Computational Intelligence Lab - Semester Project 3: Road Segmentation

Sapar Charyyev* Shinjeong Kim* Konstantinos Stavratis* Aidyn Ubingazhibov*
Department of Computer Science, ETH Zurich, Switzerland
{scharyyev, shikim, kstavratis, aubingazhibo}@student.ethz.ch
Team name on Kaggle: assk

***Abstract*—Road segmentation from aerial images is an important problem that has applications in urban planning, infrastructure development, and transportation. This paper presents our solution for the ETHZ CIL Road Segmentation 2023 challenge which requires segmenting roads in images obtained from google maps. Starting from a U-net baseline, we propose to replace its encoder with an EfficientNet-b7 encoder to increase its capacity. Additionally, to deal with data scarcity we propose to use a pretraining strategy, data augmentation, ensemble learning, and test time augmentation. We further refine the results with Conditional Random Fields. With all these components, our solution demonstrates promising results in aerial road segmentation.**

I. Introduction

In recent years, the rapid advancement of aerial imaging technology has revolutionized various industries, including urban planning, infrastructure development, and transportation. One significant application of this technology lies in the extraction of road networks from aerial images. Road segmentation, the process of identifying and delineating roadways within these images, plays a pivotal role in enhancing navigation systems, optimizing transportation networks, and facilitating comprehensive infrastructure analysis.

In recent years, with the rise of deep learning, Convolutional Neural Networks (CNNs) have emerged as the state-of-the-art approach for image segmentation tasks, including road segmentation [1] [2] [3] [4]. Fully Convolutional Networks (FCNs) have demonstrated impressive performance by preserving spatial information during the segmentation process, enabling better understanding of road structures in aerial images. Moreover, the integration of skip connections and encoder-decoder architectures, such as U-Net [5] and DeepLab [6], has shown promising results in improving segmentation accuracy and handling scale variations in road networks.

However, challenges persist in the road segmentation domain, particularly due to small datasets and time-consuming labeling process. This report presents our solution to the ETHZ CIL Road Segmentation 2023 challenge organized by ETH Zurich. Our solution consists of the following components:

- U-Net model with EfficientNet-b7 backbone
- Pretraining on the Massachusetts dataset
- Extensive data augmentations
- Model ensembling
- Conditional Random Fields
- Test time augmentation

The rest of the paper is organized as follows. Section II describes additional datasets we used for the project, section III contains our proposed solution, and section IV contains experiments and results.

II. Data

Selection: Collecting as much data as possible for a task at hand prior to training a deep neural network model (or any statistical model) is crucial for achieving high performance and robustness [7]. That was the motivation for complementing the provided by the competition dataset with two more similar datasets: the Massachusetts Aerial Image Labelling Dataset (henceforth mentioned as "former") and the Road extraction dataset from DeepGlobe Challenge (henceforth mentioned as "latter").

After conducting several experiments with both of them, it was observed that (pre)training with the former set bettered the performance of the target neural network; in contrast, the latter set (even if used in conjunction with the former set) made the target model underachieve. It was deduced that this stems from the nature of the sceneries captured in the photographs: the former contains images captured in urban areas –as does the provided dataset. On the other hand, the latter contains images of rural areas (e.g. fields or areas with very low population density). As such, we hypothesized that (pre)training with the latter has the model diverging from the optimal solution.

Consequently, a conscious decision to utilize only the former dataset was made, after withholding questionable samples using a customly-created filter: samples containing more than 10% empty space (pixel values of 255) are filtered out, to maintain data quality and minimize noise during pretraining.

Specifications: *Massachusetts Roads Dataset.* The dataset comprises 1108 aerial images for training, 16 images for validation, and 49 images for testing. Each image has a size of 1500 $\times$ 1500 pixels, accompanied by binary road segmentations serving as labels. For the training process, we

*Order of the authors does not signify contribution.

combined the original training and validation sets, treating them as the training set. Meanwhile, the testing set is preserved for validation purposes.

As outlined above, the images are of higher resolution than that of the competition. To comply with the latter's images specifications, crops of size 400×400 are randomly sampled from the images during the pretraining process. To create the validation set, we divided images into 9 non-overlapping patches, each sized 400×400 pixels.

Folding: Considering the limited amount of data available, it is crucial to avoid losing any additional samples from the training set for the validation set. Therefore, we split the data into 5 folds of roughly equal size and train separate models on each, which will be ensembled during inference.

III. Models and Methods

This section provides a description of the methods selected for tackling the problem, progressively leading up to our final solution.

A. Baselines

Prior to endeavoring a novel solution to the problem, a standard to compare to and overcome had to be implemented first. As per the project's instructions, at least two such baseline methods were implemented. In the spirit of simplicity and explainability, before opting towards Convolutional Neural Networks (CNNs), which have yielded remarkable results in the last decade, it was suggested that a more imperative-style baseline first; as such, a method based on thresholding pixel values was tested first. After this, methods based on CNNs are presented.

Threshold Based Segmentation: The purpose of the first baseline utilized is a Hue Saturation Value (HSV) scheme, which classifies each pixel of an image based on the pixel's HSV. More specifically, we find the values $h_l, h_u, s_l, s_u, v_l, v_u$ such that the pixel is classified as foreground if the following inequalities hold simultaneously.

$$(h_l \leq h \leq h_u), (s_l \leq s \leq s_u), (v_l \leq v \leq v_u)$$

Where h, s, v are HSV values of the pixel. The reason we opted for HSV values instead of RGB is that the former separate color information (chroma) from intensity (luma), which leads to more robust color thresholding [8]. To determine optimal values for $h_l, h_u, s_l, s_u, v_l, v_u$, grid search is enacted. To perform a more effective grid search, a heuristic method of initializing the grid was devised. Instead of randomly initializing grid values, firstly the mean HSV values, i.e. h_m, s_m, v_m, of all foreground pixels across the entire training set are computed. The grid is then formed by evenly-spaced intervals dictated by the standard deviation, which are centered around h_m, s_m, and v_m. As an accuracy metric, F1 score is used when selecting optimal $h_l, h_u, s_l, s_u, v_l, v_u$ values.

U-Net: Furthermore, we implement vanilla U-Net [5] as our second baseline. It is a Fully Convolutional Neural Network model that was first proposed for medical image segmentation. Our implementation of this model consists of a contracting path and an expansive path 1. The contracting path consists of 4 blocks, each of which consists of a batch normalization, 3x3 (with padding) convolution, and max pooling operations in succession. As a result, each block downsamples features by a factor of 2. The expansive path reverts the process by incorporating inverse-like operations: it consists of 4 blocks which in turn each consist of batch normalization, upsampling, and 3x3 convolutions. Additionally, features from the contracting path are concatenated with features from an expansive path, which is what distinguishes the U-Net architecture. Finally, the network head consists of three 3x3 convolutions followed by a 1x1 convolution that produces final masks. All outputs of a convolutional layer are passed through a ReLU activation function [9] to capture non-linearities. The model is trained with an ADAM optimizer [10] and smooth L1 loss.

B. Proposed solution

Model Architecture: As our main model, we replaced the backbone of U-net with the backbone of EfficientNet-b7 [11] model. EfficientNet is a Convolutional Neural Network (CNN) architecture that uses a *compound coefficient* to uniformly scale all dimensions of network depth, width, and resolution. While conventional CNNs use arbitrary scales for these factors, EfficientNet uses a set of fixed scaling coefficients. This is justified by the intuition that more layers are needed in case the image is larger for larger receptive field.

Data Augmentation: To reduce overfitting, we applied several geometric and non-geometric data augmentations during training. The geometric augmentations we used are random rotation, shift, scale and horizontal flip and non-geometric ones are blur, gaussian noise, image compression and random brightness.

Transfer Learning: The motivating idea behind transfer learning is that a pre-trained model trained on a large source dataset is fine-tuned on a smaller target dataset. This approach allows the target model to exploit the knowledge learned from the source data, which can improve its performance on the target task.

Towards this end, we trained U-net with EfficientNet-B7 backbone on Massachusetts Roads Dataset. The backbone of the model is initialized with ImageNet pretrained weights which requires the inputs to be divisible by 32. Thus, images are padded to 416×416. The training process spans 60 epochs, incorporating learning rate drops by a factor of 10 after the 45th and 55th epochs to enhance convergence and performance.

Conditional Random Fields (CRF): Conditional Random Fields (CRF) are widely employed in semantic segmen-

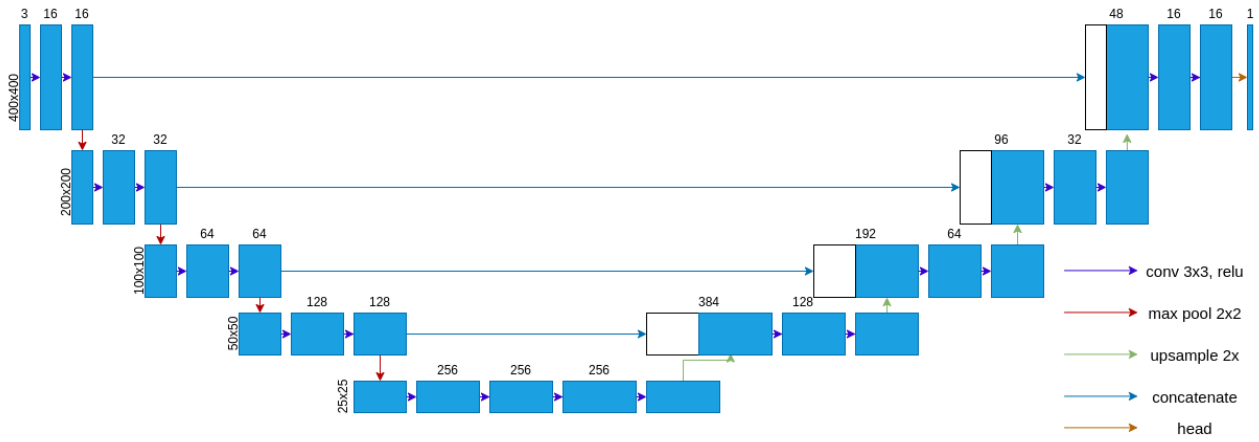

Figure 1. Structure of our baseline model, U-Net[5].

tation as a post-processing method that refines the predicted pixel-wise scores by modeling the relationship among the scores and local relationships of pixels[12], [6]. Although advanced approaches with more complex modeling of these local relationships have been proposed[13], [14], we opt to use the fully connected CRF[15] as we found its capacity to capture both fine-edge information and long-range dependencies would be especially beneficial to refine our predicted score maps.

Our main work on this part is, (1) to integrate the CRF into our pipeline as a post-processing step by utilizing an open-source implementation of fully connected CRF[16] and (2) to find the best-performing hyperparameters with grid search.

The main benefits of the use of CRFs have been known as (1) expanding the class region to fully cover the object of interest and (2) making the class boundaries detailed. However, these benefits are not quantitatively significant in our study. Thus, we would discuss how these known benefits become insignificant in our pipeline.

Inference: All the trained models are ensembled across different folds by simply averaging the per-pixel probabilities. This is achieved by passing the logits through the sigmoid function, resulting in a combined probability output. **Test-Time Augmentation** (TTA) is performed by predicting on the original and horizontally flipped images and averaging the per-pixel probabilities.

IV. Experiments & Results

A. Baseline Comparison

Despite its low computational cost and tractability, the Threshold Based Segmentation scheme we developed produced predictions similar to fully black predictions. These results are justified by the fact that averages of road pixels –which consist the initialization of the training variables– are computed per image; however the colour of roads differ among different images (due to lighting) or even inside the same image (due to the recency of the asphalt). This peculiarity of the data renders it very noisy for this method, thus leading to non-robust results. In addition, most of the label pixels are black, while we are using an F1 score to evaluate the performance of the model at each variables configuration. Consequently, the model ”spirals down” to favouring predicting black pixels.

Table II showcases that the gradual incorporation of techniques on our U-Net baseline leads to increased accuracy, as anticipated. This is an indication that, although the U-Net baseline is not sufficient to tackle the problem, it still consists a reliable foundation for developing a superior solution.

B. Conditional Random Fields (CRF)

To verify the effectiveness of the fully connected CRF as a post-processing of our pipeline, we both quantitatively and qualitatively compared the segmentation results of our model and ones refined with CRF.

In order to find the optimal combination of hyper-parameters.We conducted a grid search concerning to find the optimal combination of hyper-parameters. The grid is centered on the settings provided in the example code of the open source implementation [16] of the fully connected. The optimal setting was found in two ways: quantitative manner and qualitative manner.

For the quantitative evaluation, we considered the F1 score as our metric as it is widely used as an evaluation metric for semantic segmentation. The score is evaluated using 15 validation images. For the qualitative evaluation, we randomly chose 5 images among the validation images. We then proceeded to manually assign a rank score to every image with the different settings (i.e. grid cells of the grid search) in ascending order (best image gets a value of 1, second-best image gets a value of 2, and so on). Finally,

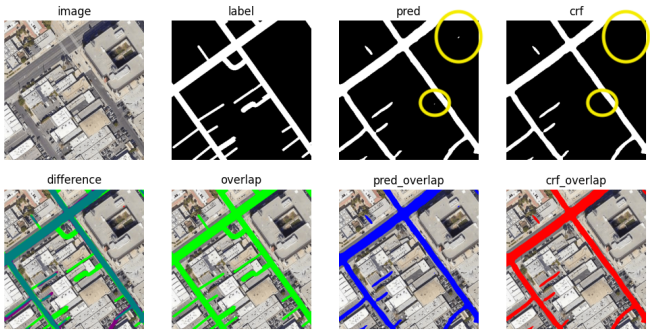

Figure 2. Qualitative comparison of CRF. The first column shows the original image and the image overlapped with predictions and labels. From the second column, the upper row shows the label, predictions, and the CRF refined predictions, and the bottom row shows an overlapping of that with the original image. The yellow circle shows where the CRF has benefits. The small mispredicted area at the center of the yellow circle on the prediction row is correctly predicted after CRF. However, CRF cannot refine more than tiny mispredictions, as explained in section IV.B.

Method	EfficientNet	Quantitative	Qualitative
F1 score	0.87353	**0.87550**	0.87383
Kaggle score	0.92675	0.92628	**0.92706**

Table I
THE PERFORMANCE WITHOUT AND WITH CRF

Method	Kaggle score
Threshold based Baseline	0.75451
U-Net Baseline	0.8952
U-Net EfficientNet backbone	0.92268
+ Pretraining	0.92675
+ Ensemble	0.93026
+ TTA	**0.9310**

Table II
RESULTS OF OUR PROPOSED SOLUTION.

Method	Kaggle score
No augmentations	0.91284
Only geometric augmentations	0.9237
All augmentations	**0.92675**

Table III
DATA AUGMENTATIONS ON U-NET WITH EFFICIENNET-B7 BACKBONE

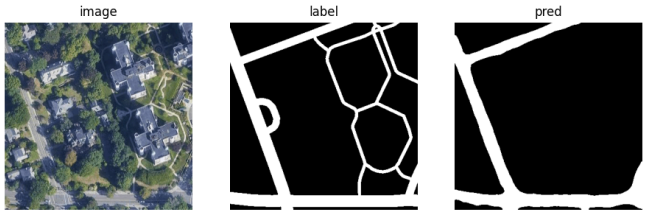

Figure 3. Failure case of the proposed pipeline. Our failure cases are mainly thin roads. Thin roads are not only easily confused by the surrounding objects but also easily inconsistently labeled.

we sum up the rankings of each setting along the different images. The setting with the smallest sum of the ranks is considered the best qualitative setting.

Table I shows the performance of our pipeline with and without CRF. The column "Qualitative" shows the fully connected CRF tuned with the abovementioned qualitative evaluation and the column "Quantitative" shows that with quantitative evaluation. Note that the F1 score is evaluated on our validation set.

The insignificance of the benefits resulting from the CRF is mainly because the roads are sometimes visually indistinguishable, especially where our model cannot correctly classify and thus the CRF is expected to refine. Also, the visually distinguished roads have clear visual boundaries with the surrounding objects and thus even the models can correctly separate them. Figure 2 shows this.

Note that the qualitatively found hyperparameter setting generalizes to the test set better, while the quantitatively found one generalizes worse than the sole model's. One possible explanation is the validation set's inability to perfectly reflect the true data distribution. And also it is possibly supporting that human-in-the-loop with a carefully-designed metric tuning can be better in terms of generalization. However, this should be more thoroughly investigated, and we thus leave this as our future work.

C. Ablation studies

As shown in Table II, the integration of the EfficientNet backbone into the U-Net architecture leads to a substantial performance boost. Furthermore, pretraining on the Massachusetts Roads Dataset results in an approximate increase of 0.5% on the public test set. Leveraging the power of ensembling and TTA further enhances the results, providing an additional improvement of 0.4%. Table III showcases the impact of augmentations on the overall performance. Interestingly, when augmentations are absent or limited to only geometric transformations (without blurring, changing brightness and contrast, etc.), the results are noticeably inferior compared to utilizing the full range of online augmentations. This highlights the crucial role that diverse and comprehensive augmentations play in achieving superior performance.

V. CONCLUSION

We started with a modest baseline and effectively enhanced it through a series of strategic steps. First, we incorporated an ImageNet pretrained EfficientNet-b7 into the architecture and used heavy augmentations to regularize it. Next, we conducted pretraining on an external dataset to address the limited dataset. We also conducted a widely-accepted postprocessing step, CRF, to remove small mispredictions (as roads) by expanding the non-roads area. To make our approach more robust we employed model ensembling and TTA. As a result, the proposed pipeline is shown to easily classify the thick and visually distinguishable roads; however, the thin, hidden, and/or visually indistinguishable ones cannot be properly classified (shown in Figure 3), which should be our future work.

REFERENCES

[1] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in sar satellite images with deep fully convolutional neural

networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 12, pp. 1867–1871, 2018.

[2] H. Chen, W. Guo, and J. Yan, "New method of sar image road recognition based on deep learning," *J. Jilin Univ*, vol. 50, pp. 1778–1787, 2020.

[3] V. John, K. Kidono, C. Guo, H. Tehrani, S. Mita, and K. Ishimaru, "Fast road scene segmentation using deep learning and scene-based models," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 3763–3768.

[4] H. Li, Y. Chen, Y. Yang, P. Liu, and C. Zhong, "Deep learning road extraction model based on similarity mapping relationship," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 9799–9802.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[7] P. Dube, B. Bhattacharjee, E. Petit-Bois, and M. Hill, "Improving transferability of deep neural networks," 2018.

[8] N. Ali, N. K. A. M. Rashid, and M. Y. Mohd, "Performance comparison between rgb and hsv color segmentations for road signs detection," *Applied Mechanics and Materials 393*, 2013. [Online]. Available: https://doi.org/10.4028/www.scientific.net/amm.393.550

[9] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," 2022.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[11] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.

[12] C. Rother, V. Kolmogorov, and A. Blake, "" grabcut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[13] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2800–2809.

[14] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1529–1537.

[15] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, 2011.

[16] A. Pedersen, G. Wang, M. Asad, and J. Jerphanion, "Simplecrf," https://github.com/HiLab-git/SimpleCRF/tree/master, 2021.