DUAL PYRAMIDS ENCODER-DECODER NETWORK FOR SEMANTIC SEGMENTATION IN GROUND AND AERIAL VIEW IMAGES

S. L. Jiang [1,3], G. Li [3], W. Yao [1,2*], Z. H. Hong[4], T. Y. Kuc [3]

[1]Department of Land Surveying and Geo-informatics, The Hongkong Polytechnic University, Hong Kong
[2]Research Institute for Sustainable Urban Development, The Hong Kong Polytechnic University, Hong Kong
[3]College of Information and Communication Engineering, Sungkyunkwan University, Suwon, Korea
[4]College of Information Technology, Shanghai Ocean University, Shanghai, China

KEY WORDS: Semantic segmentation, Encoder-decoder network, Convolution neural network, aerial and ground view image

ABSTRACT:

Semantic segmentation is a fundamental research task in computer vision, which intends to assign a certain category to every pixel. Currently, most existing methods only utilize the deepest feature map for decoding, while high-level features get inevitably lost during the procedure of down-sampling. In the decoder section, transposed convolution or bilinear interpolation was widely used to restore the size of the encoded feature map; however, few optimizations are applied during up-sampling process which is detrimental to the performance for grouping and classification. In this work, we proposed a dual pyramids encoder-decoder deep neural network (DPEDNet) to tackle the above issues. The first pyramid integrated and encoded multi-resolution features through sequentially stacked merging, and the second pyramid decoded the features through dense atrous convolution with chained up-sampling. Without post-processing and multi-scale testing, the proposed network has achieved state-of-the-art performances on two challenging benchmark image datasets for both ground and aerial view scenes.

1. INTRODUCTION

Semantic image segmentation is a dense classification task for image understanding, which has many practical applications such as autonomous driving and augmented reality devices. Since the proposal of fully convolutional network (FCN) (Long et al., 2015) has led to an end-to-end trend for semantic image segmentation, most of the state-of-the-art models are based on the FCN to implement dense classification of images. FCN-based architectures (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Treml et al., 2016; Jiang et al., 2019; Jiang et al., 2020) utilized several pooling layers to extract high-level features and restored the extracted feature map to original resolution through transposed convolution. However, this process inevitably lost some information during each down-sampling layer, and it is difficult for low-resolution feature maps to group the inline relationship during up-sampling.

Therefore, a great number of strategies were proposed to solve this contradictory, they mainly aim at how to minimize information loss during down-sampling and how to effectively aggregate different feature maps during up-sampling. Skipping connection architecture (Ronneberger et al., 2015, Badrinarayanan et al., 2017) was first proposed to compensate for the output of CNNs by connecting the feature maps between different layers. However, the feature maps in early stage of the neural network contain a large amount of noise, which sharply reduces the accuracy during classifying objects. Some neural networks (e.g., FCN-DenseNet (J´egou et al., 2017) and DenseASPP (Yang et al., 2018) utilized the concept of DenseNet (Huang et al., 2017) for the purpose of maximally increasing the inline connection among features of different scales. However, dense connections brought heavy computational cost coupled with increasing depth of the neural network. DeepLab series (Chen et al., 2015, Chen et al., 2017, Chen et al., 2018a, Chen et al., 2018b) proposed the ASPP module to enlarge the receptive field while maintaining the resolution. Specifically, atrous convolution (Holschneideret al., 1990) with various dilation rates are utilized to extract features in parallel. Although this kind of pyramid structure is effective in multi-scale feature extraction and can enhance the ability to classify and group ambiguous objects, it only captures contextual information from the deepest feature map by conducting a context module after the encoding stage. Therefore, we hold the view that the contextual information in early and middle stages can be further extracted to enhance feature extraction.

Summarizing above proposals, the current methods are mainly confronted with followed issues. 1.) Low level features in early or middle stages of neural networks are insufficiently utilized, and it leads to ambiguous classification during final outputting. 2.) Up-sampling through transposed convolution from shallow resolution feature map to original resolution results in the difficulty of grouping the objects. 3.) Receptive field size is difficult to be determined in high level feature maps, i.e., using large kernel size can cover large scale objects but the small-scale objects can be hardly detected. Vice versa, utilizing small kernel size can benefit to decoding small scale objects but it leads to the challenges for grouping of large-scale objects.

Based on above observations, the Dual Pyramids Encoder-Decoder Network (DPEDNet) is proposed. As shown in Fig.1 (a), our proposed deep neural network consists of two pyramids: multi-resolution feature aggregation pyramid for the encoder and multi-scale dense atrous convolution pyramid for the decoder. The first pyramid fuses the neighboring feature maps in order to sufficiently encode the different scale features. Different from FCN based neural network, our proposed neural network employs the multi-scale merging technology, which maximizes the utilization rate of the features from basenet and adapts them to the current scene. And the other pyramid aims at enlarging the receptive field and decoding the final feature map with minimal information loss. Different from the original usage of DenseASPP, we employ it as the decoder. As the multi-scale features are encoded as one feature block, the decoder scans the feature block in multi-receptive fields and densely decodes the features from multi-scales. The chained upscaling process is employed to verify the multi-scale strategy and avoid the noises of early stage feature maps (The features are not fine-tuned but directly concatenated with the deconvolution layers) from using

*Corresponding author

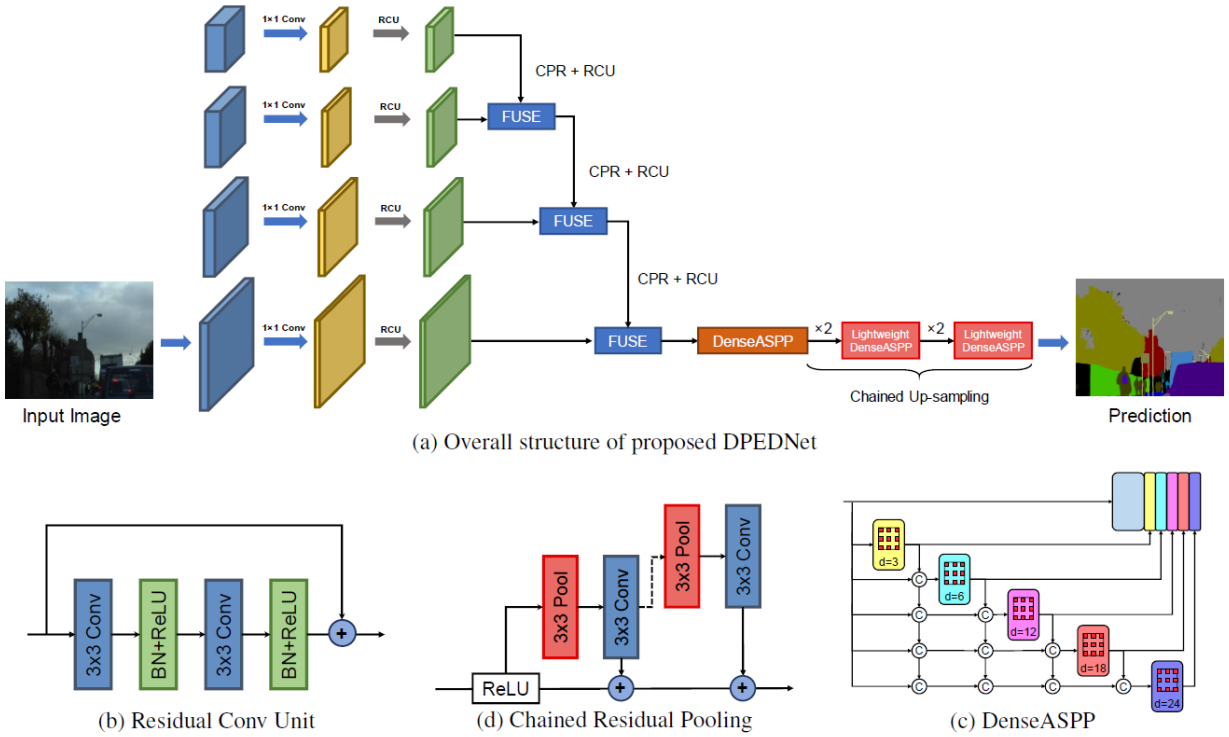

Figure 1. The overall structure of the proposed DPEDNet

skipping connection.

To prove the efficiency and effectiveness of our proposed deep neural network, we compared the DPED network with other state-of-the-art methods on two common image datasets from both ground and aerial view - CamVid (Brostow et al., 2008) and Crowded mapping (Mapping Challenge). Our proposed neural network has obtained 60.4% mIoU in the CamVid dataset, and 90.8% in the Crowded mapping challenge. Apart from the ordinary evaluation of the proposed neural network, we also evaluated the performance loss between RGB model and gray scale model. Specifically, we directly train the model with only gray-value image dataset to evaluate the model on RGB dataset. And we found that there is very limited accuracy loss on our model compared with training on RGB dataset, which means that the model shows a good ability to generalize. Our contributions in this paper can be summarized as follows:

1. Presenting a pyramid encoding architecture through stack merging strategy which sufficiently utilizes the multi-resolution features
2. Combined with the multi-scale dense atrous convolution pyramid, the whole network turns into a dense encoder-decoder structure with a denser receptive field through {3 - 15} dense atrous pyramid decoding for dense prediction tasks.
3. Evaluating the proposed network on two challenging benchmark datasets for both ground and aerial view image segmentation tasks by achieving state-of-the-art performances.

2. PROPOSED NETWORK

In this section, we elaborate the architecture and preliminary knowledge of our proposed deep neural network. The overall architecture of our proposed network is shown in Fig.1 and it consists of four parts, multi-resolution feature aggregation pyramid, residual convolution unit, chained residual pooling and multi-scale dense atrous convolution pyramid.

2.1 Multi-resolution feature aggregation pyramid

Currently, most deep neural networks only employ the deepest feature map after the encoding stage, and then use it as the sole input to context modules such as ASPP (Chen et al., 2018a) and SPP (Zhao et al., 2017). This process is not efficient enough to extract the dense feature for the decoding stage, and certain information can get lost during the encoding process.

In order to improve the feature utilization and the flow of information, we propose a multi-resolution feature aggregation pyramid to merge all feature maps from different resolutions. Firstly, we employed the widely used ResNet (He et al., 2016a, He et al., 2016b) as backbone to extract initial features, and the backbone is pretrained in ImageNet (Russakovsky et al., 2015) dataset to secure improved result. Referring to the RefineNet (Lin et al., 2017), we use the residual convolution unit (RCU) and chained residual pooling (CRP) to refine the feature, which can greatly increase the receptive field without much extra computational cost, and we add batch normalization in RCU as well as change the pooling size of CRP to 3x3 for smoothing results. The Fig.1 (b) and (c) represent the detailed structure of RCU and CRP.

As shown in Fig.1 (a), four different layers from ResNet backbone first go through a 1x1 convolution for input adaptation and followed by two RCU blocks to further refine the information, and CRP is used before every feature fusion step. Then, a stack architecture is employed to fuse every two feature maps of different resolution, the fusion can be expressed as the following formula:

$$F_n = f(m_{n-1}) + F_{n-1} \quad (1)$$

Here m_{n-1} denotes the feature map with the lower resolution, f is the function used to enlarge the image size by making it the same as feature map F_{n-1}, here we use 3x3 convolution followed by the transposed convolution to upsample the feature map. Finally, we combine these two feature maps together and generate a new feature map F_n. Once the size of the feature map reaches one fourth of the original input, we stop the feature fusion and use the final layer as input for the decoding stage. The overall architecture allows an effective encoding procedure, which sufficiently utilizes the features from different resolutions and outputs an encoded feature map with large resolution (i.e., the 1/4 of the original size).

2.2 Multi-scale dense atrous convolution pyramid

To effectively decode the encoded feature map, an effective scale-variant decoder with large receptive field is needed, and here we selected the DenseASPP as decoder. To address the issue of limited inline connection among the scales, Yang et.al (Yang et al., 2018) proposed the DenseASPP which takes advantages of both Atrous Spatial Pyramid Pooling (Chen et al., 2018a) and DenseNet (Huang et al., 2017). Using the $H_{k,d}(x)$ to represent an atrous convolution, then the general ASPP can be written as:

$$y = H_{3,6}(x) + H_{3,12}(x) + H_{3,18}(x) + H_{3,24}(x) \quad (2)$$

k represents the kernel size and d denotes the dilation rate of atrous convolution. The neural network employed the atrous convolution block with the dilation ratio series {6,12,18,24}. Such convolutional filter has the ability to cover 122 receptive field, which is quite larger than ASPP (e.g., 55 in the same ratio). Apart from the receptive field, the dense connection among the different scale feature maps also allow a closer inline connection to further decode the multi-scale features. Therefore, a flexible receptive field provides a global information decoder for different scales' objects, which allows the DenseASPP succeeding in decoding the fused feature map from various scales.

As shown in Fig.1 (d), the DenseASPP utilizes the dense feature extraction pyramid to decode the feature map, and dense connections are employed to aggregate diverse layers with different dilation rates and receptive fields. The procedure can be summarized as follows:

$$y_l = H_{k,d_l}([y_{l-1}, y_{l-2}, \ldots, y_0]) \quad (3)$$

where d_l represents the dilation rate of layer l, and [...] denotes the concatenation operation. $[y_{l-1}, y_{l-2}, \ldots, y_0]$ means the feature map formed by concatenating the outputs from all previous layers.

After the DenseASPP, two transposed convolution layers were employed to upsample the output and restore the resolution to the same as the original input, and we termed this process as Chained up-sampling. Specifically, every transposed convolution is followed by a lightweight DenseASPP to refine the upsampled feature map (dilation rates 3, 6, 9 with convolution depth 32). The lightweight chained up-sampling step allows the neural network to further decode the feature maps during the de-convolution steps. Currently, other methods generally classify the image in the deeply down-sampled feature map, and the DNN recover the size of feature map by simple bilinear up-sampling operation. Through the procedure, the object boundary is hard to be accurately delineated. Another trend is to employ the deconvolution combined with skipping connection, but this strategy usually leads to an ambiguous output. The proposed chained up-sampling employed a continuous and gradual up-scaling step. The light-weight DenseASPP effectively enlarges the shallow-sized feature map. The procedure avoids the noises from skipping connection and enables to produce the feature map to classify in proper resolution without information loss. Through our process, the decoded feature map can be further refined with clear boundary, and the gradually up-sampling enables to classify the images with better performance. In the end of our DNN, a softmax layer was employed to produce pixel-level segmentation results. Our proposed method can be regarded as two parts, the first pyramid encodes the features from different scales and densely merges them. The second pyramid decodes the encoded information in large receptive field. The chained up-sampling scale procedures enables the neural networks to adjust the shallow feature map to original size with gradually enlarging step-size.

Figure 2. Qualitative examples of the semantic segmentation on CamVid dataset

3. EXPERIMENTAL RESULTS

To evaluate the proposed method, extensive experiments are implemented on CamVid and Crowded mapping challenge. We describe detailed experimental settings in Sec. 3.1 and present both qualitative and quantitative results of two datasets in Sec.3.2.

3.1 Experimental Settings

CamVid: This is a street scene dataset for autonomous driving applications. It contains 421 training image, 112 validation images and 168 test images, and all images have a resolution of 720 x 960 with 32 semantic categories (which is distinguished from simple CamVid dataset with 19 classes and 360 x 480 resolution). We use this dataset for ground view segmentation task. Employing very limited number of images to train the model in various items in complicated street scene is the major challenge of the CamVid test.

Crowded mapping challenge: This is an open source competition dataset, which contains 281,423 training images and 60,314 validation images with 300 x 300 resolutions. There are two categories contained by the dataset, i.e., buildings and background. We use this dataset for testing aerial view images. As there are no open-access ground truth labels for test set, we utilized the validation set to serve as test set and the number of validation images is supposed to be large enough for objective evaluation. The shadow, complicated shape and changeable illumination situations restrict the models to obtain very promising accuracy in the building segmentation task.

Implementation details: For fair comparison, all the experiments were deployed on the TensorFlow platform under Ubuntu OS. Our desktop uses the I7 8700 CPU with 16GB memory and a single GTX 1080Ti GPU. To maximize the GPU memory usage and fast convergence, we maximized the batch size as large as possible during the training. We use the Adam optimization with a large minibatch size 5 to make full use of the GPU memory and set an initial learning rate as 0.0001 with learning rate decay 0.995 in every 1000 steps for the two datasets.

The number of training epochs is 400 for the CamVid and 10 for the Crowded mapping challenge. Finally, we random cropped the images to 256 x 256 in the Crowded mapping challenge and 512 x 640 for the CamVid dataset.

3.2 Evaluation Results on RGB Dataset

We report the quantitative segmentation examples in Fig.2 and Fig.3, and the results of both ground and aerial view images have proven that our DPEDNet effectively captured the contextual as well as the detailed information. Following the common procedure of semantic segmentation, we reported the precision, recall and mean Intersection over Union (IoU).

Segmentation results for ground view data are shown in Table 1. Our proposed DPEDNet obtained 90.7% precision, 89.5% recall, and 60.4% mIoU, which are the highest among all the methods. The high precision and recall represent that our approach detects most items in the challenging road scene only predicted a limited number of false negative samples. On figure 2, the visualization images show that our proposed DPEDNet enables to accurately detect and segment the objects in various scales, complicated scene and very challenging illuminate situation. This is benefitted by the double pyramids' encoder-decoder architecture. The first pyramid allows the neural network to encode the features to adapt street scene. The second pyramid decodes the engaged features, and output by the chained decoding DenseASPP, which makes the network segment the objects with proper boundary. In addition, our model can process every image in 0.11 seconds (9.1 FPS), which almost reaches the real-time requirement.

Table 1. Evaluation results on the CamVid test set.

Method	Precision (%)	Recall (%)	mIOU (%)
DenseASPP	84.3	82.6	44.8
DeepLabV3	87.0	85.1	46.6
PSPNet	86.8	85.0	52.9
UNet	88.7	87.2	53.0
DeepLabV3+	88.0	86.8	53.1
FCN-DenseNet	90.0	88.4	54.8
GCNet	89.4	87.9	56.1
RefineNet	90.4	89.0	57.8
DPEDNet	**90.7**	**89.5**	**60.4**

Segmentation results for aerial view data are shown in Table 2. Our approach again achieved the leading performance. We obtained 91.6% precision, 91.9% recall, and 90.8% mean IoU, which is much superior to the second-place FCN-DenseNet. Fig.3 shows that the proposed network successfully segmented the edges of the buildings and overcame the occlusion problem of trees and shadow. This proves that the DPED architecture allows to detect and segment with large scale receptive field. Even the buildings are sheltered by the trees or other background noises, the DPED architecture enables to discover the inline relation among the neighbor pixels, and accurately classify the objects. Both the results on ground and aerial view data prove that our DPEDNet sufficiently extract the multi-scale information through the double pyramid encoder decoder architecture, and allows better adaption to the environment compared with other state-of-art methods.

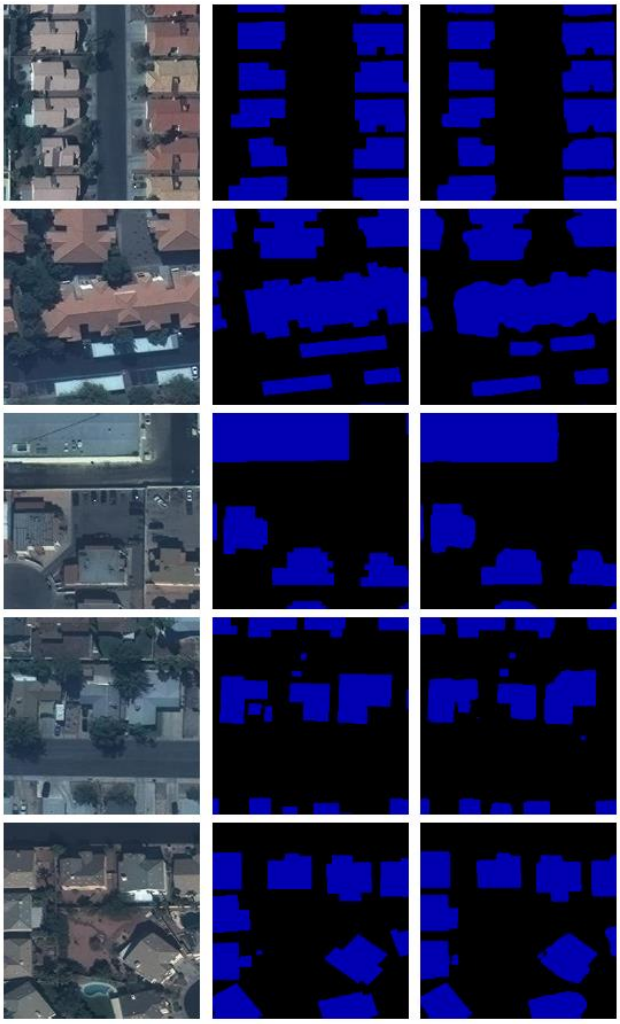

Figure 3. Qualitative examples of the segmentation results on Crowed mapping challenge Dataset

Table 2. Evaluation results on Crowded mapping challenge validation set.

Method	Precision (%)	Recall (%)	mIOU (%)
UNet	82.9	82.3	81.2
GCNet	83.2	82.9	82.2
PSPNet	84.9	84.6	83.2
DeepLabV3	85.1	84.9	83.8
DenseASPP	86.7	87.0	85.6
DeepLabV3+	87.5	87.6	85.8
RefineNet	87.2	87.3	86.2
FCN-DenseNet	87.2	87.1	86.4
DPEDNet	**91.6**	**91.9**	**90.8**

Table 3. Evaluation results on Crowded mapping challenge validation set with one channel image.

Method	Precision (%)	Recall (%)	mIOU (%)
UNet	82.9	82.4	81.1
GCNet	83.2	83.1	82.2
PSPNet	84.9	84.5	83.1
DeepLabV3	85.0	84.5	83.8
DenseASPP	86.6	85.9	85.5
DeepLabV3+	87.4	87.2	85.7
RefineNet	87.2	86.5	86.2
FCN-DenseNet	87.1	87.2	86.3
DPEDNet	**91.5**	**91.2**	**90.8**

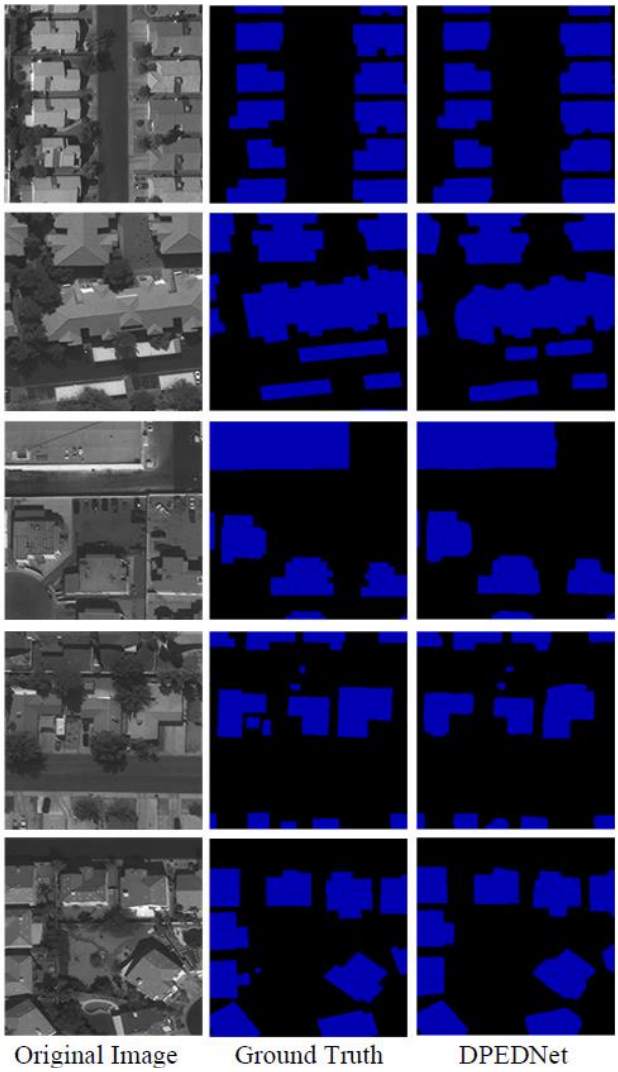

Figure 4. Qualitative examples of the segmentation results on Crowed mapping challenge dataset with single channel images

3.3 Evaluation Results on One-Channel Dataset

Apart from the RGB image, the single channel imagery is widely used as the source input in the community of remote sensing. Therefore, we evaluated the performance loss in the corresponding one channel dataset on the building segmentation dataset. The same parameters as the RGB dataset are employed in one channel evaluation.

The results are shown in table 4, the models trained and tested by the gray scale imagery retained the same level accuracy as the RGB dataset. All the deep neural networks retained similar accuracy both in precision rate, recall rate and mIoU in the corresponding gray dataset. These results illustrate that the models trained by the gray scale features also allow effectively pixel-level segmentation and classification in the aerial image scene.

One hypothesis arises that if the single channel gray-value image results can be transferred and benefit to the usage of RGB channels image. Thus, we can directly use the model trained by single channel dataset to test in RGB dataset. As shown on Table 4, the results which we have thoroughly checked are greatly beyond the common sense. The results do not have any significantly accuracy loss even directly using the gray scale model to test in RGB dataset for all deep neural networks. The results illustrated that the difference between using RGB and gray scale models are slight, which implies that they could be directly interoperated.

Table 4. Evaluation results on the Crowded mapping challenge validation set by using the model trained by single channel images testing in RGB images

Method	Precision (%)	Recall (%)	mIOU (%)
UNet	80.5	79.9	79.3
GCNet	79.8	79.4	78.5
PSPNet	82.8	82.4	80.8
DeepLabV3	83.6	83.4	82.5
DenseASPP	83.4	82.9	82.4
DeepLabV3+	85.9	85.4	84.5
RefineNet	85.6	85.4	84.5
FCN-DenseNet	86.3	86.1	85.5
DPEDNet	**90.4**	**89.6**	**89.1**

The phenomenon is surprising and different from our common sense. The destruction process from converting RGB channels to one channel imagery has unrecoverable information loss thus the rule of decoding the image for the two types of images should be different. The models trained by one channel imagery only learn the rule of decoding in gray scale imagery, but it has the strength to decode in RGB channels imagery. The reason is also not attributed to the basenet as the FCN-DenseNet and UNet have similar phenomenon.

Currently, we have one hypothesis about it, i.e., if the decoding strength can break the dimensionality once the number of low dimension training dataset is adequate. However, the quantitative test is not suitable to be deployed as the decrease in accuracy is difficult to be summarized as the reason for accuracy reducing or because of other factors. Currently, we have not found the answer and it is good to open the discussion to the community.

4. DISCUSSION

The proposed DPED network has proved its benefits in the two public and challenging benchmark datasets (i.e. one of them features a large class number but with limited training samples, another one only has binary classification but with sufficient training samples) compared with state-of-art deep neural networks. The DPED network employed multi-resolution feature aggregation pyramid to densely and maximally utilize the image samples. Then, the encoded feature map is decoded by the DenseASPP which enlarged the receptive field in {3, 6, 9 …15} dilated convolution to enable the detection in the street view scene, and the chained upscale strategy allows gradual up-sampling operation to avoid the information loss and ambiguous

classification caused by deconvolution from shallow size to large size feature map.

The results also support our opinions. In the street view scene imagery (CamVid), the DPED obtained 60.9% MaP which outperformed another multi-scale encoder (RefineNet) by 2.6 %. The result illustrates that our proposed method has the strength on maximally utilizing the features, which allows it to segment and classify the scene with limited number of training samples. It also obtained 90.8% MIoU for the crowdedAI dataset, and surpassed the second one by 4.5%. The DenseASPP and chained up-sampling strategy also contributed to allowing the DPED network to decode the image in large receptive fields which is shown in the crowdedAi building detection. It helps our DPED network to segment the objects by overcoming shadows, occlusions etc, due to the fact that the two architectures could enhance the inline relationship among the pixels. The overall architecture makes our DPED network to become a competitive neural network toward semantic segmentation.

5. CONCLUSION

In this paper, we proposed the DPEDNet which integrates multi-resolution feature aggregation from multi-scale feature extraction with contextual decoding capacity from DenseASPP, which allows the neural network to sufficiently utilize the features from multi-layer convolution with limited information loss. We have demonstrated the effectiveness of our model by carrying out comprehensive experiments on both ground view and aerial view image to perform the semantic segmentation task, and the proposed DPEDNet achieved remarkable results on two challenging benchmark datasets.

ACKNOWLEDGEMENTS

This work was supported in part by the Korea Evaluation Institute of Industrial Technology (KEIT) funded by the Ministry of Trade, Industry & Energy (MOTIE) under Grant 1415164140, in part by a grant PolyU 1-BBWD from the Research Institute for Sustainable Urban Development, The Hong Kong Polytechnic University.

REFERENCES

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2015. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. CoRR, abs/1412.7062.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018a. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40, 834-848.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. CoRR, abs/1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. ECCV.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks. ECCV.

Holschneider, M., Kronland-Martinet, R., Morlet, J., Tchamitchian, P., 1990. A real-time algorithm for signal analysis with the help of the wavelet transform. Wavelets, Springer, 286–297.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q., 2017. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708.

Jiang, S., Yao, W., and Heurich, M.: DEAD WOOD DETECTION BASED ON SEMANTIC SEGMENTATION OF VHR AERIAL CIR IMAGERY USING OPTIMIZED FCN-DENSENET, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W16, 127–133, https://doi.org/10.5194/isprs-archives-XLII-2-W16-127-2019, 2019.

Jiang, S., Yao, W., Wong, M. S., Li, G., Hong, Z., Kuc, T. Y., & Tong, X. 2020. An Optimized Deep Neural Network Detecting Small and Narrow Rectangular Objects in Google Earth Images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 1068-1081.

J´egou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 11–19.

Lin, G., Milan, A., Shen, C., Reid, I., 2017. Refinenet: Multipath refinement networks for high-resolution semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 1925–1934.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional

networks for semantic segmentation. CVPR.

Mapping Challenge, n.d.
https://www.crowdai.org/challenges/mappingchallenge/

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computerassisted intervention, Springer, 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115, 211-252.

Treml, M., Arjona-Medina, J., Unterthiner, T., Durgesh, R., Friedmann, F., Schuberth, P., Mayr, A., Heusel, M., Hofmarcher, M., Widrich, M. et al., 2016. Speeding up semantic segmentation for autonomous driving. MLITS, NIPS Workshop, 2, 7.

Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K., 2018. Denseaspp for semantic segmentation in street scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3684–3692.

Zhao, H., Shi, J., Qi, X.,Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6230-6239.