Deep Learning for Road Segmentation

Aziz Laadhar, Mohamed Charfi, Yassine Chaouch
École Polytechnique Fédérale de Lausanne (EPFL)

Abstract

This paper investigates the use of three deep learning models (U-Net, Link-Net, GC-DCNN) for automatic road segmentation in satellite images. It focuses on enhancing model performances using a diverse dataset, which includes data augmentation applied to the original dataset and the incorporation of public datasets with appropriate modifications. Additionally, we explore post-processing methods to further improve segmentation results. The results show that a variant of Linknet achieved the best F1 score of 88.7% on the test dataset.

I. Introduction

Automatic road segmentation from remote sensing images is an important research hot-spot in the remote sensing and pattern recognition fields. It plays a crucial role in urban planning and traffic management, environmental monitoring, emergency response planning and many other applications. Traditionally, labeling roads in these images manually was laborious and time-consuming. In the rapid development of computer vision, many attempts to use machine learning for automated road detection in remote sensing images didn't meet desired accuracy levels. However, the emergence of deep learning, especially convolutional neural networks (CNNs), has led to breakthroughs in tasks like image classification [1], object detection [2], and semantic segmentation [3]. Several iterations of CNNs have been developed through the years that kept improving the state of the art models used in segmentation. This work investigates the performances of three of these models and reproduce the most recent one in order to compare their results and adds post-processing methods to better adapt them to our needs. This work will also collect data from several sources in the aim to generalize our model by including different regions across the globe in our data. The rest of the paper is organized as follows. Section II describes the data collecting and engineering. Section III describes the entire architectures of the models used IV describes the loss functions, changes and extensions applied. Section V presents the evaluation results and several ablation experiments. Lastly, Section VII discusses the ethics implication of the project and VI concludes this study.

II. Data Engineering

The project utilizes a collection of 100 satellite images, each measuring 400 x 400 pixels and featuring the standard RGB color channels, sourced from GoogleMaps. These images are paired with ground truth images, which are black and white masks where white indicates the presence of a road. To illustrate, an example of a satellite image next to its ground truth provides insight into the real-world complexity of this task.

The testing data contains 50 images and differs from the training set in terms of the image size, presenting images of 608 x 608 pixels. In terms of data characteristics, the dataset reveals several challenges. Variations in road color and texture, along with environmental factors like shadows and trees or sometimes elevated railway completely covering the roads, add complexity to the segmentation process. Moreover, the dataset includes roads in various orientations, predominantly vertical or horizontal, which could lead to orientation bias in the model. These factors, combined with the presence of similar structures like walkways and parking lots, underscore the importance of a robust and nuanced approach to road segmentation in satellite imagery.

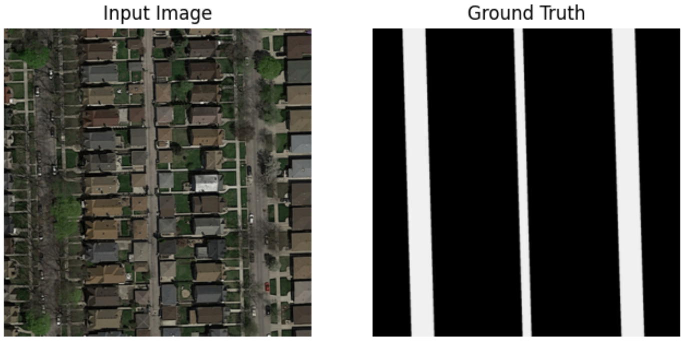

Figure 1: An example of the dataset where on the left we have the original satellite image and on the right its ground truth image

A. *Data Pre-processing and Augmentation*

To enhance our dataset, we employed multiple distinct strategies to increase our training sample space.

*1) **Data Augmentation***

We used the existing images to generate new ones. This was done in two different ways. The initial approach involved random rotations and adjustments to brightness and contrast to introduce variability in lighting and orientation. This randomness in transformation encourages the model to learn features invariant to such changes, which are commonplace in real-world scenarios. Each image was augmented 3 times randomly resulting in 300 new images in total.

The second strategy focused on deterministic rotations at specific angles 15, 30, 45, 60, 90, 180, and 270 degrees. This choice was informed by an analysis of the training dataset, which indicated that the majority of roads were aligned either perfectly horizontally or vertically. By rotating images at these fixed angles, we ensure the inclusion of road orientations that the model might encounter beyond the typical cardinal directions. In addition to rotation, we added some noise through *random Gaussian with 5% standard deviation* and *Pepper and Salt with 1% corruption ratio*. This resulted in 700 new augmented training images.

In both augmentation methods, care was taken to preserve the integrity of the image data. A reflection mode was utilized to fill the corners that emerge after rotation, avoiding the introduction of artificial black borders that could mislead the model. Through these methods, our dataset has been effectively expanded and diversified to achieve a total of the initial 100 + 300 + 700 = 1100 images, laying a robust foundation for the subsequent phases of model training and evaluation.

B. Additional Datasets

1) Datasets Extraction

Since the provided data was only taken from the same region of the world, to generalize our model we tried adding three different publicly available datasets:

- **Massachusetts Dataset**[4]: This dataset consists of 1171 aerial images of the state of Massachusetts. Each image is 1500×1500 pixels in size, covering an area of 2.25 square kilometers. Despite its large size this dataset was not used since the roads were all the same width, which didn't match our main dataset.
- **Deep-Globe Dataset**[5]: This dataset contains 803 satellite imagery in RGB of size 2448x2448. This data was also disregarded since it had mostly rural roads and most of them weren't considered as roads in our original dataset.
- **Learning Aerial Image Segmentation From Online Maps Dataset** [6]: Dataset consits of 3379 satellite images of different shapes from different urban data from cities such as Chicago, Zurich, Berlin, Paris, Potsdam, and Tokyo. This was the only data aligned with our goal so we proceeded to addapt it to our needs.

2) Datasets Pre-processing

When exploring The Learning Aerial Image Segmentation From Online Maps dataset we observed that the images resolution differs from our original dataset. This suggests an image resizing by a factor of 3.3 to ensure consistency. Additionally, the dataset presented another challenge, its labels contained segmentation for both roads and buildings. To address this, we created a mask that isolates the blue pixels as illustrated in figure.[2].

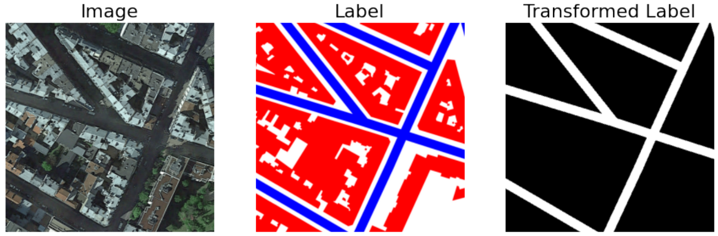

Figure 2: Learning Aerial Image Segmentation From Online Maps Dataset Image, Label, and Processed Label

III. Models

A. U-Net

- U-Net is renowned for its effectiveness in biomedical image segmentation, which is in some way similar to road segmentation. This model, illustrated in figure [3], consists of two key sections: an encoder and a decoder. The encoder reduces the size of the image and finds important features. The decoder then uses these features to make the image big again, like the original. They are connected by skip connections. Our design of U-Net is based on the work of [7]Ronneberger, Philipp Fischer, and Thomas Brox.

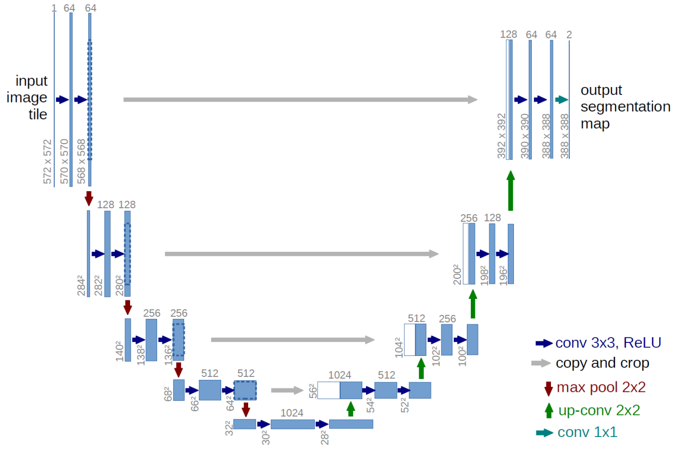

Figure 3: Unet Architecture and its Components.

B. Link-Net

LinkNet, as introduced in [8], brings a novel approach to semantic segmentation by efficiently connecting encoder and decoder layers, effectively preserving spatial information lost during downsampling. This design not only enhances segmentation accuracy but also ensures a more parameter-efficient network conducive to real-time applications. D-LinkNet [9] improves upon this by replacing pooling with dilated convolution layers to retain detailed information. Non-Local LinkNet [10] enhances the concept further with non-local neural operations, capturing long-range dependencies. As shown in Figure [4] LinkNet inputs must be a multiple of 32 x 32 for the 5-stage encoding and decoding with a stride of 2. To maintain contextual information, we resize inputs to 408 x 408. The encoder uses consecutive 3×3 convolutional layers to compactly represent essential features, while the decoder employs 1×1 convolutions for refinement and 3×3 convolutions for detailed segmentation map reconstruction, restoring lost spatial information.

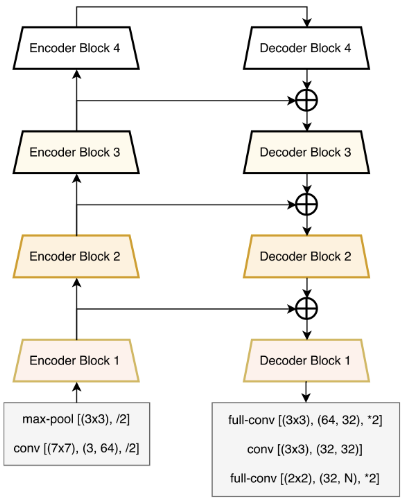

Figure 4: LinkNet Architecture and its Components.

C. GC-DCNN

Although deep CNN-based methods have progressed considerably, existing deep learning-based methods fail to address the view occlusion problem to obtain coherent and smooth segmentation results because they are simple applications of CNN to the road segmentation task. This is why we also opted for the Global Context-based Dilated CNN (GC-DCNN) a neural network introduced by Zhang et al.[11]. It's important to note that we independently reconstructed this architecture from the ground up, coding it entirely from scratch. We made this choice because we could not find publicly available code for the GC-DCNN model, and we believed it was essential to ensure our implementation closely adhered to the model as described in the original paper. During this process, we implemented various modules such as:

- **Residual Diluted Block** The RDB unit, shown in Fig.[5], uses pre-activation feature maps, undergoes BN, ReLU activation, and three consecutive 3x3 convolutions. The last two convolutions use dilated convolution with a ratio of 2. A shortcut connection, inspired by the residual block, adds the input to the final feature maps.
- **Pyramid Pooling Module** Zhao et al.'s[12] PPM (Pyramid Pooling Module) enhances scene parsing. PPM utilizes four levels with global average pooling, sub-region pooling, and a 1x1 convolutional layer to maintain consistent weight across different-level features. The low-dimensional maps undergo up-sampling and are concatenated with input features to obtain the final global context features.

So the model architecture as can be seen in Fig.5 is a U-type network that is composed of an encoder network, PPM, and a decoder network. The encoder part contains two initial convolutional layers that convert the input RGB image into the primary high-dimensional features, and then the features are fed into subsequent blocks to generate the multiscale hierarchical feature. Instead of using the max pooling operation to downsample the feature maps, the model uses the stride of the first convolutional layer to 2 in each RDB unit. Therefore, the total stride of the encoder network is 8. PPM works similar to a bridge, which uses the final features of the encoder as the input and produces the features with global context representation for the decoder part. The decoder network is also composed of three special RDB units, whose dilation ratios are set to 1 for refinement. These block units are connected via the upsampling operation, which is implemented by the transposed convolution operator in Pytorch with a kernel size and stride of 2. The upsampled features in each level are concatenated with the corresponding hierarchical features of the encoder in the depth dimension to obtain features with rich spatial details. The last 1x1 convolutional layer converts the high-dimensional features into single-channel features, and the loss function is calculated with the ground truth after the sigmoid activation function.

IV. Methods

A. Loss Function

In our exploration of road segmentation from satellite imagery, the choice and analysis of loss functions were pivotal, with a particular focus on Binary Cross-Entropy (BCE) and Dice loss. The Binary Cross-Entropy loss, represented as L_{BCE} is a distribution-based loss function and is expressed as:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i))$$

where $\hat{y}_i$ is the true label, $\hat{p}_i$ the predicted probability, and N the total number of pixels. Conversely, the Dice loss, a region-based loss function, is denoted as L_{Dice} and is defined as :

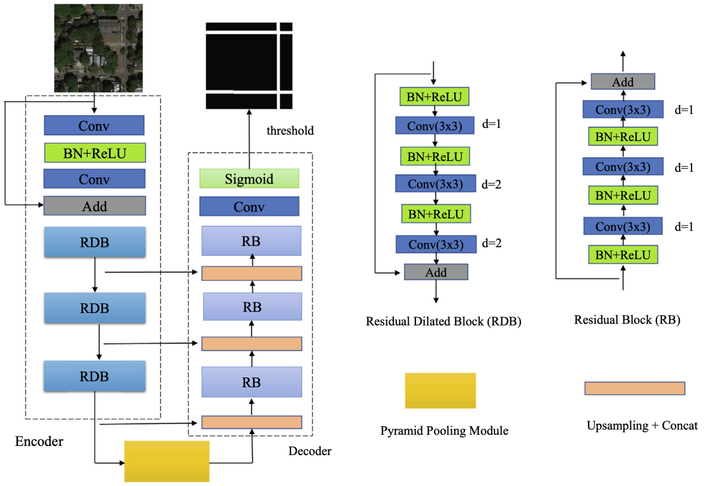

Figure 5: Global Context-based Dilated Convolutional Neural Network Architecture

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i\hat{p}_i + 1}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N}\hat{p}_i + 1}$$

This function emphasizes the overlap between the predicted segmentation and the ground truth. Our comprehensive testing included various combinations of these loss functions, including hybrid approaches. Notably, the region-based Dice loss alone demonstrated superior performance in effectively delineating road segments, highlighting the importance of selecting an appropriate loss function that aligns with the specific challenges of the segmentation task.

B. Hyperparameters and K-fold

In our road segmentation study, we opted for 5-fold cross-validation, which provides a balanced assessment without the extensive computational load of exhaustive cross-validation. The learning rate was set at $3 * 10^{-4}$ after empirical testing, which offered a good compromise between convergence speed and stability. Regularization was managed through inherent batch normalization in our model architectures and a dropout rate of 0.2 in GCDCNN to mitigate overfitting. These choices were aimed at delivering a reliable and computationally efficient road segmentation solution.

C. Test time augmentation(TTA)

In our pursuit to refine the predictive capacity of our model we explored the potential of Test Time Augmentation [13]. TTA is a technique designed to enhance the robustness and accuracy of the model's inference phase. By applying the *d4 transform*, which encompasses a horizontal flip and rotations at angles of 0, 90, 180, and 270 degrees, we were able to simulate a variety of viewing perspectives and orientations that roads may present in real-world scenarios. Subsequently, we restored the orientation of the model's outputs to align with the original images, ensuring consistency in our predictions and averaged the probabilities across all augmented predictions for each pixel, thereby consolidating the model's confidence in its inference.

D. Cropping and Predicting

Given the dissimilarity in shapes between the training and testing images, we investigated the potential of partitioning the test images into overlapping patches, each having the same size as the training images. Subsequently, we processed each patch individually and assembled the overall prediction by computing the average of predictions within overlapping regions.

V. Results

A. Experiments

1) Experimental settings

- **Implementation details** For this experiment we will use a LinkNet model with ResNet18 as backbone which has been pretrained on the ImageNet [14] dataset to compare it to a UNet and GC-DCNN models which we implemented from scratch using the Pytorch framework. These three models will be trained on NVIDIA A100 Tensor Core GPU using 3843 images of size 400x400 randomly sampled from the ”Learning Aerial Image Segmentation From Online Maps Dataset” with our initial data from AIcrowd and validated on 100 images with size 400x400 from the same dataset. Batch size is set as 10 for each model, and the Adam optimizer with betas of 0.9 and 0.999 is adopted to optimize all the models with the original learning rate of 3e-4 and using Dice Loss function. We train the model for 30 epochs and drop the learning rate by a factor of 0.095 every 5 epochs. In the inference process, we set the threshold at 0.5, indicating that the final value of each position in the output probability map is 1 if the predicted value is greater than the threshold and 0 otherwise.
- **Evaluation Metrics** We mainly use F1 score which combines the two numbers of precision and recall to evaluate. The reported results are the Score and Secondary Score obtained when testing on AIcrowd’s test dataset mentioned in II.

2) Predictions and Results

Figure 6 presents the training and validation loss history during the training process with different models. We can see that the LinkNet starts at a lower loss than the other models and is the fastest model to converge during training. Such result is expected since it is the only pre-trained model of the three. We can also see that the Unet model is the second fastest model to converge due to its simplicity and fewer number of parameters compared to the GC-DCNN (35.66 M vs 38.57 M). Lastly we can see that the GC-DCNN model couldn’t show its true potential in these experimental settings which can be due to the low number of epochs and size of data.

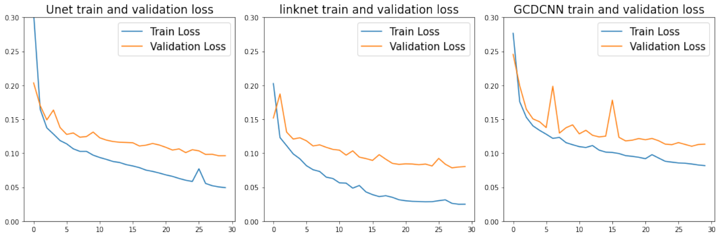

Figure 6: Comparaison of the learning curves of the models trained on 3843 Images of shape 400x400 and validated on 100 images with the same shape from the same dataset

Table I presents the AIcrowd’s results report for the three models where Params is the number of parameters of the model and Speed is the time it takes to run on epoch in training. These results confirm that the LinkNet model was the best performing model since it achieved higher score than the other two and the three models combined where we take the average of the predictions of the three models. Hence, to further investigate the effect of our post-processing methods we added TTA and Cropping extensions to the LinkNet model. The TTA achieved less Score and Secondary score than the base model.However, the Linknet increased its Secondary Score by 0.001 when using the cropping techniqueIV-D.

Model	**Epochs**	**lr**	**Score**	**Sec. score**	**Params**	**Speed**
Unet	30	3e-4	0.883	0.930	35.66M	2m13s
GCDCNN	30	3e-4	0.874	0.925	38.57M	2m44s
LinkNet	30	3e-4	**0.887**	0.932	11.5M	1m38s
Combined models	30	3e-4	0.882	0.929	-	-
LinkNet + TTA	30	3e-4	0.875	0.929	-	-
LinkNet+Cropped	30	3e-4	**0.887**	**0.933**	-	-

Table I: AIcrowd Test Results Table For Models Trained on 3843 Images of shape 400x400 and Tested on 50 Images of shape 608x608

To get a pixel level perspective on the prediction we randomly choose an image from the test set and plot the prediction of each of the models we tested in figure 7. These images confirms the effectiveness of our post processing methods in these types of instances. The combination of the three models clearly outperformed each of the three models this is can be due to the spread of the miss-classifications of the Unet and Linknet for the parking lots for which the probability of each pixel has been divided by a factor of three. We also can notice the effect of the TTA that also cleared the miss-classifications of these parking lots but also miss-classified a portion of the road.

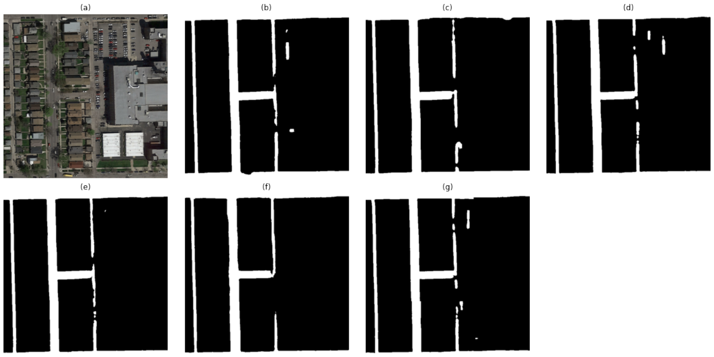

Figure 7: Pixel Level Results. (a): test image, (b): Unet result, (c): GCDCNN result, (d): linknet result, (e): combined result, (f): linknet with TTA result, (g): linknet with cropping result

VI. Conclusion

In conclusion, the research demonstrates the potential of deep learning models in accurately segmenting roads from satellite images. By exploiting U-Net, Link-Net, and GC-DCNN architectures potential in remote sensing and pattern recognition field, and incorporating data augmentation of diverse datasets, the study reveals significant insights on road detection using neural networks. The findings suggest promising directions for future research in automated road mapping using advanced machine learning methods.

VII. Ethics discussion

In our road segmentation project using satellite images, ethical considerations demand our undivided attention. Firstly, we must address the aspect of privacy. Satellite imagery can inadvertently capture private properties, raising concerns about intrusions into individuals’ privacy. To mitigate this, one potential solution is to apply image blurring techniques to obscure identifiable private locations. Second, it's important to consider how the labels produced by our machine learning algorithm are used. For instance, if the algorithm is applied to suggest routes, like in a map application, accuracy is critical. Incorrect predictions of roads, specifically false positives, could mislead users or potentially lead to unsafe routes. In the case of false positives, there are tangible consequences to consider. Identifying roads that do not exist in a region could result in individuals paying higher taxes for non-existent infrastructure, or conversely, it may prevent certain areas or municipalities from receiving the necessary investments to improve their infrastructure. Hence, the ethical implications of our road segmentation project extend beyond privacy concerns to encompass broader societal and economic impacts.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[2] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[4] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, University of Toronto, 2013.

[5] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[6] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6054–6068, 2017.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, 2015.

[8] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," *arXiv preprint arXiv:1707.03718*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1707.03718

[9] 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)*. IEEE, 2018, pp. 192–1924. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018_workshops/papers/w4/Zhou_D-LinkNet_LinkNet_With_CVPR_2018_paper.pdf

[10] Y. Wang, J. Seo, and T. Jeon, "Nl-linknet: Toward lighter but more accurate road extraction with non-local operations," *arXiv preprint arXiv:1908.08223*, 2019, iEEE Geoscience and Remote Sensing Letters (2020, to appear). [Online]. Available: https://doi.org/10.48550/arXiv.1908.08223

[11] M. Lan, Y. Zhang, L. Zhang, and B. Du, "Global context based automatic road segmentation via dilated convolutional neural network," *Information Sciences*, vol. 535, pp. 156–171, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025520304862

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2017.

[13] Test time augmentation tutorial. https://mmengine.readthedocs.io/en/latest/advanced_tutorials/test_time_augmentation.html. MMEngine. Accessed: Date you accessed the tutorial.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385 [cs.CV]*, 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1512.03385