Non-local D-LinkNet with Separable Convolutions for Road Segmentation

Eiman Alnuaimi Alessandro Cabodi Dimitri Francolla Rafael Wanner
Department of Computer Science, ETH Zurich, Switzerland
Group: HighwayToCIL
{ealnuaimi,acabodi,dfrancolla,rwanner}@ethz.ch

***Abstract*—The widespread use of autonomous vehicles across different industries has made road segmentation a crucial application of semantic image segmentation. Recent approaches apply Convolutional Neural Networks (CNNs) to achieve this task. However, CNNs can lose fine-grained details which can be crucial for the segmentation task. In this paper, we propose a combination of two novel variation of D-LinkNet to features non-local operations and separable convolutions to preserve image details while improving the model's resources consumption. This paper describes our pipeline, including: data pre-processing; baselines selection; improving the existing D-LinkNet architecture; training with different loss functions; ensembling the best performing models. Results demonstrate the efficacy of our pipeline with a public F1 score of 0.92157 and reduction of GPU time during training by $\sim 29\%$.**

I. Introduction

In recent years, there has been a substantial surge in the adoption of autonomous vehicles across various domains, including rescue missions and fast delivery solutions within e-commerce [1], [2]. Powered by recent breakthroughs in Deep Learning, these devices exhibit the capability to effectively perceive and navigate their environment, hence successfully completing their mission [3]. A common task these devices have to perform is road segmentation, which is an application of semantic image segmentation [4]. Such a task entails employing pixel-wise classification where categories, "road" or "no road", are assigned to individual pixels within an image. This renders the road segmentation task more intricate in comparison with conventional image classification tasks.

II. Related Work

In this section, we conduct a review of some relevant contemporary architectural approaches employed for road segmentation, highlighting the progress made in mitigating spatial information loss within CNN architectures.

CNN Architectural Breakthroughs: Recent advancements in image segmentation employ Fully-Convolutional Networks (FCNs) [5], [6], [7], which is a family of Convolutional Neural Networks (CNNs) that is capable of generating dense pixel-wise predictions from arbitrary-sized images [8]. Nevertheless, FCNs may disregard scene-level semantic context, leading to low-resolution feature maps in the final output [9], [10]. Dilated convolutions address the resolution decrease due to stride convolutions in FCNs but also entail heavy computations [9]. On the other hand, depthwise separable convolutions [11] have often been employed in neural networks design to reduce parameters and computational complexity, thereby enhancing representational efficiency [12], [13], [14]. Joint Pyramid up-sampling (JPU) effectively employs separable convolutions to decrease the computation cost of dilated convolutions [10]. Additionally, non-local networks [15] overcome FCNs' limitations by capturing long-range dependencies through a weighted sum of features across input feature maps to compute responses [16].

CNN for Road Segmentation: U-Net, which has been first proposed to improve accuracy in medical images segmentation [5], has since been extended to other segmentation tasks [9]. Residual U-Net (ResUNet) [17] combines residual units [18] with U-Net to extract roads from aerial images to prevent information degradation [18]. D-LinkNet [19], which builds upon LinkNet [6], is a semantic segmentation neural network designed specifically for the DeepGlobe Road Extraction Challenge. It employs dilated convolutions to better handle intricate road properties such as narrowness and connectivity.

III. Our Contribution

In this paper, we introduce: i) Two novel D-LinkNet models that, combined, improve training efficiency while accurately retaining road details ii) A quantitative comparison between our baseline models and the proposed model iii) An assessment of patch accuracy using two standard loss functions and a novel hybrid loss To the best of our knowledge, our work is the first to combine non-local blocks and separable convolutions with D-LinkNets for road segmentation. Our implementation is available on GitHub[1].

IV. Methodology

This section introduces our baselines and outlines the systematic framework employed to develop our novel models.

A. Data Pre-processing

The dataset comprises 144 aerial RGB images of resolution $3 \times 400 \times 400$, each accompanied by a corresponding ground truth $1 \times 400 \times 400$. The ground truth data represents road pixels, where a value of 0.0 signifies "not road" and 255.0 denotes "road". Due to the limited size of the

[1]https://github.com/rafalum/HighwayToCIL

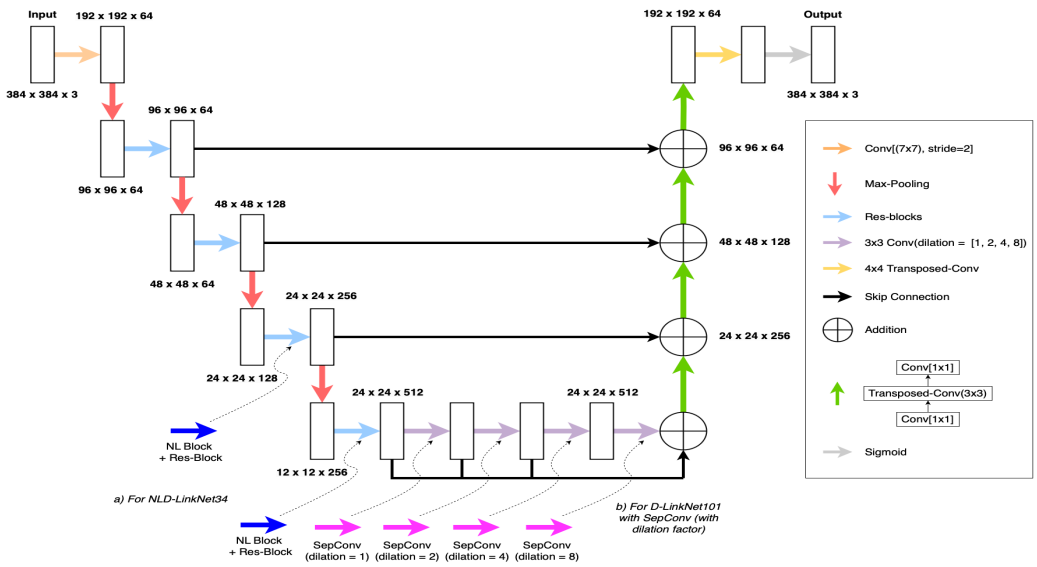

Figure 1. Architecture of D-LinkNet34 model. Dotted arrows represent the optional additional blocks to be inserted in the architecture to obtain our novel proposed architectures. Namely, (a) non-local blocks for NLD-LinkNet34, (b) separable convolutions with dilation for more efficient D-LinkNet101

dataset, we incorporate random geometric and photo-metric augmentation techniques. Initially, the data is subjected to random rotation, center cropping, resizing, horizontal and vertical flipping. Subsequently, we use randomized factors to enhance edges, blur with Gaussian filter, reduce image brightness and enhance their contrast. Notably, brightening the images and applying colour inversion did not improve results. The data is resized to 384×384 before being passed as input to the models.

B. Baseline Models

1) U-Net: Our first baseline is U-Net, as presented in [5]. U-Net has an encoder-decoder structure with identity shortcuts (skip connections) to preserve spatial details.

2) ResUNet: Our second baseline is ResUNet [17] which improves the U-Net's architecture by including residual units. We extend the original architecture with an extra encoder layer which further improves the model's capacity to capture intricate spatial information and contextual features while maintaining a comparable number of parameters with respect to U-Net.

3) LinkNet34: Our third baseline is LinkNet34 [6], which also has an encoder-decoder architecture with skip connections between layers. Differently from U-Net, LinkNet uses ResNet [20] as a backbone, thus including residual units in each of the encoder blocks. For performance reasons, we implement the LinkNet34 using ResNet34 instead of the originally proposed ResNet18 [21].

C. Model Architecture

D-LinkNet [19] builds on LinkNet by adding dilated convolution layers in the center part, between the encoder and the decoder. This increases the receptive field of feature points without sacrificing feature map resolution, thus better capturing intricate road characteristics such as narrowness, connectivity and the distance they cover. D-LinkNet combines both cascade and parallel modes of dilated convolution, leveraging shortcut connections to enhance segmentation accuracy.

There are different flavours of the architecture depending on the ResNet version used (ResNet34 or ResNet101 [21]). In the following we shall refer to D-LinkNet34 and D-LinkNet101 to specify the backbone we are using. The latter is a deeper and a more complex architecture which may help in retrieving more details, at the expense of potentially introducing more noise together with an increase in the computational cost. Figure 1 represents the D-LinkNet34 architecture. D-LinkNet101 is practically the same, with the appropriate parameter scaling. We propose two new versions based on D-LinkNet34 and of D-LinkNet101, respectively.

1) NLD-LinkNet: NL-LinkNet [22] proposes the use of non-local neural operations in order to overcome the receptive field limitations of CNN methods. The non-local operations compute feature map values as a weighted sum of the features at all positions, enabling the model to capture distant information and to account for long-range dependencies.

Each non-local block consists of the result of the non-local operation connected with an input feature via residual connection, as described in [16]:

$$z_i = W_z y_i + x_i$$

Where the non-local operation is defined as follows:

$$y_i = \frac{1}{C} \sum_{j=1}^{N} f(x_i, x_j) g(x_j)$$

For the definition of f, g we choose an embedded Gaussian for the former and a simple form of linear embedding for the latter:

$$g : g(x_j) = W_g x_j$$
$$f : f(x_i, x_j) = e^{(W_\theta x_i)^T (W_\phi x_j)}$$

Inspired by this work, we add non-local blocks to the standard D-LinkNet architecture (see Figure 1a).

2) D-LinkNet101 with Separable Convolutions (SepConv): D-LinkNet101 is significantly more costly in terms of resources when compared to the smaller D-LinkNet34. Inspired by the work in [10], which replaces the time and memory costly dilated convolutions by proposing a novel joint up-sampling module named Joint Pyramid up-sampling (JPU), we modify the center part of the D-LinkNet101 architecture (see Figure 1b) to render dilated convolutions more efficient. In particular we replace the standard 2D convolutions with separable convolutions as presented in [12], [13], [14]:

$$Conv(W, x)_{(i,j)} = \sum_{k,l,m}^{K,L,M} W_{(k,l,m)} \cdot x_{(i+k,j+l,m)}$$

$$PointwiseConv(W, x)_{(i,j)} = \sum_{m}^{M} W_m \cdot x_{(i,j,m)}$$

$$DepthwiseConv(W, x)_{(i,j)} = \sum_{k,l}^{K,L} W_{(k,l)} \odot x_{(i+k,j+l)}$$

$$SepConv(W_p, W_d, x)_{(i,j)} = PointwiseConv_{(i,j)}(W_p, DepthwiseConv_{(i,j)}(W_d, x))$$

In section V we show that our modification does in fact significantly reduce memory demands of the model .

D. Ensemble Learning

In Ensemble learning, predictions from two or more base models are combined, for example by voting mechanisms or by averaging. The non-local blocks enable NLD-LinkNet34 to retain fine grained details while maintaining a similar efficiency compared to the original D-LinkNet34. Meanwhile, using separable convolutions in D-LinkNet101 improves efficiency despite using a deeper and more complex architecture as backbone, leading to increased precision and reduced noise. The complementary nature of D-LinkNet101 SepConv and NLD-LinkNet34 renders them ideal for combining predictions, effectively mitigating their individual weaknesses.

E. Loss Functions

The default loss function for segmentation tasks is Binary Cross Entropy (BCE) which measures the difference between two probability distributions of a random variable [23]. We define BCE where y is the ground truth and $\hat{y}$ is the predicted value:

$$L_{BCE}(y, \hat{y}) = -\left(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right)$$

To handle the highly imbalanced class of our ground truths, we use Focal Loss, a variant of BCE that down-weighs the contribution of simple examples while considering harder examples [24]:

$$L_{Focal}(p_t) = -(1 - p_t)^\gamma \log(p_t)$$

We are interested to see what the result is when we combine these two loss functions into a hybrid loss. β is a contribution factor that helps us decide how much weight is given to the BCE contribution. We define the hybrid loss below:

$$L_{Hybrid}(p_t) = \beta L_{BCE} + L_{Focal}$$

In section V, we compare the accuracy of different loss functions on our proposed models.

F. Training

The selected framework for the implementation is PyTorch. For training we use Adam optimizer with a learning rate scheduler that dynamically reduces the learning rate of the optimizer if the patch validation accuracy stops improving over 50 epochs. To reduce over-fitting, we use an early stopping strategy based on the improvement of the patch accuracy of the validation over 30 epochs. We have a data split of 90:10, where 134 images are used for training and 10 images for validation.

G. Model Evaluation

To evaluate the performance of the models, we use an F1 score which captures both the recall and accuracy of the test set. We further evaluate the Jaccard Index, referred to as Intersection over Union (IoU), which is a measure of the similarity between two samples. The IoU is calculated with 10 validation images that were withheld from training.

V. Results

A. Analytical Results

Comparison with Baselines: In Table I we compare our three baselines to the introduced models. All models were trained on 134 training images, processed as mentioned in section IV-A, using BCE. The IoU was measured using the 10 validation images.

Comparison between losses: We compare the proposed models when trained with the three different loss functions introduced in section IV-E. In Table II, one can see that the Binary Cross Entropy results in a better F1 score compared to the two more sophisticated losses.

Model	F1 score	IoU
Unet	0.89653	0.458
ResUnet (not trained)	0.90475	0.504
ResUnet pre-trained	0.90526	0.5174
LinkNet34	0.9114	0.564
D-LinkNet34	0.9126	054
D-LinkNet101	0.90306	**0.583**
NLD-LinkNet34	0.91535	0.547
D-LinkNet101 SepConv	**0.92097**	0.58

Table I
F1 SCORE AND IOU FOR EACH OF THE MODELS TRAINED ON BINARY CROSS ENTROPY

Model	BCE Loss	Focal Loss	Hybrid Loss
NLD-LinkNet34	**0.91535**	0.90764	0.91391
DLinkNet101 SepConv	**0.92097**	0.91857	0.91857

Table II
F1 SCORE FOR THE PROPOSED MODELS ON BINARY CROSS ENTROPY (BCE), FOCAL AND HYBRID LOSS

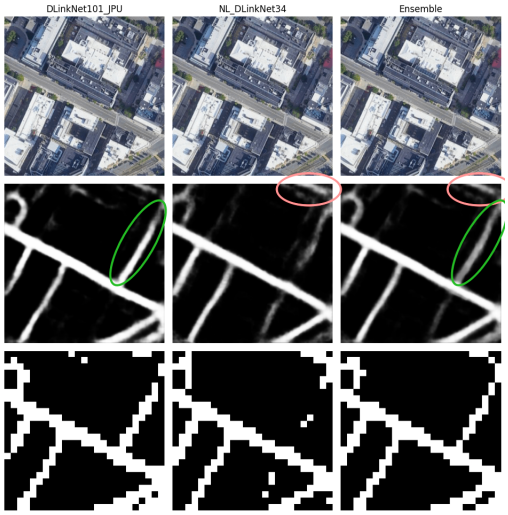

Figure 2. For each of the three models, the top row shows the input to the model, the middle row shows the raw output prediction of the model and the bottom row shows the patched prediction.

Ensemble model: For the ensemble models, we perform a convex combination with the raw predictions of the two best performing models, D-Link101 with SepConv and the NLD-LinkNet34. The combined prediction resulted in an F1 score of **0.92157**.

Memory Analysis: By replacing the dilated convolutions in D-LinkNet101 with SepConv, we are able to reduce number of parameters by $\sim 4\times$. We also notice a considerable improvement in GPU usage (see Table III).

Model	GPU Memory	Parameters
D-LinkNet101	11,275 MB	236M
D-LinkNet101 SepConv	8,052 MB	68M

Table III
MEMORY COST OF STANDARD D-LINKNET101 AND OUR VERSION WITH SEPARABLE CONVOLUTIONS (RESULTS WERE OBTAINED BY USING BATCH SIZE OF 8)

B. Visual Results

In Figure 2, we show a sample prediction from the two proposed models. The green circle marks an area where the D-LinkNet101 SepConv is confident that there is a road, whereas the red circle marks an area where the NLD-LinkNet34 predicts road with a high probability. The ensemble model then combines both predictions as can be seen by the two circles.

VI. DISCUSSION

The ResUNet and LinkNet34 models, both using a pre-trained ResNet34 backbone, exhibit similar F1 scores, as shown in Table I. However, using a deeper backbone like ResNet101 in D-LinkNet degrades performance due to probably being better suited for higher resolution images. The model introduces excessive noise, resulting in unsatisfactory results. Despite the performance degradation, its high IoU suggests a strong similarity between the predicted and ground truth regions. In NLD-LinkNet34, the addition of the non-local blocks improves the ability of the model to gather fine-grained details. Meanwhile, separable convolutions in D-Link101 SepConv reduce the complexity of D-LinkNet101, allowing it, at least in our use-case, to focus on capturing meaningful, generalizable patterns, thus leading to improved performances compared to using standard convolutions. Furthermore, the use of separable convolutions greatly reduces the memory requirements on the GPU and thus speeds up training and inference. This is an encouraging result especially for running the models on edge devices. As expected, the ensemble, profiting from the complimentary nature of the novel models we propose (see Figure 2), improves the F1 public score. Contrary to expectations during training, the combination of BCE and Focal loss doesn't yield the desired performance increase. Table II confirms that BCE remains the best choice.

VII. CONCLUSION

In this paper, we tackle aerial image road segmentation using a novel D-LinkNet variation. We integrate non-local operations and separable convolutions to preserve intricate image details and enhance computational efficiency. For future research, we will explore D-LinkNet101 SepConv's resource efficiency on platforms with greater computational power, testing larger batch sizes. While our approaches show promise in terms of efficiency, we suggest further fine-tuning of the proposed models to enhance performance.

REFERENCES

[1] M. Silvagni, A. Tonoli, E. Zenerino, and M. Chiaberge, "Multipurpose uav for search and rescue operations in mountain avalanche events," *Geomatics, Natural Hazards and Risk*, vol. 8, no. 1, pp. 18–33, 2017.

[2] M. W. Ulmer and S. Streng, "Same-day delivery with pickup stations and autonomous vehicles," *Computers & Operations Research*, vol. 108, pp. 1–19, 2019.

[3] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, 2020.

[4] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*, vol. 52, pp. 1089–1106, 2019.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[6] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4.

[7] V. Badrinarayanan, A. Kendall, and R. C. SegNet, "A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, vol. 5, 2015.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015.

[9] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[10] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation," 2019.

[11] L. SIfre and S. Mallat, "Rigid-motion scattering for texture classification," 2014.

[12] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation," 2017.

[13] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[15] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2. Ieee, 2005, pp. 60–65.

[16] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[17] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 192–1924. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPRW.2018.00034

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[21] "Models and pre-trained weights — Torchvision 0.15 documentation." [Online]. Available: https://pytorch.org/vision/stable/models.html

[22] Y. Wang, J. Seo, and T. Jeon, "Nl-linknet: Toward lighter but more accurate road extraction with non-local operations," 2020.

[23] M. Yi-de, L. Qing, and Q. Zhi-Bai, "Automated image segmentation using improved pcnn model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.* IEEE, 2004, pp. 743–746.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.