CIL Road Segmentation Project Report 2024

Chia-Wen Chen, Ming-Han Lee, Moritz Dieing, Natcha Jengjirapas
Group: Meow
Department of Computer Science, ETH Zurich, Switzerland

***Abstract*—Accurate road segmentation is crucial for applications like navigation, autonomous driving, urban planning, and geographic information systems. Despite extensive research, achieving high accuracy with limited data remains challenging. This report aims to enhance U-Net-based architectures, specifically U-Net++, for road segmentation using a small dataset. We employ data augmentation, transfer learning with EfficientNet-B7, and advanced post-processing techniques, including Generative Adversarial Networks (GANs) [1], to achieve the f1-score of 0.93255, ranking 5th on the public leaderboard. Additionally, we propose Sutando referencing to improve GAN-based post-processing. Our evaluation shows that these strategies significantly enhance model accuracy and robustness, even with limited data, advancing road segmentation capabilities.**

I. Introduction

Acquiring accurate road information through segmentation is fundamental for various applications, including autonomous driving, navigation, urban development and planning, and geographic information systems (GIS). As a result, significant research on road segmentation has been conducted [2] [3], leading to the development of numerous methods and architectures with varying degrees of extraction accuracy. Among these, the U-Net architecture has emerged as one of the most popular choices for road segmentation due to its strong performance and low complexity. However, the effectiveness of U-Net in achieving the desired accuracy can vary when the amount of input data is limited.

In order to overcome data constraint, our study aims to enhance the performance and generalization of U-Net-based models, particularly U-Net++, through data augmentation on a small dataset of aerial images with ground-truth masks. We used methods such as horizontal and vertical flips, image rotation, and resizing to ensure the model can handle various road orientations and focus on different parts of the images.

To further address data limitations, we employed transfer learning using the EfficientNet-B7 encoder [4] in our U-Net++ segmentation model. EfficientNet-B7, a convolutional neural network [4], leverages a compound scaling method to efficiently scale the network's width, depth, and resolution. Initialized with pre-trained weights from ImageNet, this approach enhances the model's ability to generalize from limited data. The integration of Mobile Inverted Bottleneck (MBConv) layers further augments its feature extraction capabilities, crucial for our segmentation tasks.

Additionally, we implemented advanced post-processing techniques to mitigate issues such as disconnected road segments, and obscure boundaries. Specifically, we explored the use of Cycle Generative Adversarial Networks (CycleGAN) [5] and the Pix2Pix framework [6] for image-to-image translation, aiming to refine the predicted segmentation maps and improve their alignment with the ground truth.

Furthermore, we introduced Sutando referencing to enhance the post-processing task, as traditional methods typically condition the input noise solely on the input image, which does not yield satisfactory results for our specific needs. Sutando referencing involves incorporating the original aerial image into the input of the generator, but not the discriminator. This approach is crucial for models such as Pix2Pix, CycleGAN, and similar architectures, as it significantly enhances their performance. By providing additional contextual information, Sutando referencing aims to address these limitations and improve the accuracy and robustness of the segmentation outcomes.

Our experimental results demonstrates that combining U-Net++ with EfficientNet-B7, comprehensive data augmentation, and advanced post-processing techniques significantly improves road segmentation performance. The final model achieved a public score of 0.93255 on the Kaggle leaderboard, a substantial improvement over baseline models. This report highlights the potential of these methods to enhance the accuracy and robustness of segmentation models, even with limited datasets.

In summary, this study's contribution includes:

- We proposed Sutando referencing to modify the Pix2Pix and CycleGAN models for post-processing, refining the segmentation masks and improving overall model performance.
- We extensively used data augmentation and integrated EfficientNet-B7 with U-Net++, demonstrating a practical approach to enhance model performance with limited data.
- We achieved a f1-score of 0.93255, ranking 5th on the public leaderboard of the CIL competition.

II. Models and Methods

A. Data Preparation

The initial training set comprises 144 aerial images, each accompanied by corresponding ground-truth masks, with a resolution of 400x400 pixels. To enhance our model's performance and generalization capability, we aimed to

increase the training set size and diversity through extensive data augmentation techniques.

The initial images were augmented using several methods: horizontal and vertical flips with a 50% probability, random rotations between -90 and 90 degrees, followed by resizing to 384x384 pixels. These transformations ensured that the model could handle various orientations and focus on different parts of the images, thereby increasing its robustness.

B. Network Architecture

We utilize U-Net++ as our segmentation model, with EfficientNet-B7 [4] serving as the encoder. The encoder is frozen and initialized with pre-trained weights from ImageNet [7], allowing us to focus solely on training the decoder.

EfficientNet-B7. The EfficientNet architecture proposed by Mingxing Tan and Quoc V. Le [4] is a special CNN-based architecture which uses a novel technique called compound coefficient. Compound coefficient enables to efficiently scale up width, depth and resolution of the network depending on the input images. The main building block of the network is the Mobile Inverted Bottleneck (MBConv) layer, which combines convolutions and inverted residual blocks. Depending on the type of EfficientNet model, the number of consecutive MBConv layers varies. The EfficientNet-B7 which is used as an encoder in our architecture is the biggest model of the EfficientNet family.

U-Net++. Using the latent representations obtained by feeding the images through the encoder network, we train a U-Net++ architecture for the prediction of the segmentation masks. The U-Net++ architecture proposed by Zhou et al. in [8] is an architecture for segmentation tasks based on U-Net. The main improvement of U-Net++ is the introduction of Nested Skip Pathways. These are multiple connections from the encoder to the decoder which link different levels of both components to each other. This enables a better feature aggregation resulting in a reduced information loss and hence improves the accuracy of the segmentation masks.

For the implementation of our base model, we make use of the segmentation models library [9].

C. Post-processing

Despite high accuracy from our base model, the raw segmentation maps suffer from inaccuracies such as false positives, disconnected road segments, and obscure boundaries. We aim to address these issues by post-processing.

Gaussian Kernel. To address anomalous white points in our raw mask predictions, we applied Gaussian Kernel Smoothing, a denoising technique that averages pixel values using a weighted Gaussian kernel. Although it did not improve our performance score, it provided a valuable baseline for comparing subsequent post-processing techniques.

Image Translation. Image-to-Image translation is a task in computer vision that convert an image in one domain into another domain. In our road segmentation setting, we formulate the post-processing problem to be translating the predicted masks of our base model to the target masks. Since both domains reflect the roads of the same aerial map, we expect the model to learn the mapping effortlessly. We applied Pix2Pix [6] and CycleGAN[5] to our task.

Pix2Pix is a conditional generative adversarial network (cGAN) [10] framework where the primary goal is to learn a mapping from input domain to output domain using paired training samples. The training of the cGAN is the same process as an ordinary GAN with conditioned input noise. Namely, a process of the generator and the discriminator playing a minimax game:

$$\min_G \max_D E_{x\sim p_{\text{data}}(x)}[\log D(x,y)]+E_{z\sim p_z(z)}[1-\log D(x,G(x,z))]$$

where G is generator, D is discriminator, (x, y) is a sample pair drawn from the data and z is random noise. The objective of Pix2Pix is formulated as:

$$L_{pix2pix}(G,D) = L_{cGAN}(G,D) + \lambda L_1(G)$$

$$L_{cGAN}(G,D) = E_{x,y}[\log D(x,y)]+E_{x,z}[1-\log D(x,G(x,z))]$$

$$L_1(G) = E_{x,y,z}[\|y - G(x,z)\|_1]$$

The L1 loss encourages the generator to generate the samples that are similar to the ground truth. Although L1 loss obscuring the samples violates our goal of enhancing the edges, we observed that this could be mitigated by training for longer epochs.

CycleGAN, on the other hand, aims to deal with the same task when there is no existing paired dataset. CycleGAN is composed of two sets of GANs and trained with an additional cycle loss term:

$$\begin{aligned} L_{cycleGAN}(G,D) = {} & L_{GAN}(G,D_Y,X,Y) \\ & + L_{GAN}(F,D_X,X,Y) \\ & + \lambda L_{cycle}(G,F) \end{aligned}$$

$$\begin{aligned} L_{cycle}(G,F) = {} & E_{x\sim p_{data}(x)}\left[\|F(G(x)) - x\|_1\right] \\ & + E_{y\sim p_{data}(y)}\left[\|G(F(y)) - y\|_1\right] \end{aligned}$$

where G and F are generators for translating images back and forth, and D_{domain} is the discriminator for that domain. Intuitively, cycle loss acts as a regularization term that encourages the model to maintain consistency by ensuring that translating an image to another domain and then back to the original domain results in the same image. This helps to preserve the key features of the input image and prevents the model from making excessive changes. Since we have existing paired dataset, we applied the paired dataset to CycleGAN.

Sutando Referencing. Sutando, originates from the popular anime "Jojo's Bizzare Adventure", is an energetic form of

entity existing behind its owner and cannot be seen by other people unless the owner shows it. Although Pix2Pix and CycleGAN work in the normal image-to-image translation setting, we found that directly applying it to the post-processing task would worsen the performance of our base model. To solve this, we propose Sutando referencing, a technique that additionally introduces original aerial map to the input of the generator but not the discriminator (thus Sutando). The idea behind is quite simple, to let the generator has references when editing input masks. We assume it helps the generator learn to inpaint the disconnected road by looking at the original image. Therefore, our final objective for Pix2Pix is:

$$\begin{aligned} L_{cGAN}(G, D) &= E_{x,y}[\log D(x, y)] \\ &\quad + E_{x,z}[1 - \log D(x, G(\mathbf{o}, x, z))] \\ L_1(G) &= E_{x,y,z}[\|y - G(\mathbf{o}, x, z)\|_1] \end{aligned}$$

where we introduce $\mathbf{o}$ to be original aerial image. The idea turns out to work quite well. We recorded a 0.924% improvement to the public test score of Pix2Pix model. We will look into this in the results and discussion section.

D. Training Details

Base Model Training. To train our base model, we divided the original training data into an 80% training set and a 20% validation set. For the sake of comparability, we decided to train the base model as well as the baseline models using a fixed number of 50 epochs with the batch size set to 4. We used a learning rate of 0.001 and the Adam optimizer to facilitate the training process. The sigmoid activation function was applied in the network, and we utilized Dice loss function to optimize the performance of our model. In Section II-E, we will discuss our choice of loss function.

In order to achieve the highest possible score in the Kaggle competition, we increased the training duration for our final models up to 150 epochs. All scores resulting from this longer training procedure are marked with a subscript $+$ in the report.

Pix2Pix Training. We use U-Net++ as our generator and conditioned the noise on the predicted masks of our base model as well as the aerial map and the discriminator is a PatchGAN implemented as a convolutional neural network. We trained the discriminator and the generator alternatively. λ is set to 1000 since we would like to force the generator to inpaint the disconnected roads and remove noises. The model was trained for 50 epochs with a batch size of 1. We used the Adam optimizer with a learning rate of 0.0002.

E. Dice Loss

This dataset, like most road segmentation datasets, is typically imbalanced, with more background areas than road areas. To address this, we employed Dice Loss[11] during training to mitigate the model's tendency to favor background pixel predictions. Dice Loss, as defined in Equation (1), measures the overlap between the true mask (y) and the predicted mask ($\hat{p}$), rather than performing pixel-wise classification like Binary Cross-Entropy. We evaluated the effectiveness of Dice Loss, as shown in Table I.

$$DiceLoss(y, \hat{p}) = 1 - \frac{2y\hat{p} + 1}{y + \hat{p} + 1} \tag{1}$$

F. Other Unsuccessful Approaches

We attempted normalization, the addition of an external dataset, and dilation, but none yielded positive results. Due to the varying color and contrast in our dataset, we tried histogram normalization and adjustments using ImageNet's mean and standard deviation, but these did not improve the score. This may be due to the dataset's inherent characteristics or our model's robustness to color variation. We also integrated the Massachusetts Dataset[12] to address data limitations, but it slightly lowered performance, so we excluded it from the final submission.

Incorporating dilated convolutions improved our basic U-Net's performance by about 4%, from 0.82620 to 0.85877. However, after upgrading the encoder to a pretrained ResNet18, adding dilation to the decoder reduced performance. The decoder's role is to upsample and refine features for accurate segmentation mask reconstruction, and dilation might cause a loss of fine details and disrupt skip connections. Future research could explore alternating regular and dilated blocks, using dilation only centrally, or fine-tuning the dilation rate to address these issues.

III. Results and Discussion

Our final method before ensembling achieved a public score of 0.92851_+, representing a 7.32% improvement over the ETHZ CIL Road Segmentation 2024 baseline and a 12.38% enhancement compared to our basic U-Net trained from scratch. On the Kaggle leaderboard, our final performance showed a 12.87% improvement over the basic U-Net. We conducted a series of ablation experiments to evaluate the contributions of different components, including model architecture, loss functions, data augmentation, and notably, post-processing methods. The results of these experiments are detailed in Table I and II. Note that the experiments were rerun on a separate day and thus have slight difference with our best public test score in Table III. We will discuss them in two sections: Base Model and Post-processing.

A. Base Model

The initial U-Net model achieved a public score of approximately 0.82620. Upgrading to the U-Net++ with pretrained EfficientNet-B7 encoder (EffNet-b7) improved the performance significantly, reaching a score of around 0.91668. Further enhancements, including the use of more stable loss functions and comprehensive augmentation techniques (Aug), elevated the score to approximately 0.92333.

Although the model's performance is already impressive, there remains potential for further refinement and optimization to achieve even higher accuracy, which led us to explore post-processing techniques.

Model	Loss	EffNet-b7	Aug	Public Test Score
U-Net	BCE			0.82620
U-Net++	BCE	V		0.91668
U-Net++	DICE	V		0.91788
U-Net++	DICE	V	V	**0.92333**

Table I: Ablation Studies for Base Model

B. Post-processing

Table II demonstrates the effectiveness of Sutando referencing on GAN-based image-to-image translation models applied in post-processing tasks. While the original CycleGAN and Pix2Pix worsen the scores, incorporating Sutando referencing leads to substantial improvements, making these kinds of models become effective in the post-processing tasks. Specifically, our base model scores 0.92242 without any post-processing. While after applying CycleGAN and Pix2Pix for post-processing worsen the scores to 0.92188 and 0.91402, with Sutando referencing, these two models start to enhance the results and score 0.92280 and 0.92313, respectively. We chose Pix2Pix with Sutando Referencing as our final post-processing module due to its superior performance in inpainting disconnected roads, straightening them, and maintaining the high resolution of the original mask, as shown in Figure 1 (red box). While the original Pix2Pix paper used $\lambda = 100$, we chose $\lambda = 1000$ to enforce the generator focus on inpainting disconnected roads and removing noises, which further enhanced the score from 0.92313 to 0.92326.

Using GAN-based image-to-image translation models with Sutando referencing for post-processing, in particular, highlighted the effectiveness of generative adversarial networks in enhancing segmentation accuracy and robustness, suggesting that similar approaches could be applied to other post-processing tasks for improved outcomes.

Methods	Public Test Score
Base Model - UNet++ w/Aug, External Dataset	0.92242
Gaussian Kernel Smoothing	0.92242
CycleGAN	0.92188
CycleGAN w/Sutando Referencing	0.92280 ↑
Pix2Pix	0.91402
Pix2Pix w/Sutando Referencing	0.92313 ↑
Pix2Pix w/Sutando Referencing, large λ (our method)	**0.92326** ↑

Table II: Ablation Studies for Post-processing: only after applying Sutando referencing the GAN-based image-to-image translation models start to improve the predicted masks of the base model (marked by ↑, otherwise no improvements or worsen)

With the goal of further increasing our prediction score we applied an ensemble method to our best scoring submission using our best performance models

Methods	Public Test Score
U-Net++ w/CycleGAN	0.92640_{+}
U-Net++ w/Aug	0.92731_{+}
U-Net++ w/Aug, Pix2Pix, Sutando Referencing, large λ	0.92851_{+}
Ensemble	**0.93255**

Table III: Top 3 Models and Final Public Test Score

The ensemble method simply iterates patch by patch through the submission and always selects the dominant prediction amongst the three models. This increased our public score from 0.92851 to 0.93255.

Despite promising results, our image-to-image translation models, while effective on training masks, failed to generalize to the validation set, resulting in minor public test score improvements. This issue stems from inconsistent defects in predicted masks, such as white point removal and disconnected road inpainting. Future research will aim to produce consistent defects in the base model or improve post-processing generalization. Continuous validation and testing on independent datasets are also essential to ensure robustness and mitigate overfitting.

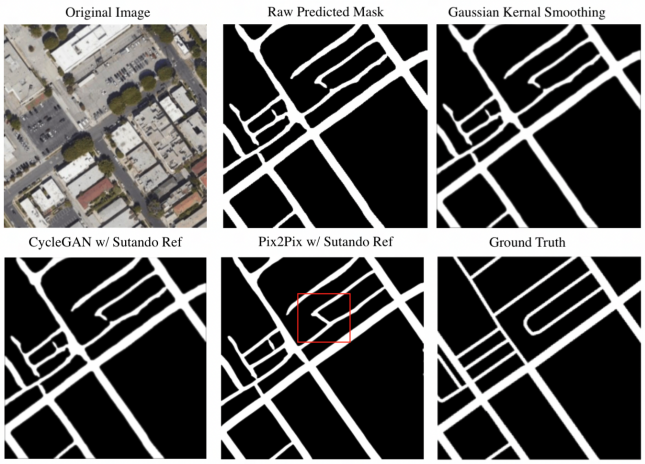

Figure 1: Example samples for post-processing techniques.

IV. Summary

This research explored novel methods for the aerial road segmentation task, specifically addressing the challenges of limited data and refining raw segmentation masks. We enhanced data diversity through augmentation at reasonable computational effort. Additionally, we investigated innovative post-processing approaches by applying Sutando referencing to the Pix2Pix model, comparing its performance to the conventional Gaussian kernel smoothing method. These findings demonstrate the potential of generative models for advanced refinement in road segmentation tasks.

REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[2] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *European Conference on Computer Vision (ECCV)*, 2010. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-15561-1_6

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018. [Online]. Available: https://arxiv.org/abs/1802.02611

[4] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," 2018. [Online]. Available: https://arxiv.org/abs/1807.10165

[9] P. Iakubovskii, "Segmentation models pytorch," https://github.com/qubvel/segmentation_models.pytorch, 2019.

[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014. [Online]. Available: https://arxiv.org/abs/1411.1784

[11] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*. Springer International Publishing, 2017, p. 240–248. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-67558-9_28

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

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

CIL ROAD SEGMENTATION PROJECT REPORT

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Chen	Chia-Wen
Lee	Ming-Han
Dieing	Moritz
Jengjirapas	Natcha

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich, 30.07.2024	陳佳雯
	李明翰
	M. Dieing

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.