

Multi-Scale Attentional Network for Multi-Focal Segmentation of Active Bleed after Pelvic Fractures

Yuyin Zhou¹, David Dreizin², Yingwei Li¹, Zhishuai Zhang¹,
Yan Wang¹, Alan Yuille¹

¹The Johns Hopkins University

²University of Maryland & R. Adams Cowley Shock Trauma Center

Abstract. Trauma is the worldwide leading cause of death and disability in those younger than 45 years, and pelvic fractures are a major source of morbidity and mortality. Automated segmentation of multiple foci of arterial bleeding from abdominopelvic trauma CT could provide rapid objective measurements of the total extent of active bleeding, potentially augmenting outcome prediction at the point of care, while improving patient triage, allocation of appropriate resources, and time to definitive intervention. In spite of the importance of active bleeding in the quick tempo of trauma care, the task is still quite challenging due to the variable contrast, intensity, location, size, shape, and multiplicity of bleeding foci. Existing work presents a heuristic rule-based segmentation technique which requires multiple stages and cannot be efficiently optimized end-to-end. To this end, we present, Multi-Scale Attentional Network (MSAN), the first yet reliable end-to-end network, for automated segmentation of active hemorrhage from contrast-enhanced trauma CT scans. MSAN consists of the following components: 1) an encoder which fully integrates the global contextual information from holistic 2D slices; 2) a multi-scale strategy applied both in the training stage and the inference stage to handle the challenges induced by variation of target sizes; 3) an attentional module to further refine the deep features, leading to better segmentation quality; and 4) a multi-view mechanism to leverage the 3D information. MSAN reports a significant improvement of more than 7% compared to prior arts in terms of DSC.

1 Introduction

High-energy pelvic fractures, which are usually related to motor vehicle accidents, falls from height, or crush injury, are the second leading cause of death from acute physical trauma after brain injury. The mortality rate of pelvic fractures ranges from 5% ~ 15%, overall, increasing from 36% to 54% in those with hemorrhagic shock [12]. With the widespread availability of CT in trauma bays, the majority of patients with severe pelvic trauma admitted to level I trauma centers currently undergo an examination with contrast-enhanced trauma CT, in part to assess for foci of active bleeding, manifesting as contrast extravasation [3]. The size of foci of contrast extravasation from bleeding vessels correlates with

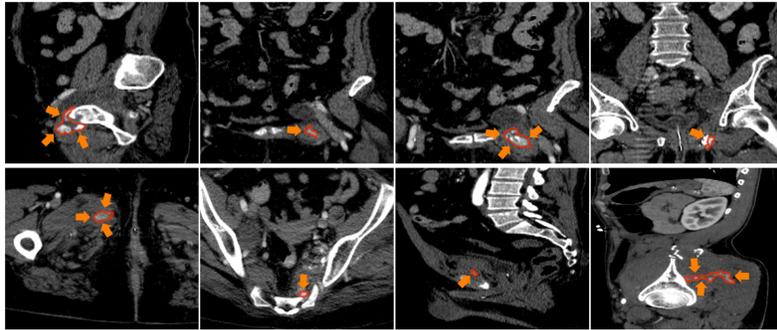


Fig. 1. Visual examples of pelvic CT scans from axial/coronal/sagittal views. Red contour denotes the boundaries of the active hemorrhage, where we can observe large variations of shape and textures.

the need for blood transfusion, angiographic or surgical hemostatic intervention, and mortality, but reliable measurements of contrast extravasation volume cannot be derived at the point of care using manual, semi-automated, or shorthand diameter-based methods. Fully automated methods are necessary for real-time point-of-care decision making, treatment planning, and prognostication.

In this paper, we focus on volumetric segmentation of foci of active bleeding (i.e. contrast extravasation) after pelvic fractures. This task is of vital importance yet challenging for the following reasons: 1) hemorrhage gray levels vary from patient to patient, depending on a variety of factors (*e.g.*, the rate of bleeding, the timing of the scan, and the patients physiologic state after trauma), 2) hemorrhage boundaries are often very poorly defined and highly irregular; and 3) the intensity levels are inconsistent throughout the region of a hemorrhagic focus. Prior works have utilized semi-automated threshold- or region growing-based methods using post-processing software [5]. However, these techniques are too time-consuming for clinical use in the trauma radiology setting. To overcome this difficulty, a method [4] was previously proposed to first utilize spatial contextual information from artery and bone to detect the hemorrhage, and then employ a rule-based strategy to refine the segmentation results. This heuristic approach requires multiple stages which cannot be efficiently optimized end-to-end. Moreover, this method cannot properly handle other challenges such as variation of target sizes and ambiguous boundaries.

Recently, the emerge of deep learning has largely advanced the field of computer aided diagnosis (CAD). Riding on the success of convolutional neural networks, *e.g.*, fully convolutional networks [9], researchers have achieved accurate segmentation on many medical image analysis tasks [10, 11, 15, 16]. Existing coarse-to-fine methods [15, 14], which propose to refine segmentation results through explicit cropping of a single region of interest (ROI) are more suitable for single connected structures such as the pancreas or liver, while sites of active bleeding are frequently discontinuous and multi-focal and occur in widely disparate vascular territories. Herein, we present a multi-scale attentional network (MSAN), for segmenting active bleed after pelvic features, the first yet reliable framework, for segmenting active bleed after pelvic features. Specifically, our

framework is able to 1) fully exploit contextual information from holistic 2D slices via using an encoder which is capable of extracting the global contextual information across different levels of image features; 2) efficiently handle the variation of active hemorrhage sizes by adopting multi-scale strategies during the training phase and the testing phase; 3) deal with the ambiguous boundaries by utilizing an attentional mechanism to better enhance the discrimination between trauma region and non-trauma region; 4) utilize the aggregation of multiple views (*i.e.*, Coronal, Sagittal and Axial views) to further leverage the 3D information. To assess the effectiveness of our framework, we collect a dataset of 65 patients with pelvic fractures and active hemorrhage with widely varying degrees of severity. For each case, every pixel/voxel of active hemorrhage was manually labeled by an experienced radiologist. Unlike the previously described heuristic method which used crude and not widely adopted measurements of accuracy such as missegmented area [4], we employed the Dice-Sørensen coefficient (DSC) for evaluation based on pixel/voxel-wise predictions. Experimental results demonstrate the superiority of our framework compared with a series of 2D/3D state-of-the-art deep learning algorithms.

2 Multi-Scale Attentional Network

2.1 Overall Framework

We denote a 3D CT-scanned image as \mathbf{X} with size $W \times H \times L$, where each element of \mathbf{X} indicated the Housefield Unit (HU) of a voxel. The corresponding binary ground-truth segmentation mask is denoted as \mathbf{Y} where $y_i = 1$ indicates a foreground voxel. Consider a segmentation model $M : \mathbf{Z} = \mathbf{f}(\mathbf{X}; \Theta)$, where M is parameterized by Θ , our goal is to predict a binary output volume \mathbf{Z} of the same dimension as \mathbf{X} . We denote \mathcal{Y} and \mathcal{Z} as the set of foreground voxels in the ground-truth and prediction, *i.e.*, $\mathcal{Y} = \{i \mid y_i = 1\}$ and $\mathcal{Z} = \{i \mid z_i = 1\}$. The accuracy of segmentation is evaluated by the Dice-Sørensen coefficient (DSC): $\text{DSC}(\mathcal{Y}, \mathcal{Z}) = \frac{2 \times |\mathcal{Y} \cap \mathcal{Z}|}{|\mathcal{Y}| + |\mathcal{Z}|}$. This metric falls in the range of $[0, 1]$, and $\text{DSC} = 1$ implies a perfect segmentation.

Following [15, 14, 11], 3 sets of images, *i.e.*, $\mathbf{X}_{C,w}$ ($w = 1, 2, \dots, W$), $\mathbf{X}_{S,h}$ ($h = 1, 2, \dots, H$) and $\mathbf{X}_{A,l}$ ($l = 1, 2, \dots, L$) are obtained along three axes. The subscripts C, S and A stand for “coronal”, “sagittal” and “axial”, respectively. We train an individual model M for each of the three viewpoints. Without loss of generality, we consider a 2D slice along the *axial* view, denoted by $\mathbf{X}_{A,l}$. Our goal is to infer a binary segmentation mask $\mathbf{Z}_{A,l}$ of the same dimensionality. In the context of deep networks [9, 1], it is achieved by computing a *probability map* $\mathbf{P}_{A,l} = \mathbf{f}[\mathbf{X}_{A,l}; \theta]$, where $\mathbf{f}[\cdot; \theta]$ is the architecture as in Fig. 2(a). This network contains an encoder (Sec. 2.2) to extract different levels of features for distilling global context and an attentional module (Sec. 2.3) as further refinement.

Specifically, we apply Atrous Spatial Pyramid Pooling (ASPP) [1] at the end of the backbone model to extract high-level features with enriched global context. Meanwhile, the low-level features extracted from earlier layers which

contain local information are fed to an attentional module to distill more useful information. The refined low-level features are then concatenated with high-level features extracted by ASPP and fed to the final classifier layer, which outputs probabilities $\mathbf{P}_{A,l}$, $\mathbf{P}_{C,l}$ and $\mathbf{P}_{S,l}$ which are then binarized into $\mathbf{Z}_{A,l}$, $\mathbf{Z}_{C,l}$ and $\mathbf{Z}_{S,l}$ respectively. The final segmentation outcome can be fused from the three views via majority voting [15, 14]. Multi-scale processing [8, 1] is used in both the training stage and the inference stage to further enhance the segmentation accuracy, especially for small targets. As illustrated in Fig. 2, different rescaled version of the original image are fed to the network during training. During the testing stage, to produce the final segmentation mask, the output from different scales are fused by taking at each position the average response. If the average probability is larger than a certain threshold ρ it is regarded as foreground otherwise it is regarded as background.

2.2 Encoder Backbone Architecture

Atrous Convolution has been widely applied in computer vision problems, which can efficiently allow for larger receptive field via controlling atrous rates. Given an input feature map x , atrous convolution is applied over x as follows:

$$y[i] = \sum_k x[i + r \cdot k]w[k], \quad (1)$$

where i and w denote the spatial location and the convolution filter, respectively. r stands for the atrous rate.

Atrous Spatial Pyramid Pooling (ASPP) is originated from Spatial Pyramid Pooling [7]. The main difference is that ASPP uses atrous convolution which allows for larger field-of-view during training and thus can efficiently integrate global contextual information. As a strong contextual aggregation module [1], ASPP is applied (see Fig. 2(a)) so that the contextual information from artery and bone can be better exploited. In our experiment, we set the atrous rates to be $\{12, 24, 36\}$, respectively.

2.3 Attentional Module

We adapt the non-local block [13] as the attentional module in our framework. Specifically, it first computes an attention map y of an input feature map x by taking a weighted average of features in all spatial locations \mathcal{L} :

$$y_i = \frac{1}{\mathcal{C}(x)} \sum_{\forall j \in \mathcal{L}} f(x_i, x_j) \cdot x_j, \quad (2)$$

where i and j are spatial indices. A pairwise function $f(x_i, x_j)$ is used to compute the spatial attention coefficients between each i and all j . And these coefficients are applied as the weighting of the input feature to better prune out irrelevant background features and thereby distinguish salient image regions. $\mathcal{C}(x)$

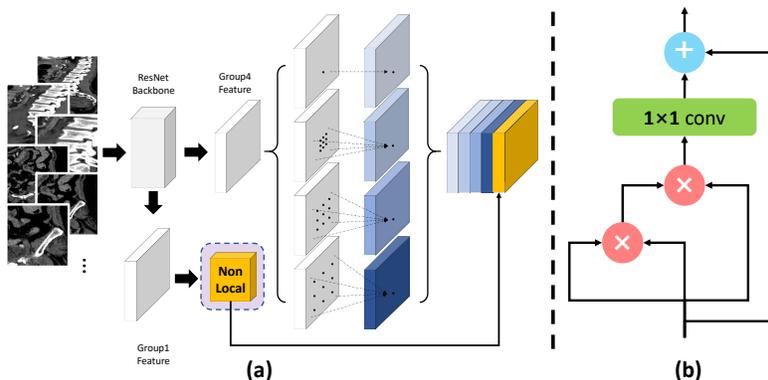


Fig. 2. (a) The network architecture structure of MSAN. Low-level features are refined by an attentional module. Meanwhile ASPP is applied at the end of the backbone model to extract high-level features with enriched global context. (b) Our implementation of the attentional module, where we use nonlocal means [13] as the main operation.

is a normalization function. We use the dot product version in [13] by setting $f(x_i, x_j) = x_i^T x_j$ and $\mathcal{C}(x) = N$, where N is the number of pixels in x .

Following [13], the attention map y is then processed by a 1×1 convolutional layer and added to the input feature map x to obtain the final output z , *i.e.*, $z = wy + x$, where w is the weight of the convolutional layer. An illustration our attentional module can be found in Fig. 2(b).

3 Experiments

3.1 Dataset and Evaluation

We have collected 65 studies were routinely acquired with 64 section or higher MDCT scanners in the trauma bay in either the late arterial or portal venous phase of enhancement. We use 45 cases for training and evaluate the segmentation performance on the rest 20 cases. Note that [4] was studied on only 12 cases, which, to the best of our knowledge, was the first and only curated dataset with manual ground truth label masks. Therefore our dataset can be considered as a valid set for evaluation. The metric we use is DSC, which measures the similarity between the prediction voxel set \mathcal{Z} and the ground-truth set \mathcal{Y} , with the mathematical form of $DSC(\mathcal{Z}, \mathcal{Y}) = \frac{2 \times |\mathcal{Z} \cap \mathcal{Y}|}{|\mathcal{Z}| + |\mathcal{Y}|}$.

3.2 Implementation details

Our implementations are based on Tensorflow. We used two standard architectures, *i.e.*, ResNet-50 and ResNet-101 [6] as backbone models. All our segmentation experiments were performed on the whole pelvic CT scan and were run on Tesla V100 GPU. For data pre-processing, following [11], we simply truncated the raw intensity values to be within the range of $[-80, 320]$ HU and then normalized each raw CT case to $[0, 255.0]$. Random rotation of $[0, 15]$ is used as

Model	scale=1.0	scale=1.25	scale=1.5	scale=1.75	Avg. Dice
ResNet50-single-scale	-	-	-	-	35.96%
ResNet50-single-scale	-	-	-	-	48.14%
ResNet50-single-scale	-	-	-	-	47.71%
ResNet50-single-scale	-	-	-	-	46.29%
ResNet50-2-scale	-	-	-	-	52.75%
ResNet50-MSAN-2-scale	-	-	-	-	54.31%
ResNet50-3-scale	-	-	-	-	54.40%
ResNet50-MSAN-3-scale	-	-	-	-	55.61%
ResNet101-single-scale	-	-	-	-	37.53%
ResNet101-single-scale	-	-	-	-	46.38%
ResNet101-single-scale	-	-	-	-	52.67%
ResNet101-single-scale	-	-	-	-	54.56%
ResNet101-2-scale	-	-	-	-	54.98%
ResNet101-MSAN-2-scale	-	-	-	-	55.70%
ResNet101-3-scale	-	-	-	-	58.72%
ResNet101-MSAN-3-scale	-	-	-	-	59.89%
Zhou <i>et al.</i> [15]	-	-	-	-	50.15%
Yu <i>et al.</i> [14]	-	-	-	-	52.12%
3D-UNet [2]	-	-	-	-	40.81%

Table 1. DSC comparison of active bleed segmentation. ResNet101-MSAN-3-scale achieves the best performance of 59.89%, surpassing the prior art by more than 7%.

online data augmentation. A *poly* learning policy is applied with an initial learning rate of 0.05 with a decay power of 0.9. We follow [15, 11, 14] to use ImageNet pretrained model for initialization.

3.3 Results and Discussions

All results are summarized in Table 1, where we list thorough comparisons under different configuration of network architecture (*i.e.*, ResNet50 and ResNet101 [6]) and scales (*i.e.*, $scales = \{1.0, 1.25, 1.5, 1.75\}$). Note that we use larger scales (≥ 1.0) since our goal is to segment small targets. Under different settings, our method consistently outperforms others, indicating the effectiveness of MSAN.

Efficacy of multi-scale processing. As shown in Table 1, larger scales generally lead to better results. For instance, using ResNet50 as the backbone model, the performance under $scale = 1.0$ is $\sim 10\%$ lower than that under other larger scales. ResNet101-single-scale yields the best result of 54.56% under $scale = 1.75$, which is more than 17% better than using the scale of 1.0. These facts all indicate the efficacy of utilizing larger scales. Another observation is that the integration of more scales also leads to better segmentation quality than using just one scale. Using either ResNet50 or ResNet101 as the backbone, 3-scales always yield better results than 2-scales/single-scale, which shows that the learned knowledge from these different scales is complementary to each other. Therefore combining the information from these different scales can be beneficial for handling targets with a large variety of sizes, such as active bleed in our study.

Efficacy of the attentional module. Meanwhile, we also witness additional benefit from the attentional module. For instance, ResNet50-MSAN-3-scale observes an improvement of 1.17% compared with ResNet101-3-scale; ResNet101-MSAN-2-scale) observes an improvement of 0.72% compared with ResNet101-2-scale. A similar improvement can be also witnessed for ResNet-50. Three qualitative examples are shown in Fig. 3, where MSAN consistently outperforms other existing

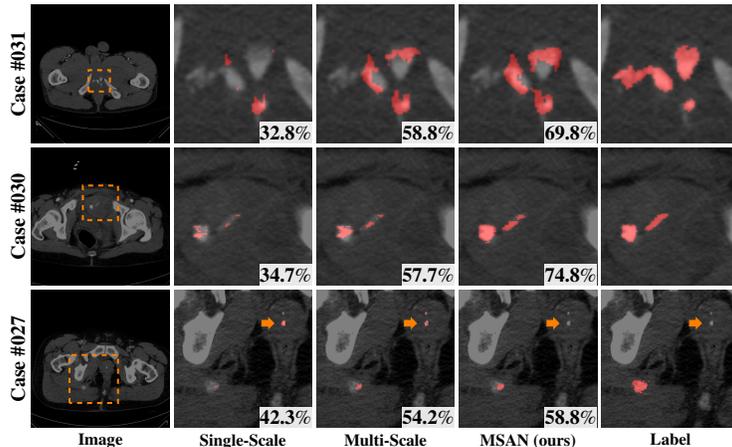


Fig. 3. Qualitative comparison of different methods. from left to right: original CT image, predictions of single-scale method ($scale = 1.50$), multi-scale method ($scale = \{1.25, 1.50, 1.75\}$), MSAN and the manual label. (Best viewed in color)

methods. For case 027, our MSAN successfully removes the outlier (indicated by the orange arrows) which is detected as false positives by other methods. This further justifies that the usage of attentional mechanisms can indeed refine the results and diminish non-trauma outliers.

Overall, our proposed MSAN observes a significant performance gain under different settings, which shows the generality and soundness of our approach. Additionally, we also compare our method with other state-of-art 3D segmentation methods including [15], [14] and [2]. Our method outperforms all these methods significantly (p -values for testing significant difference satisfy $p < 0.0001$), which further demonstrates the effectiveness of our approach. In order to further validate the generality and stability of MSAN, we directly test on a newly collected additional 15 cases without any retraining. Our method obtains an average DSC of 50.19%, whereas prior arts report 44.15% ([14]), 35.14% ([15]) and 27.32% ([2]). MSAN significantly outperforms these methods.

4 Conclusions

In this paper, we present Multi-Scale Attentional Network (MSAN), an end-to-end framework for automated segmentation of active hemorrhage from pelvic CT scans. Our proposed MSAN substantially improves the segmentation accuracy by more than 7% compared with prior arts. We note this framework can be practical in assisting radiologists for clinical applications, since the annotation in 3D volumes requires massive labor from radiologists.

Acknowledgements. This work was supported by NIBIB (National Institute of Biomedical Imaging and Bioengineering)/NIH under award number K08EB027141, University of Maryland Institute for Clinical and Translational Research Accelerated Translational Incubator Pilot (ATIP) award and Radiologic Society of North America (RSNA) Research Scholar Award #1605.

References

1. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *International Conference on Learning Representations* (2015)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: *MICCAI*. pp. 424–432. Springer (2016)
3. Cullinane, D.C., Schiller, H.J., Zielinski, M.D., Bilaniuk, J.W., Collier, B.R., Como, J., Holevar, M., Sabater, E.A., Sems, S.A., Vassy, W.M., et al.: Eastern association for the surgery of trauma practice management guidelines for hemorrhage in pelvic fracture update and systematic review. *Journal of Trauma and Acute Care Surgery* **71**(6), 1850–1868 (2011)
4. Davuluri, P., Wu, J., Tang, Y., Cockrell, C.H., Ward, K.R., Najarian, K., Hargraves, R.H.: Hemorrhage detection and segmentation in traumatic pelvic injuries. *Computational and mathematical methods in medicine* **2012** (2012)
5. Dreizin, D., Bodanapally, U., Boscak, A., Tirada, N., Issa, G., Nascone, J.W., Bivona, L., Mascarenhas, D., OToole, R.V., Nixon, E., et al.: Ct prediction model for major arterial injury after blunt pelvic ring disruption. *Radiology* **287**(3), 1061–1069 (2018)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. *CVPR* (2016)
7. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **37**(9), 1904–1916 (2015)
8. Kamnitsas, K., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Rueckert, D., Glocker, B.: Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *arXiv* (2016)
9. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. In: *CVPR* (2015)
10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI* (2015)
11. Roth, H., Lu, L., Farag, A., Sohn, A., Summers, R.: Spatial Aggregation of Holistically-Nested Networks for Automated Pancreas Segmentation. *MICCAI* (2016)
12. Sathy, A.K., Starr, A.J., Smith, W.R., Elliott, A., Agudelo, J., Reinert, C.M., Minei, J.P.: The effect of pelvic fracture on mortality after trauma: an analysis of 63,000 trauma patients. *JBJS* **91**(12), 2803–2810 (2009)
13. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: *CVPR* (2018)
14. Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E.K., Yuille, A.L.: Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In: *CVPR*. pp. 8280–8289 (2018)
15. Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E.K., Yuille, A.L.: A fixed-point model for pancreas segmentation in abdominal ct scans. In: *MICCAI*. pp. 693–701. Springer (2017)
16. Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., Du, N., Fan, W., Xie, X.: Anatomicnet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics* **46**(2), 576–589 (2019)