Skipped-Hierarchical Feature Pyramid Networks for Nuclei Instance Segmentation

Hyekyoung Hwang, Toan Duc Bui, Sang-il Ahn, Jitae Shin*
SungKyunKwan Univ. Department of Electrical and Computer Engineering, Suwon, Republic of Korea
E-mail:{ristar1234, toanhoi, il2s, jtshin}@skku.edu Tel: +82-31-290-7994

***Abstract*— Dealing with multiple scale of object is main problem in computer vision. Feature Pyramid Networks (FPN) has widely used in instance segmentation area to utilize multiple scales of features. Using different scale of feature maps, the method enables to capture a various sizes of objects in a scene. However, FPN still cannot propagate semantic information of deeper layer into the shallow layer which contains spatial information strongly. In this paper, we propose a novel network which consists of stage residual connection and aggregation between C_i and P_{i-1} above the FPN to improve the imperfectness of original FPNs for the instance segmentation. Our proposed network is called Skipped-Hierarchical Feature Pyramid Networks (SH-FPN), integrated on Mask R-CNN. Experimental results of SH-FPN show that it has significant improvement on Data Science Bowl 2018 benchmark dataset on nuclei segmentation, compared to FPN.**

I. INTRODUCTION

The human body consists of tens of thousands of cells, each of them contains 'nuclei' that contains genetic information and regulates the cell function. By identifying the nucleus in input image, (e.g. microscopy images) researchers can discriminate several cells respectively in the scene so that they can investigate the cell's diverse responds from various treatment and these approach make it easy to find cure for the diseases. Therefore, observing cell nuclei in different environment is basic and most important issue to study the therapy of all kinds of diseases such as cancer, heart disease, or rare disease that current treatments have not been developed. Having an automated algorithm for detecting nucleus accurately, it is possible for us to develop effective therapies for existing diseases and also to be able to respond quickly to emerging diseases.

Due to the appearance of deep learning, object detection has achieved better performance than conventional methods. Furthermore, unlike conventional detection methods, users do not have to create a descriptor or hand-crafted features for sophisticated feature extraction anymore, because deep learning models are trained on the network using vast amounts of data directly related to the problem they are trying to solve. Those achievements are mainly due to both emergence of region-based Convolutional neural network (CNN) [3] and its upgraded versions [4], [7] and development of a feature pyramidal structure [8] for detecting' objects of various sizes. They made it possible to detect more difficult objects such as cell/nuclei than general things like vehicle, person, etc.

Despite the remarkable development with CNN, there are some shortcomings. Region-based CNNs are popular and usually show high performance but their computation time is too long and they cannot detect small object well. In the case of feature pyramid network (FPN), the semantic information of the deepest stage is delivered to the shallowest stage but not that strongly connected. Furthermore, once a problem occurs at any stage during training, the information will not be delivered properly to most stages.

FPN integrated on Mask R-CNN [9] shows a superior performance in instance segmentation. In this paper, we propose an improvement of FPN, called Skipped Hierarchical Feature Pyramid Network (SH-FPN) integrated on Mask R-CNN [9]. We focused on making the information flow within the structure more flexible and allowing the deeper, semantic information to be conveyed to the shallow stage which contains spatial information strongly.

II. RELATED WORKS

A. Feature Pyramid Networks (FPNs)

To create a high-accuracy object detector, first thing to do is letting networks know objects of any size that arise due to the varying sizes of objects, which is from the different perspective or various size of object itself in the image. The first approach to handle this was forming a pyramidal structure using one image through a resolution change and find objects through the stages respectively [1], which was very slow because each layer was independently computed. However, as CNNs evolve [6], the method of recognizing objects using the feature map on the CNN has gained popularity by exhibiting great performance in many object detection benchmarks. After then instead of input image itself, consisting pyramidal structures with CNN's feature map have been proposed. It is much faster than previous methods and has better performance using more complex features. A typical example of this approach is FPNs.

The FPN is based on the fact that as the depth of a layer deepens the contextual information and contains shallow spatial information, it is designed to supplement the semantic information in relatively shallow stages with strong spatial information. It takes images of arbitrary size, uses feature maps of several stages with a certain ratio, and consists of bottom-up pathway, top-down pathway, and lateral connection.

Bottom up pathway is the process of creating layered feature map through feed-forward operation in backbone of network such as ResNet [5]. Here, the same size of convolution blocks are grouped together as a stage and the last layer in each stage is used as a reference layer to create a pyramid structure. Here, {C2, C3, C4, C5} are the reference feature map of each stage that consist of pyramid structure and each of them is the representative of each stage. {C2, C3, C4, C5} are named in order from the shallowest layer to the deepest layer and each layer has (H / 2, W / 2), (H / 4, W / 4), (H / 8, W / 8), and (H / 16, W / 16) dimensions as input has (W, H) dimension.

The top-down pathway is a construction of pyramid structure with {C2, C3, C4, C5} with lateral connection. Each stage in the pyramid consists of elementwise summation of up-sampled feature map which contains a lot of semantic information and another feature map which is extracted from backbone network and computed 1 × 1 convolution. Through this process, semantic information in deeper layer can be passed to the previous layer, which contains more spatial information. In this case, the reference feature map of each stage is called {P2, P3, P4, P5} in order from the shallowest layer to the deepest layer and those P_i are the final feature maps which are used for network's purpose (e.g. detection, segmentation).

B. *Layer Aggregation*

Layer aggregation means the combination of certain layers which consist of network so that it makes network to be deeper and meaningful. According to [10], there are various types of aggregation. For the first, as shown in Fig. 1(a), there is a shallow aggregation like FPN which is linear and aggregates the deepest one at first and the shallowest one at last. The reason why this way called 'shallow' is that this kind of aggregation only aggregates two stages, arbitrary one stage and the stage which is right before the chosen one. Next, there is tree structured aggregation which 'literally' looks like decision tree structure which is shown in Fig. 1(b). Utilizing this structure can make some hierarchies between the information which can let feature maps be more complex.

In this paper, we have applied both shallow aggregation and tree structured aggregation to the four stages used in the top-down pathway of the FPN to create Skip-Hierarchical Feature Pyramid Networks (SH-FPN) that forms a more complex and

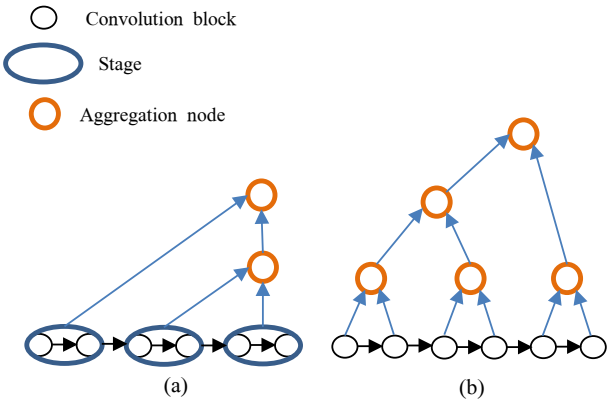

Fig. 1 (a) Shallow aggregation, (b) Tree structure aggregation

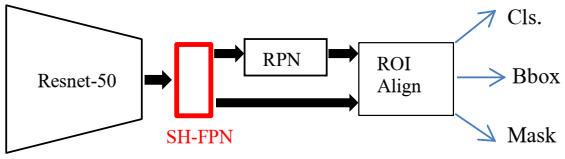

Fig. 2 Overall architecture

powerful feature map for instance segmentation.

III. PROPOSED METHOD

A. *Overall Structure*

. The proposed SH-FPN is constructed on Mask R-CNN, which shows excellent performance for instance segmentation. Considering that segmentation occurs after an object is firstly detected in Mask R-CNN, we focused on improving the performance of detection part by using SH-FPN. As shown in Fig. 2, it is easy to find out that where we put our SH-FPN in the Mask R-CNN. As shown in red square in Fig. 2, our model works behind the feature extractor, ResNet-50, which produces feature maps that will be the inputs of SH-FPN. And then, our SH-FPN makes more powerful feature maps that will be passed to RPN and ROI–Align to make a prediction for proper detection and segmentation.

B. *Skipped Hierarchical Feature Pyramid Networks*

Our proposed network composed with additional two connections to the top-down pathways of original feature pyramid network. Stage residual connection and C_i to P_{i-1} connection (i = 3, 4, 5). Here, followed the FPN's notation, {C2, C3, C4, C5} are feature maps from feature extractor and {P2, P3, P4, P5} are final feature maps that are used for network's purposes. Fig. 1(a) shows the aggregation way of original FPN and Fig. 1(b) shows tree structure. Our proposed structure can be seen in Fig. 3 and it is not hard to find out the difference between two pyramidal networks.

The main differences between two architectures are as follows: Stage residual connections and aggregation between C_i and P_{i-1}. These differences lead SH-FPN to make better performance than the original FPN.

Stage residual connection, the green arrow of Fig. 3, is just residual connection between P_i stages. It leads feature maps to contain more contextual information with locational information. That's because the semantic information of P5 and P4 will be directly attached to P3 and P2, which contain more locational information compare to the other deeper

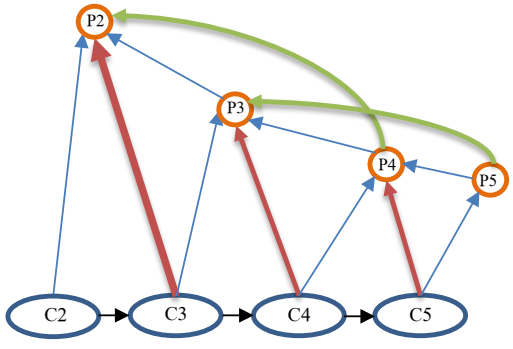

Fig. 3 Skipped Hierarchical FPN

layers. By adding these connections to existing top-down pathway of FPNs, allowing us to strengthen the way of shallow aggregation's contextual information propagation.

In the case of aggregation between C_i and P_{i-1}, the red arrow of Fig. 3, which we get some idea of tree structure, gives information hierarchy to network's feature maps.

Furthermore, interactions between C_i and P_{i-1} aggregation and stage residual connection can make more flexible network. For example, using a structure like Fig 1(a) and once problems happened during the training phase (e.g. values of the layer are nearly zero), information flow during the backward propagation will not be able to produce meaningful updates. It means that if such problem occurs at P3 in original FPN, it will not be able to update the information about P4 or C3 properly in the backward propagation procedure since P3 will give them very small update values. However, that kind of inefficiency has been solved with aggregation between C_i and P_{i-1}. For instance, let's assume the same problem occurred at P3. Having too small values or values that are nearly zero at P3, in the case of Fig. 3, still can update the information of P4 and C3 via connection from P2. That's why our proposed model has more flexible information flow.

IV. Experiment Results

A. Dataset and Training

We used image set BBBC038v1, available from the Broad Bio-image Benchmark Collection [2] to evaluate the method. It consists of 670 training samples and 65 test samples, which was created for the Kaggle 2018 Data Science Bowl benchmark.

Our proposed network (SH-FPN) is implemented based on Pytorch [11] and trained with an NVIDIA Titan X. Using Adam optimizer, the learning rate was initially set to 0.01 for 16K iterations and dropped the learning rate by a factor of γ = 0.1 for every 10K iterations. For fair comparison same condition was applied when we implement original FPN with Mask R-CNN.

B. Evaluation Metric

We evaluate our network, with mean average precision with several different intersection of union(IoU) thresholds, t.

The way to calculate the IoU between two set X, Y is as (1):

$$\mathrm{IoU}(X, Y) = \frac{X \cap Y}{X \cup Y} \quad (1)$$

And then computing IoU of ground truth and predicted mask for all threshold value t, which is the criteria of right detection and it starts from 0.5 and increased gradually with 0.05 and finished when the value of t is 0.95.

A precision value for each threshold is calculated with the number of true positives, false negatives and false positives and be calculated as (2):

$$\frac{TP(t)}{TP(t) + FP(t) + FN(t)} \quad (2)$$

Finally, we take the mean of all above precision value through the whole threshold and work out an average of all images' average precisions. In this case, the higher score means better performance.

C. Performance

Before evaluating the performance of our proposed method, we combined 4 images into 1 mosaic images for data augmentation. As a result, we used 378 gray images for training the network, 37 gray images for validation and 53 for testing. Also, we used 69 color images for training, 12 color images for validation and the other 12 color images for testing.

Since the small number of training image, we do validation process to avoid overfitting and find the early stop point.

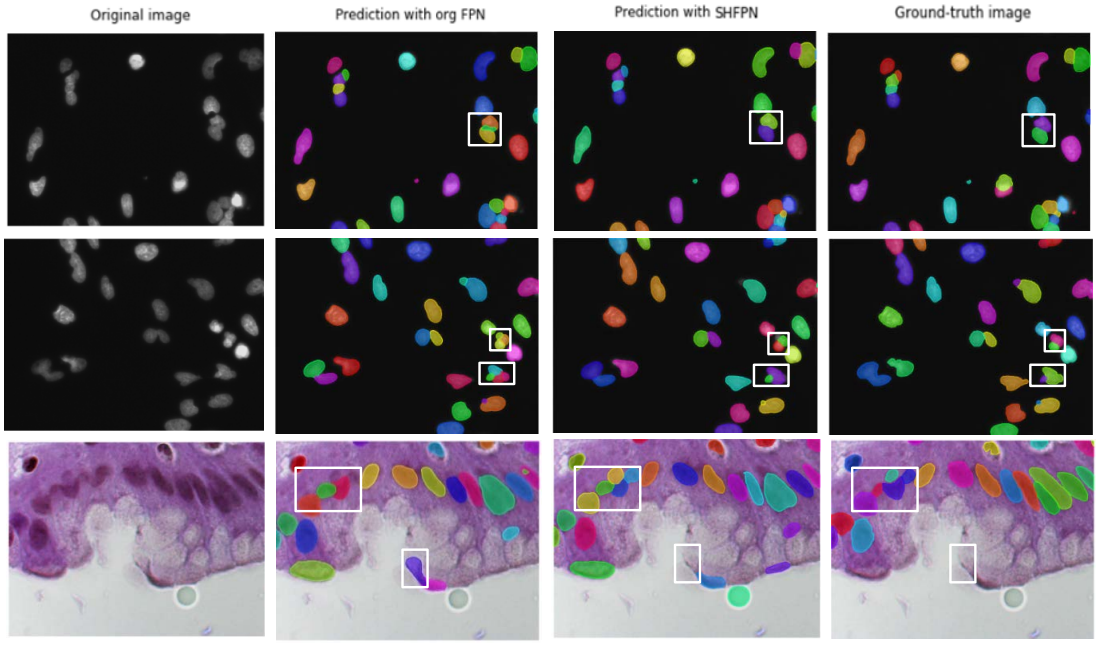

Fig 4. Detection results among validation images

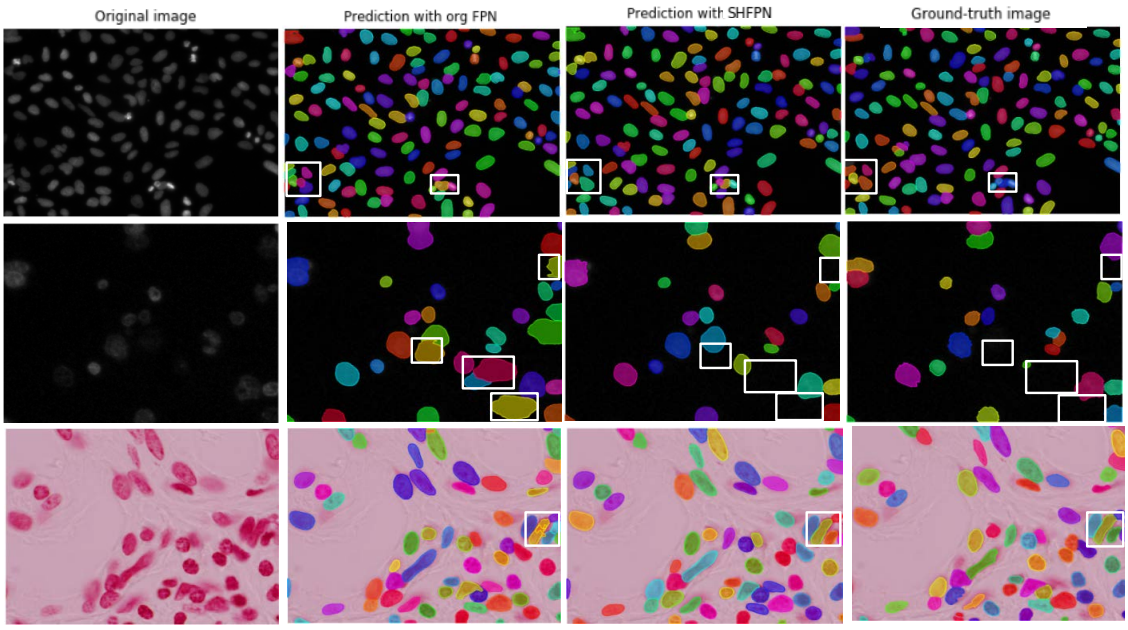

Fig.5 Detection results among test images

At Fig. 4 and Fig. 5, we visualize our detection result through the validation set and test set respectively. Each figure consists of original image, instance segmentation prediction result from FPN and SH-FPN, and the ground truth. Both Fig. 4 and Fig. 5 have white boxes which indicate the different prediction result of FPN and SH-FPN for the same original image. Comparison between the ground truth, prediction from FPN has many false positives. However, SH-FPN reduces more false positives than the original FPN by observing the white boxes in Fig. 4 and Fig. 5. For instance, by observing the 2nd row of Fig. 5, comparing with ground truth, it is easy to find many false positives inside the white boxes in the result of FPN. But the prediction result of SH-FPN doesn't have such false positives in the white boxes.

Table 1 reports the evaluated score over validation data with evaluation metric. We used original FPN with Mask R-CNN as baseline. Table 2 is the performance of networks among the test data. The result show that the total performance of SH-FPN is better than the original FPN in both of case.

Table 1 Mean average precision of state-of-the-art methods on validation set

Method	Gray	Color	Total
FPN[8]	0.4920	0.1029	0.5949
SH-FPN	**0.4958**	0.1028	**0.5986**

Table 2 Mean average precision of state-of-the-art methods on BBBC038v1 test set

Method	Gray	Color	Total
FPN[8]	0.4232	0.0574	0.4806
SH-FPN	**0.4573**	0.0560	**0.5133**

V. CONCLUSIONS

We proposed a novel network called SH-FPN which combines the extracted features with mixture of shallow aggregation and tree structures with several residual connections to FPN allows that our network can capture multiple scale features of nuclei better than original FPN. Our experimental results demonstrated the advantages of the proposed method in terms of detection accuracy comparison with accuracy of original FPN on BBBC038v1 dataset.

ACKNOWLEDGMENT

This research was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03031752), the NRF grant funded by the Korea government (MSIT) (NRF-2018R1C1B6007462) and partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2018-2018-0-01798) supervised by the IITP (Institute for Information & communications Technology Promotion).

REFERENCES

[1] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, J. M. Ogden "Pyramid methods in image processing" *RCA engineer*, 1984.

[2]] V. Ljosa, K. L. Sokolnicki, A. E. Carpenter, "Annotated high-throughput microscopy image sets for validation.", *in Nature Method,* 2012

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

[4] R. Girshick, "Fast R-CNN," *in IEEE International Conf. on Computer Vision (ICCV)*, 2015.

[5].K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image

recognition," *in IEEE International Conference on Computer Vision (ICCV)*, 2015

[6] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *in International Conference on Learning Representations (ICLR)*, 2015.

[7] S. Ren, K. He, R. Girshick, J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] T.Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie "Feature Pyramid Networks for Object Detection", in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017

[9] K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask R-CNN," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017

[10] F. Yu, D. Wang, E. Shelhamer, T. Darrell "Deep Layer Aggregation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018

[11] A. Paszke, S.Gross, S. Chintala, G. Chanan, E. Yang, Z. D.Vito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, *"*Automatic differentiation in PyTorch*"in Neural Information Processing Systems, 2*017