OBJECT BOUNDING BOX-CRITIC NETWORKS FOR OCCLUSION-ROBUST OBJECT DETECTION IN ROAD SCENE

Jung Uk Kim[1†], *Jungsu Kwon*[1†], *Hak Gu Kim*[†], *Haesung Lee*[‡], *and Yong Man Ro*[*†]

[†]Image and Video Systems Lab, School of Electrical Engineering, KAIST, Republic of Korea
[‡]KEPCO Research Institute

ABSTRACT

Object detection in a road scene has received a significant attention from research fields of developing autonomous vehicle and automatic road monitoring systems. However, object occlusion problems frequently occur in generic road scenes. Due to such occlusion problems, previous object detection methods have limitations of not being able to detect objects accurately. In this paper, we propose a novel object detection network which is robust in occlusions. For effective object detection even with occlusion, the proposed network mainly consists of two parts; 1) Object detection framework, 2) Multiple object bounding box (OBB)-Critic network for predicting a BB map which estimates both object region and occlusion region. Comprehensive experimental results on a KITTI Vision Benchmark Suite dataset showed that the proposed object detection network outperformed the state-of-the-art methods.

Index Terms— Object detection, adversarial learning, actor-critic network, plug-in, occlusion

1. INTRODUCTION

Object detection is one of the most studied problems in computer vision [1-3]. It learns a visual model of each object and finds an appropriate object category and bounding-box area. Recently, object detection in driving environment has attracted increasing interest from research fields and industry. Automatically detecting the objects such as cars and pedestrians on roads allows the driver to be aware of the road condition, traffic information, and so on. For this reason, object detection can be utilized for various applications such as autonomous vehicle [4] and automatic surveillance systems [5]. However, there are many challenges in development of reliable object detection method in driving environment: object occlusion problems [6], large variances of scale [7], etc. In particular, occlusion caused by parked cars, passing vehicles, and pedestrian is one of the most critical factors on a road. Such occlusion can make object detection on a road difficult.

Recently, with the advent of the deep convolutional neural networks (CNNs), performance of object detection has been improved significantly compared to previous hand-craft based methods [8, 9]. Basic methods of object detection using CNNs were those of region-based model [10, 11]. [10, 11] used selective search algorithm [12] to extract candidates regions. In [10], each candidate region was extracted from the image and it passed independent CNNs. In [11], each candidate region was extracted from the encoded feature map of a single CNN. Based on the candidate regions extracted from [10, 11], object classification and localization were performed. However, since the selective search algorithm [12] is a hand-craft method, it is sensitive to parameter tuning. In addition, many inferences (~2,000 proposals) have to be computed which slow down the object detection network. In [13], deep learning based region proposal network (RPN) was applied to extract object candidate regions instead of selective search algorithm [12]. In RPN, pre-defined anchor predicts object region of interests (RoI). By introducing RPN, fast and high performance were achieved.

Recently, several object detection methods in the driving environment based on RPN were introduced [14, 15]. In [14], to predict the presence of certain subcategories at a specific scale and location, scene subcategory information such as 2D pose, 3D pose, and 3D shape were encoded in the RPN. In [15], multi-task CNNs and RoI voting were introduced to detect various types of objects through the RPN. However, these were limited to partial occlusions and were difficult to apply of severe occlusions.

In this paper, we propose a novel deep learning based object detection network that is robust in occlusions in road scenes. For object detection network learning, we apply multiple actor-critic network [16] in a “plug-in” manner for robust object detection in occlusions. We introduce multiple actors to estimate object region and occlusion region in the form of bounding-box map (referred to as BB map) for encoding latent object and occlusion feature. Multiple actor consists of global and local actor. Global actor differentiates multiple objects from an input image and local actor refines the bounding-box of each object by making use of occlusion information in the RoI. In addition, predicted BB map con-

[1] Both authors are equally contribution
* Corresponding author (ymro@ee.kaist.ac.kr)

978-1-4799-7061-2/18/$31.00 ©2018 IEEE

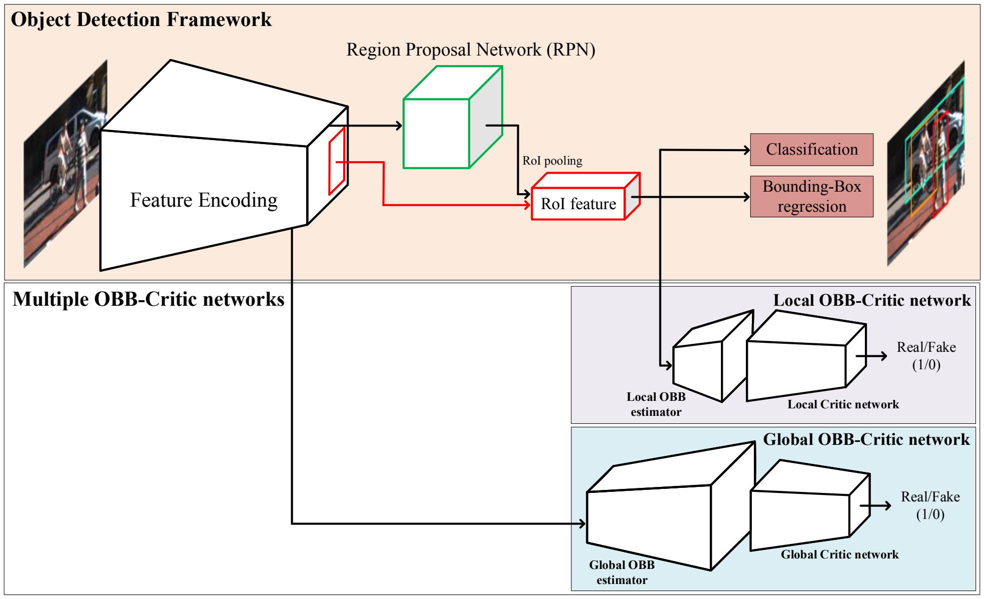

Fig. 1. An overview of the proposed object detection network. It contains object detection framework and multiple OBB-Critic networks. In the training phase, multiple OBB-Critic networks are plugged-in to the object detection framework. In the testing phase, only object detection network is evaluated.

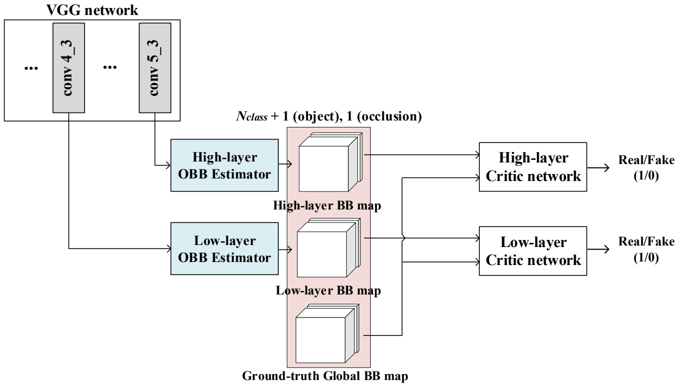

Fig. 2. Proposed global OBB-Critic network.

tains all object-BB maps and an occlusion-BB map. Also we introduce corresponding multiple critic networks. Those determine whether a predicted BB map predicted based on the multiple actor and each ground-truth are real or fake. We combine the adversarial approach [21] of the generative adversarial networks [17] (GAN) with multiple actor-critic networks to improve the performance of both networks by learning competitively.

The rest of this paper is organized as follows. In section 2, we describe the proposed deep learning based object detection network with multiple OBB-Critic networks. In section 3, the experimental results are presented to verify the performance of the proposed object detection networks. Finally, conclusions are drawn in section 4.

2. PROPOSED METHOD

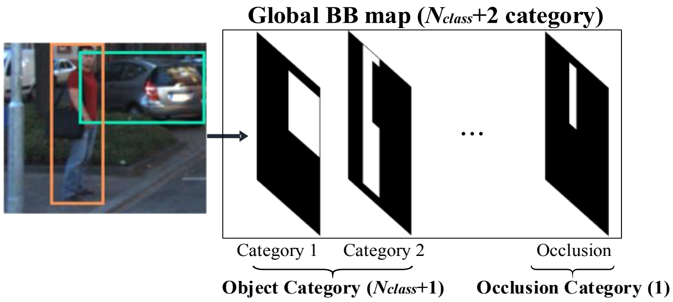
Global BB map (N_{class}+2 category)
...
Category 1
Category 2
Occlusion
Object Category (N_{class}+1)
Occlusion Category (1)

Fig. 3. Example of global BB map

Fig. 1 shows the overall procedure of the proposed object detection network. As shown in the figure, the proposed object detection network consists of the two modules. First module is object detection framework. It performs classification and bounding box regression. We use the VGG 16 [19] network in the feature encoding part of the object detection framework. Second module is multiple OBB-Critic networks. These module consist of global and local OBB actor and corresponding two critic networks. We call OBB actor as OBB estimator and it divides object areas and occlusion areas.

In the training phase, multiple OBB-Critic networks are trained in a "plug-in" manner to the object detection framework. In the testing phase, only object detection framework is used. Detailed descriptions of each part are given in the following sections.

2.1. Adversarial learning for global OBB-Critic network

Fig. 2 shows the proposed global OBB-Critic network. It consists of global OBB estimator and corresponding global critic network. Global OBB estimator globally estimates the global BB map. Fig. 3 shows an example of the global BB map. As shown in figure, the global BB map containes object BB map and occlusion BB map and has (N_{class}+2) channel. (N_{class}+1) is object categories and background and 1 is the occlusion category. In the object BB map, internal bounding-box area of each object is assigned to 1. In the occlusion BB map, overlapping inside region of object bounding-boxes is assigned to 1.

Global OBB estimator consists of high and low-layer OBB estimators. Two OBB estimators have same architecture as decoder part of the U-net [18]. As the resolution doubles, the feature map is concatenated to the same resolution feature map of VGG 16 network. Using this architecture, we estimate BB map. Inputs of high and low-layer OBB estimator are conv5_3 and conv 4_3 feature map of VGG16, respectively. Outputs are high and low-layer BB maps, respectively. After high and low-layer BB maps are estimated, global critic network (D^{Global}) discriminates whether high and low-layer BB maps are close to real (i.e., ground-truth) or fake (i.e., estimated). Global critic networks have same architecture to the critic network of [23].

By competitively learning between estimators and critic networks in an adversarial manner, global OBB estimator can better predict object BB map and occlusion BB map and feature encoding part of the object detection framwork encodes latent object and occlusion feature to aware of the object region and occlusion region.

2.2. Adversarial learning for local OBB-Critic network

Local OBB-Critic network refines the bounding-box of RoI considering occlusion information. Input of the local OBB estimator (S^{Local}) is RoI feature which is extracted from the RPN. Local OBB estimator performs 1×1 convolution with RoI feature. Output of local OBB estimator is local BB map which is same resolution of RoI feature. Like the global BB map, the number of chnnel is (N_{class}+2). The object BB map and occlusion BB map of local BB map is the internal bounding-box area and overlapping area of RoI.

Predicted local BB map and ground-truth local BB map pass through local critic network (D^{Local}). Local critic network distinguishes two BB maps as real or fake. It has same structure as the global critic network. Similar to global OBB-Critic network,RoI feature can be aware of the specific RoI object region and occlusion region by plugging-in local OBB-Critic network to the object detection framework.

2.3. Training objectives

In the proposed method, multiple OBB-Critic networks are trained jointly through a mini-max scheme that alternates optimizing OBB estimator and corresponding critic network. OBB estimator is trained after learning critic network. First, the loss of critic network can be written as

$$L_{Critic} = \lambda_{Global-Critic} L_{Global-Critic} + \lambda_{Local-Critic} L_{Local-Critic}, \quad (1)$$

where $L_{Global\text{-}Critic}$ and $L_{Local\text{-}Critic}$ are global and local critic network losses, respectively. $\lambda_{Global-Critic}$ and $\lambda_{Local-Critic}$ are hyper-parameters to control crtic network loss Each loss function can be written as

$$L_{Global-Critic} = \sum_{t=1}^{2} (J_D(D^{Global}(x_i, S^t(x_i)),0) + J_D(D^{Global}(x_i, y_i^{Global}),1)), \quad (2)$$

$$L_{Local-Critic} = J_D(D^{Local}(x_i, S^{Local}(x_i)),0) + J_D(D^{Local}(x_i, y_i^{Local}),1), \quad (3)$$

where $J_D(\hat{t}, t) = -t \ln \hat{t} + (1-t)\ln(1-\hat{t})$ is binary logistic loss for critic prediction. x_i is i-th input image. Ground-truth global and local BB maps are y_i^{Global} and y_i^{Local}, respectively. $S^t(x_i)$ is the i-th predicted low or high-layer BB map. We train two critic networks by minimizing the Eq. (1) with respect to D^{Global} and D^{Local} for a fixed $S^t(\mathrm{x_i})$ and $S^{Local}(x_i)$.

Table 1. Criteria for dividing the three levels in the [20]

Difficulty	Easy	Moderate	Hard
Min Height (pixel)	40	25	25
Occlusion	Fully Visible	Partially Occluded	Difficult to see
Truncation (%)	15	30	50

Second, when training OBB estimator, learning is performed with the object detection framework. Total loss function can be written as

$$L_{Total} = L_{Estimator} + \lambda_{OD} L_{OD}, \quad (4)$$

where $L_{Estimator}$ and L_{OD} is loss of estimator and the object detection framework, respectively. λ_{OD} is hyper-parameter to control the object detection framework. Each loss function can be written as

$$L_{Estimator} = \sum_{t=1}^{2} \lambda_{Global} J_D(D^{Global}(x_i, S^t(x_i)),1) + \lambda_{Local} J_D(D^{Local}(x_i, y_i^{Local}),1)), \quad (5)$$

$$L_{OD} = L_{RPN-CLS} + L_{RPN-REG} + L_{CLS} + L_{REG} + L_{Global} + L_{Local}, \quad (6)$$

where λ_{Global} and λ_{Local} are hyper-parameters to control estimator loss term. $L_{RPN\text{-}CLS}$, $L_{RPN\text{-}REG}$, L_{CLS}, and L_{REG} are classification and regression losses of the RPN and the fully connected layer of the RoI feature, respectively. L_{Global} and L_{Local} are the losses of global and local OBB estimator. We used cross-entropy loss for classification loss and OBB estimator loss and smooth L_1 loss for regression loss.

3. EXPERIMENTS

3.1. Experimental setup

In our experiments, we used object detection framework as Faster R-CNN [13] which is state-of-the-art method and evaluated proposed method on the KITTI Vision Benchmark Suite dataset [20]. It consisted of videos in road scene with 7,481 images and had three classes; car, pedestrian, and cy-

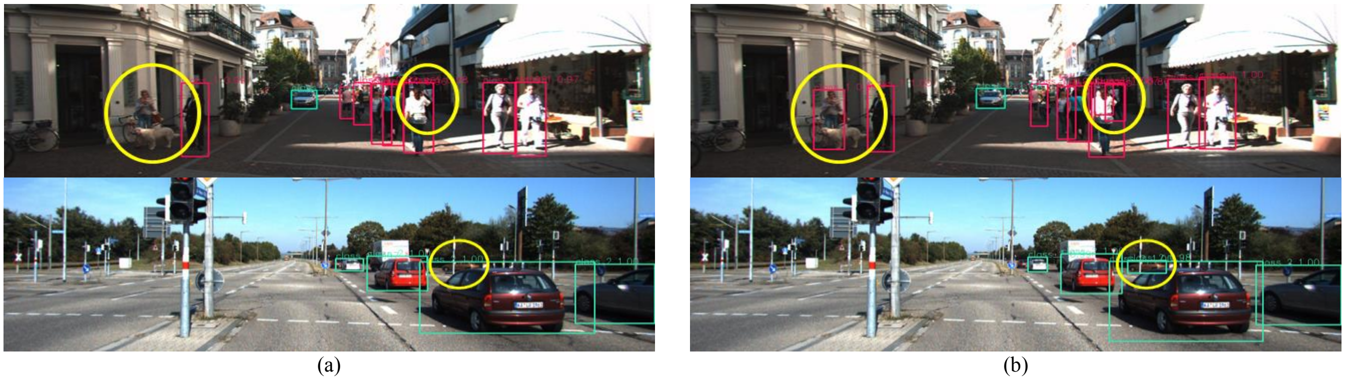

Fig. 3. Examples of the detection results of Faster R-CNN [13] and the proposed method. (a) Result of Faster R-CNN, (b) Result of the proposed method.

Table 2. Experimental results of the KITTI object detection benchmark in three classes (Car, Pedestrian, and Cyclist)

	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Faster R-CNN [13]	0.829	0.778	0.663	0.833	0.684	0.626	0.564	0.464	0.428
X. Yuan [22]	0.805	0.679	0.582	N/A	N/A	N/A	N/A	N/A	N/A
W. Chu [15]	0.908	0.907	0.843	N/A	N/A	N/A	N/A	N/A	N/A
Y. Xiang [14]	**0.951**	0.852	0.721	0.859	0.685	0.625	0.710	0.559	0.517
Proposed method	0.950	**0.910**	**0.863**	**0.862**	**0.695**	**0.655**	**0.730**	**0.666**	**0.630**

cyclist. We divided 3,682 images and 3,799 images into training and testing in the same way as [14]. The training images and the test images were taken different video. We used input images with 384×1248 sized image. Output size of global BB maps were 96×312. We used 9 types of the anchors that are same aspect ratios and scales of RPN in the [13]. We used ensembles of the proposed networks.

We evaluated the detection performance into three different levels following [20]; easy, moderate, and hard. The criteria for dividing the three three levels were shown in the Table 1. Following [20], we used area under precision-recall curve (AUC) to compare numerically. To measure AUC, overlap thresholds for car, pedestrian, and cyclist were 70%, 50%, and 50%, respectively. If the IoU overlap between ground-truth bounding box and predicted bounding-box was higher than overlap thresholds, it was considered as true positive.

3.2. Experimental results

To show the effectiveness of the proposed multiple OBB-Critic network, we compared proposed method with the baseline Faster R-CNN [13]. Fig. 2 showed the detection results of the two methods. Fig. 2 (a) and (b) showed Faster R-CNN and proposed method, respectively. As shown in the Fig. 2 (b), proposed method detected occluded object more robust than Faster R-CNN (shown by yellow circles). Small objects hidden by a large objects missed in the Faster R-CNN. However these were detected in the proposed method.

We numerically compared with Faster R-CNN and two deep learning based state-of-the-art methods. As shown in Table 2, the propose method were outperformed than Faster R-CNN in all cases and two state-of-the-art deep learning methods except for easy in the car category in [14]. Unlike [14], additional 3D information were not used the proposed method In the pedestrian category, performaces of the proposed method were (0.862, 0.695, 0.655), Faster R-CNN were (0.833, 0.684, 0.626), and [14] were (0.859, 0.685, 0.625) in the three levels. Especially, in the cyclist category, performace of the proposed method were (0.730, 0.666, 0.630), Faster R-CNN were (0.564, 0.464, 0.428) and [14] were (0.710, 0.559, 0.517) in the three levels. When we compared with Faster R-CNN , it verified that when multiple OBB-Critic network were plugged-in to Faster R-CNN in the training phase, Faster R-CNN detected object more robust. In addition, when we compared with the state-of-the-art deep learning methods, it verified that considering occlusion was robust than adding additional 3D informations to RPN.

4. CONCLUSIONS

In this paper, we proposed a novel object detection network which considers occlusions in road scenes. To deal with occlusions, we proposed the multiple OBB-Critic network in a "plug-in" manner. In the proposed multiple OBB estimator, BB maps were predicted for feature encoding network to encode latent feature of objects and ab occlusion. Corresponding critic networks determined whether the predicted BB map is real or fake. To effectively cope with occlusions, we adopt OBB-Critic network in an adversarial manner. The experimental results showed that the proposed method outperformed baseline object detection framework and state-of-the art methods.

5. ACKNOWLEDGEMENT

This research was supported by Development of Deep Learning based Electric Equipment State Inference and Maintenance Technologies of the Korea Electric Power Research Institute (KEPRI) funded by the Korea Electric Power Corporation (KEPCO).

6. REFERENCES

[1] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Computer Vision and Patter Recognition (CVPR)*, pp. 2117-2125, 2017.

[2] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," *arXiv preprint arXiv: 1605.06409*, 2016.

[3] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. "Scalable, high-quality object detection," *arXiv preprint arXiv: 1412.1441*, 2014.

[4] X. Wen, L. Shao, W. Fang, and Y. Xue, "Efficient feature selection and classification for vehicle detection," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 508-517, 2014.

[5] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," in *Multimedia Tools and Applications*, vol. 68, pp. 5-21, 2012.

[6] T. Wu, B. Li, and S. C. Zhu, "Learning and-or model to represent context and occlusion for car detection and viewpoint estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1829-1843, 2016.

[7] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. European Conf. Computer Vision (ECCV),* pp. 354-370, 2016.

[8] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 13-24, 2013.

[9] H. Azizpour, and I. Laptev, "Object detection using strongly-supervised deformable part models," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 836-849, 2012.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 580-587, 2014.

[11] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 1440-1448, 2015.

[12] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object detection," *Int. Journals of Computer Visions*, vol. 104, no. 2, pp. 154-171, 2013.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[14] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," *arXiv preprint arXiv:1604.04693*, 2016.

[15] W. Chu, Y. Liu, C. Shen, D. Cai, and X. S. Hua, "Multi-task vehicle detection with region-of-interest voting," *IEEE Trans. Image Processing*, vol. 27, no. 1, pp. 432-441, 2018.

[16] V. R. Konda, J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. Neural Information Processing Systems (NIPS)*, pp. 1008-1014, 2000.

[17] I. J. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D W. Farley, S. Ozair, et al., "Generative adversarial nets," in *Proc. Neural Information Processing Systems (NIPS)*, pp. 2672-2680, 2014.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *arXiv preprint arXiv: 1505.04597*, 2015.

[19] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv: 1409.1556*, 2014.

[20] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3354-3361, 2012.

[21] D. Pfau, and O. Vinyals, "Connecting generative adversarial networks and actor-critic methods," *arXiv preprint arXiv: 1610.01945*, 2016.

[22] X. Yuan, X. Cao, X. Hao, H. Chen, X. Wei, "Vehicle detection by a context-aware multichannel feature pyramid," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1348-1357, 2016.

[23] W. Dai, J. Doyle, X. Liang, H. Zhang, N. Dong, Y. Li, and E. P. Xing, "Scan: structure correcting adversarial network for organ segmentation in chest x-rays," *arXiv preprint arXiv: 1703.08770*, 2017.