PAN: Projective Adversarial Network for Medical Image Segmentation

Naji Khosravan[2], Aliasghar Mortazi[2], Michael Wallace[1], Ulas Bagci[2]

[1] Mayo Clinic Cancer Center, Jacksonville, FL.
[2] Center for Research in Computer Vision (CRCV),
School of Computer Science, University of Central Florida, Orlando, FL.

Abstract. Adversarial learning has been proven to be effective for capturing long-range and high-level label consistencies in semantic segmentation. Unique to medical imaging, capturing 3D semantics in an effective yet computationally efficient way remains an open problem. In this study, we address this computational burden by proposing a novel projective adversarial network, called PAN, which incorporates high-level 3D information through 2D projections. Furthermore, we introduce an attention module into our framework that helps for a selective integration of global information directly from our *segmentor* to our adversarial network. For the clinical application we chose pancreas segmentation from CT scans. Our proposed framework achieved state-of-the-art performance without adding to the complexity of the segmentor.

Keywords: Object Segmentation · Deep Learning · Adversarial Learning · Attention · Projective · Pancreas.

1 Introduction

Segmentation has been a major area of interest within the fields of computer vision and medical imaging for years. Owing to their success, deep learning based algorithms have become the standard choice for semantic segmentation in the literature. Most state-of-the-art studies model segmentation as a pixel-level classification problem [2–4]. Pixel-level loss is a promising direction but, it fails to incorporate global semantics and relations. To address this issue researchers have proposed a variety of strategies. A great deal of previous research uses a post-processing step to capture pairwise or higher level relations. Conditional Random Field (CRF) was used in [2] as an offline post-processing step to modify edges of objects and remove false positives in CNN output. In other studies, to avoid offline post-processing and provide an end-to-end framework for segmentation, mean-field approximate inference for CRF with Gaussian pairwise potentials was modeled through Recurrent Neural Network (RNN) [17].

In parallel to post processing attempts, another branch of research tried to capture this global context through multi-scale or pyramid frameworks. In [2–4], several spatial pyramid pooling at different scales with both conventional convolution layers and *Atrous* convolution layers were used to keep both contextual

and pixel-level information. Despite such efforts, combining local and global information in an optimal manner is not a solved problem, yet.

Following by the seminal work by Goodfellow et.al. in [7] a great deal of research has been done on adversarial learning [8, 10, 14, 15]. Specific to segmentation, for the first time, Luc et. al. [8] proposed the use of a discriminator along with a segmentor in an adversarial min-max game to capture long-range label consistencies. In another study $SegAN$ was introduced, in which the segmentor plays the role of generator being in a min-max game with a discriminator with a multi-scale $L1$ loss [14]. A similar approach was taken for structure correction in chest X-rays segmentation in [5]. A conditional GAN approach was taken in [10] for brain tumor segmentation.

In this paper, we focused on the challenging problem of pancreas segmentation from CT images, although our framework is generic and can be applied to any 3D object segmentgation problem. This particular application has unique challenges due to the complex shape and orientation of pancreas, having low contrast with neighbouring tissues and relatively small and varying size. Pancreas segmentation were studied widely in the literature. Yu et al. introduced a recurrence saliency transformation network, which uses the information from previous iteration as a spatial weight for current iteration [16]. In another attempt, U-Net with an attention gate was proposed in [9]. Similarly, a two-cascaded-stage based method was used to localize and segment pancreas from CT scans in [13]. A prediction-segmentation mask was used in [18] for constraining the segmentation with a coarse-to-fine strategy. Furthermore, a segmentation network with RNN was proposed in [1] to capture the spatial information among slices. The unique challenges of pancreas segmentation (complex shape and small organ) shifted the literature towards methods with coarse-to-fine and multi-stage frameworks, promising but computationally expensive.

Summary of our contributions: The current literature on segmentation fails to capture 3D high-level shape and semantics with a low-computation and effective framework. In this paper, for the fist time in the literature, we propose a projective adversarial network (PAN) for segmentation to fill this research gap. Our method is able to capture 3D relations through 2D projections of objects, without relying on 3D images or adding to the complexity of the segmentor. Furthermore, we introduce an attention module to selectively integrate high-level, whole-image features from the *segmentor* into our adversarial network. With comprehensive evaluations, we showed that our proposed framework achieves the state-of-the-art performance on publicly available CT pancreas segmentation dataset [11] even when a simple encoder-decoder network was used as *segmentor*.

2   Method

Our proposed method is built upon the adversarial networks. The proposed framework's overview is illustrated in Figure 1. We have three networks: a segmentor (S in Figure 1), which is our main network and was used during the test phase, and two adversarial networks (D_s and D_p in Figure 1), each with a

specific task. The first adversarial network (D_s) captures high-level *spatial* label contiguity while the second adversarial network (D_p) enforces the *3D semantics* through a 2D projection learning strategy. The adversarial networks were used only during the training phase to boost the performance of the segmentor without adding to its complexity.
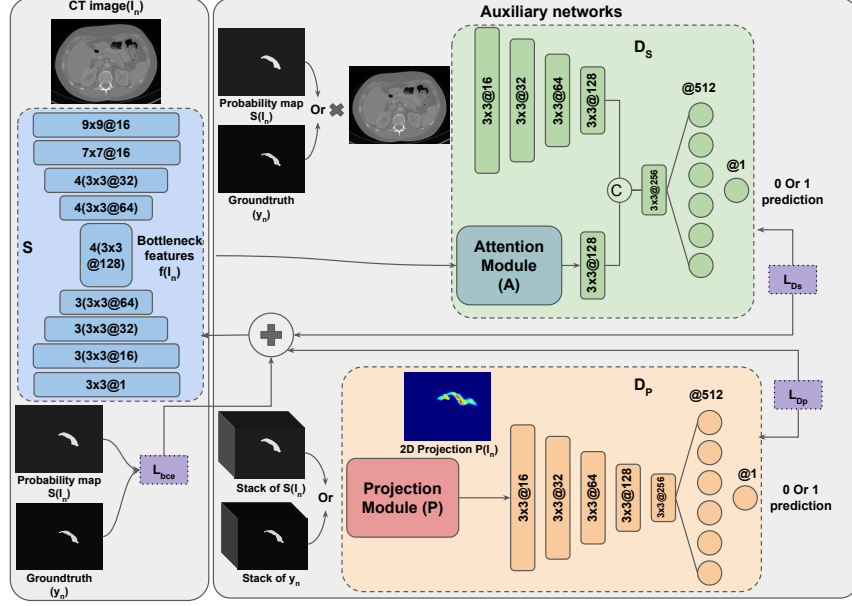

Fig. 1. The proposed framework consists of a segmentor S and two adversarial networks, D_s and D_p. S was trained with a hybrid loss from D_s, D_p and the ground-truth.

2.1   Segmentor (S)

Our base network is a simple fully convolutoinal network with an encoder-decoder architecture. The input to the segmentor is a 2D grey-scale image and the output is a pixel-level probability map. The probability map shows probability of presence of the object at each pixel. We use a hybrid loss function (explained in details in Section 2.3) to update the parameters our segmentor (S). This loss function is composed of three terms enforcing: (1) pixel-level high-resolution details, (2) spatial and high-range label continuity, (3) 3D shape and semantics, through our novel projective learning strategy.

As can be seen in Figure 1, the segmentor contains 10 conv layers in the encoder, 10 conv layers in the decoder and 4 conv layers as the bottleneck. The last conv layer is a 1×1 conv layer with the channel output of 1, combining channel-wise information in the highest scale. This layer is followed by a sigmoid function to create the notion of probability.

2.2   Adversarial Networks

Our adversarial networks are designed with the goal of compensating for the missing global relations and correcting higher-order inconsistencies, produced by a single pixel-level loss. Each of these networks produces an adversarial signal and apply it to the segmentor as a term in the overall loss function (Equation 2). The details of each network is described below:

Spatial semantics network (D_s): This network is designed to capture spatial consistencies within each frame. The input to this network is either the segmented object by the ground-truth or by the segmentor's prediction. The Spatial semantics network (D_s) is trained to discriminate between these two inputs with a binary cross-entropy loss, formulated as in Equation 4. The adversarial signal produced by the negative loss of D_s to S forces S to produce predictions closer to ground-truth in terms of spatial semantics.

As illustrated in Figure 1 top right, D_s has a two-branch architecture with a late fusion. The top branch processes the segmented objects by ground-truth or segmentor's prediction. We propose an extra branch of processing, getting the bottleneck features corresponding to the original gray-scale input image, and passing them to an attention module for an information selection. The processed features are then concatenated with the first branch and passed through the shared layers. We believe that having the high-level features of whole image along with the segmentations improves the performance of D_s.

Our attention module learns where to attend in the feature space to have a more discriminative information selection and processing. The details of the attention module are described in the following.

Attention module (A): We feed the high-level features form the segmentor's bottleneck to D_s. These features contain global information about the whole frame. We use a soft-attention mechanism, in which our attention module assigns a weight to each feature based on its importance for discrimination. The attention module gets the features with shape $w \times h \times c$, as input, and outputs a weight set with a shape of $w \times h \times 1$. A is

Fig. 2. Attention module assigns a weight to each feature allowing for a soft selection of information.

composed of two 1×1 convolution layers followed by a softmax layer (Figure 2). The softmax layer introduces the notion of *soft selection* to this module. The output of A is then multiplied to the features before being passed to the rest of the network.
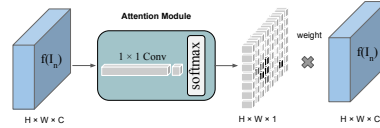
Projective network (D_p): Any 3D object can be projected into 2D planes from specific viewpoints, resulting in multiple 2D images. The 2D projection contains 3D semantics information, to be retrieved. In this section, we introduce our projective network (D_p). The main task of D_p is to capture 3D semantics

without relying on 3D data and from the 2D projections. Inducing 3D shapes form 2D images has previously been done for 3D shape generation [6]. Unlike existing notions, however, in this paper we propose 3D semantics induction from 2D projections, to benefit segmentation for the first time in the literature.

The projection module (P) projects a 3D volume (V) on a 2D plane as:

$$P((i,j),V) = 1 - \exp^{-\sum_k V(i,j,k)},\tag{1}$$

where each pixel in the 2D projection $P((i,j),V)$ gets a value in the range of $[0,1]$ based on the number of voxel occupancy in the third dimension of corresponding 3D volume (V). For the sake of simplicity, we refer to the projection of a 3D volume V as $P(V)$. We pass each 3D image through our segmentor (S) slice by slice and stack the corresponding prediction maps. Then, these maps are fed to the projection module (P) and are projected in the axial view.

The input to D_p is either the projected ground-truth or projected prediction map produced by S. D_p is trained to discriminate these inputs using the loss function defined in Equation 5. The adversarial term produced by D_p in Equation 2 forces S to create predictions which are closer to ground-truth in terms of 3D semantics. Incorporating D_p as an adversarial network to our segmentation framework helps S to capture 3D information through a very simple 2D architecture and without adding to its complexity in the test time.

2.3   Adversarial training

To train our framework, we use a hybrid loss function, which is a weighted sum of three terms. For a dataset of N training samples of images and ground truths (I_n, y_n), we define our hybrid loss function as:

$$l_{hybrid} = \sum_{n=1}^{N} l_{bce}(S(I_n), y_n) - \lambda l_{D_s} - \beta l_{D_p},\tag{2}$$

where l_{D_s} and l_{D_p} are the losses corresponding to D_s and D_p and $S(I_n)$ is the segmentor's prediction. The first term in Equation 2 is a weighted binary cross-entropy loss. This loss is the state-of-the-art loss function for semantic segmentation and for a grey-scale image I with size $H \times W \times 1$ is defined as:

$$l_{bce}(\hat{y}, y) = - \sum_{i=1}^{H \times W} (wy_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)),\tag{3}$$

where w is the weight for positive samples, y is the ground-truth label and $\hat{y}$ is the network's prediction. Equation 3 encourages S to produce predictions similar to ground-truth and penalizes each pixel *independently*. High-order relations and semantics cannot be captured by this term.

To account for this drawback, the second and third terms are added to train our auxiliary networks. l_{D_s} and l_{D_p} are defined below, respectively:

$$l_{D_s} = l_{bce}(D_s(I_n, y_n), 1) + l_{bce}(D_s(I_n, S(I_n)), 0), \tag{4}$$

$$l_{D_p} = l_{bce}(D_p(P_{I_n}, Py_n), 1) + l_{bce}(D_p(P_{I_n}, P_{S(I_n)}), 0). \tag{5}$$

Here P is the projection module, l_{bce} is the binary cross-entropy loss with $w = 1$ in Equation 3 corresponding to a single number (0 or 1) as the output.

3   Experiments and Results

We evaluated the efficacy of our proposed system with the challenging problem of pancreas segmentation. This particular problem was selected due to the complex and varying shape of pancreas and relatively more difficult nature of the segmentation problem compared to other abdominal organs. In our experiments we show that our proposed framework outperforms other state-of-the-art methods and captures the complex 3D semantics with a simple encoder-decoder. Furthermore, we have created an extensive comparison to some baselines, designed specifically to show the effects of each block of our framework.

Data and evaluation: We used the publicly available TCIA CT dataset from NIH [11]. This dataset contains a total of 82 CT scans. The resolution of scans is $512 \times 512 \times Z$, $Z \in [181, 466]$ is the number of slices in the axial plane. The voxels spacing ranges from $0.5mm$ to $1.0mm$. We used a randomly selected set of 62 images for training and 20 for testing to perform a 4-fold cross-validation. Dice Similarity Coefficient (DSC) is used as the metric of evaluation.

Comparison to baselines:

To show the effect of each building block of our framework we designed an extensive set of experiments. In our experiments we start from only training a single segmentor (S) and go to our final proposed framework. Furthermore, we show comparison of encoder-decoder architecture with other state-of-the-art semantic segmentation architectures.

Table 1 shows the results adding of each building block of our framework. The

	Model	DSC%
	Encoder-decoder (S)	57.7
	Atrous pyramid	48.2
1-fold	$S + D_s$	85.0
	$S + D_s + A$	85.9
	$S + D_s + A + D_p$	**86.8**

Table 1. Comparison with baselines.

eccoder-decoder architecture is the one showed in Figure 1 as S, while the Atrous pyramid architecture is similar to the recent work of [4]. This architecture is currently state-of-the-art for semantic segmentation. In which an Atrous pyramid is used to capture global context. We added an Atrous pyramid with 5 different scales: 4 Atrous convolutions at rates of $1, 2, 6, 12$, with the global image pooling. We also replaced the decoder with 2 simple upsampling and conv layers similar to the main paper [4]. We refer the readers to the main paper for more details about this architecture due to space limitations [4]. We found out having an extensive processing in the decoder improves the results compared to the Atrous pyramid architecture (possibly a better choice for segmentation of objects at multiple scales). This is because our object of interest is relatively small.

Moreover, we showed that adding a spatial adversarial notwork (D_s) can boost the performance of S dramatically, in our task. Introducing attention (A) helps for a better information selection (as described in section 2.2) and boosts the performance further. Finally, our best results is achieved by adding the projective adversarial network (D_p), which adds integration of 3D semantics into the framework. This supports our hypothesis that our segmentor has enough capacity in terms of parameters to capture all this information and with proper and explicit supervision can achieve state-of-the-art results.

Comparison to the state-of-the-art: We provide the comparison of our method's performance with current state-of-the-art literature on the same TCIA CT dataset for pancreas segmentation. As can be seen from experimental validation, our method outperforms the state-of-the-art with dice scores, provides better efficiency (less computational burden). Of a note, the proposed algorithm's least achievement is consistently higher than the state of the art methods.

	Approach	Average DSC%	Max DSC%	Min DSC%
	Roth et al. [11]	71.42 ± 10.11	86.29	23.99
	Roth et al. [12]	78.01 ± 8.20	88.65	34.11
	Roth et al. [13]	81.27 ± 6.27	88.96	50.69
4-fold cross validation	Zhou et al. [18]	82.37 ± 5.68	90.85	62.43
	Cai et al. [1]	82.40 ± 6.70	90.10	60.00
	Yu et al. [16]	84.50 ± 4.97	**91.02**	62.81
	Ours	$\mathbf{85.53 \pm 1.23}$	88.71	**83.20**

Table 2. Comparison with state-of-the-art on TCIA dataset.

4   Conclusion

In this paper we proposed a novel adversarial framework for 3D object segmentation. We introduced a novel projective adversarial network, inferring 3D shape and semantics form 2D projections. The motivation behind our idea is that integration of 3D information through a fully 3D network, having all slices as input, is computationally infeasible. Possible workarounds are: 1)down-sampling the data or 2)sacrificing number of parameters, which are sacrificing information or computational capacity, respectively. We also introduced an attention module to selectively pass whole-frame high-level feature from the segmentor's bottleneck to the adversarial network, for a better information processing. We showed that with proper and guided supervision through adversarial signals a simple encoder-decoder architecture, with enough parameters, achieves state-of-the-art performance on the challenging problem of pancreas segmentation. We achieved a dice score of **85.53%**, which is state-of-the art performance on pancreas segmentation task, outperforming previous methods. Furthermore, we argue that our framework is general and can be applied to any 3D object segmentation problem and is not specific to a single application.

References

1. Cai, J., Lu, L., Xie, Y., Xing, F., Yang, L.: Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function. arXiv preprint arXiv:1707.04912 (2017)
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. IEEE transactions on pattern analysis and machine intelligence 40(4), 834–848 (2018)
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 801–818 (2018)
5. Dai, W., Dong, N., Wang, Z., Liang, X., Zhang, H., Xing, E.P.: Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 263–273. Springer (2018)
6. Gadelha, M., Maji, S., Wang, R.: 3d shape induction from 2d views of multiple objects. In: 2017 International Conference on 3D Vision (3DV). pp. 402–411. IEEE (2017)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
8. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)
9. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
10. Rezaei, M., Harmuth, K., Gierke, W., Kellermeier, T., Fischer, M., Yang, H., Meinel, C.: A conditional adversarial network for semantic segmentation of brain tumor. In: International MICCAI Brainlesion Workshop. pp. 241–252. Springer (2017)
11. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 556–564. Springer (2015)
12. Roth, H.R., Lu, L., Farag, A., Sohn, A., Summers, R.M.: Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 451–459. Springer (2016)
13. Roth, H.R., Lu, L., Lay, N., Harrison, A.P., Farag, A., Sohn, A., Summers, R.M.: Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Medical image analysis 45, 94–107 (2018)
14. Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X.: Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. Neuroinformatics 16(3-4), 383–392 (2018)
15. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. arXiv preprint arXiv:1809.07294 (2018)

16. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8280–8289 (2018)
17. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1529–1537 (2015)
18. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 693–701. Springer (2017)