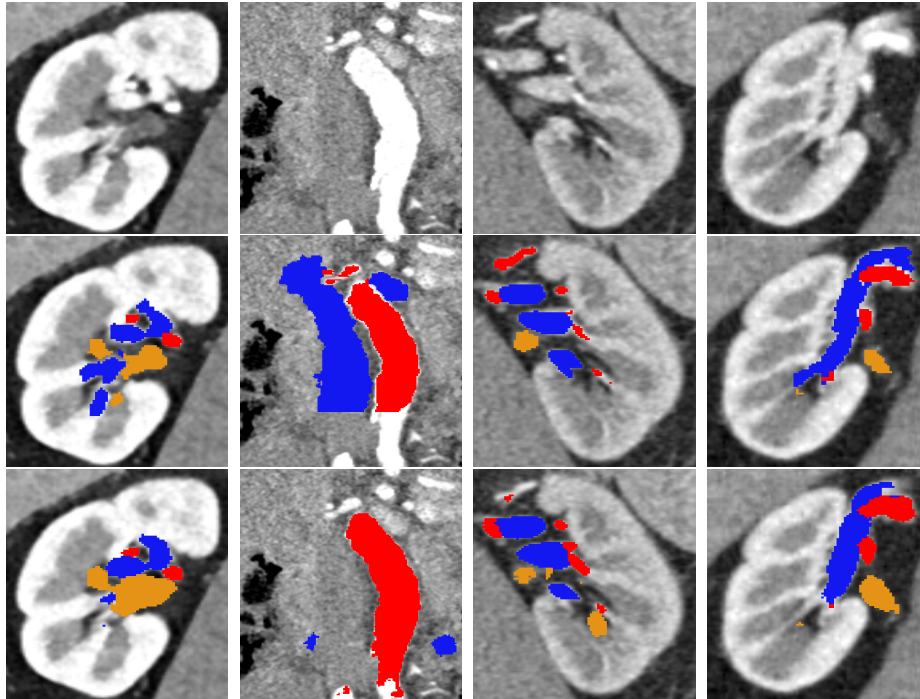
Fig. 4. Qualitative evaluation results. Rows contain raw CT slice, ground truth annotation, and Kid-Net segmentation respectively. Artery, vein, and ureter are highlighted in red, blue, and orange. Best seen in color and zoom

Among the three kidney vessels, the collecting system is the most challenging. Due to their tiny size, it is difficult to manually annotate or automatically segment. Ureter vessels ground truth annotations are available only within the kidney proximity– far ureter are less relevant. This explains why ureter F1 score is similar in both evaluation regions. Ureter class is assigned the highest weight due to its relative small size. This leads to a lot of false ureter positives. While random sampling has limited effect on artery and vein F1 score, its merits manifest in ureter segmentation. It boosts segmentation accuracy by 30% and 22% in the whole ROI and kidney bounding box respectively. Thus, It is concluded that both dynamic weights and random sampling are essential to achieve accurate tiny vessels segmentation.

Figure 4 shows qualitative results and highlights vessels around the kidney. All CT-slices are rendered using soft tissue window–level=40, width=400. Vein and collecting system segmentation are the most challenging. The second column shows a shortcoming case due to a concealed vein. Fine vessels near kidney are the most difficult to annotate. Manually annotating such vessels can be cumbersome and time consuming. Thats why Kid-Net is valuable; its training schema enables fine anatomy segmentation in high resolution CT-volumes, and voids GPU memory constraints.

4 Conclusion

We propose Kid-Net, a convolution neural network for kidney vessels segmentation. Kid-Net achieves great performance segmenting different vessels by processing $3D$ patches. Fitting a whole CT-volume in memory, to train a neural network, is no longer required. We propose a two-fold solution to balance foreground and background distributions in a training dataset. Dynamically weighting voxels resolves unbalanced data. Assigning higher weights to randomly sampled background voxels effectively reduces false positives. The proposed concepts are applicable to other fine segmentation tasks.

References

1. Bertasius, G., Shi, J., Torresani, L.: Semantic segmentation with boundary neural fields. In: CVPR (2016)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915 (2016)
3. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: MICCAI (2016)
5. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence (2013)
6. Kaiser, L., Gomez, A.N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., Uszkoreit, J.: One model to learn them all. arXiv preprint arXiv:1706.05137 (2017)
7. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS (2011)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
10. Li, J., Lo, P., Taha, A., Wu, H., Zhao, T.: Segmentation of renal structures for image-guided surgery. MICCAI (2018)
11. Merkow, J., Marsden, A., Kriegman, D., Tu, Z.: Dense volume-to-volume vascular boundary detection. In: IJCARS (2016)
12. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV) (2016)
13. Porter, C.R., Crawford, E.D.: Combining artificial neural networks and transrectal ultrasound in the diagnosis of prostate cancer. Oncology (Williston Park, NY) (2003)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. arXiv preprint arXiv:1505.04597 (2015)
15. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE TMI (2016)
16. Sørensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. Biol. Skr. (1948)
17. Zettinig, O., Shah, A., Hennersperger, C., Eiber, M., Kroll, C., Kübler, H., Maurer, T., Milletarì, F., Rackerseder, J., zu Berge, C.S., et al.: Multimodal image-guided prostate fusion biopsy based on automatic deformable registration. IJCARS (2015)