Blending Generative and Discriminative U-nets for Road Segmentation of Aerial Images

Bastian Morath[1], Daniel Peter[1], Michael Seeber[1], Kam-Ming Mark Tam[2], Group: Mandarinfish
[1]Department of Computer Science, ETH Zürich, Switzerland
[2]Department of Architecture, ETH Zürich, Switzerland

***Abstract*—Research in Machine Learning on image segmentation has devoted significant attention to the development of novel modelling architectures based on Convolutional Neural Networks (CNN). Rather than creating a new model architecture, this project aims to maximise the use of established approaches–synthesising their predictive power while exploring the application of additional learning techniques that may contribute to better prediction. Three techniques are presented: the blending of predictions from a model for different test-time augmentation of the inputs, predictions from a model obtained at various locally optimum states during the training process, or *Snapshot Ensemble*, and predictions from two different modelling approaches explored–encompassing a discriminative, and a generative formulation of a encoder-decoder CNN architecture featuring skip connections, informed by U-net and Pix2Pix. Preliminary results offer evidences that the blended approach may have contributed to improve generalisation, as reflected in the validation score provided by the submission system.**

1 Introduction

Image segmentation is the process of attributing every pixel in an image to a certain category, in order to create a simplified representation that is more meaningful to analyse. In the case of road segmentation, each pixel of an aerial RGB image is assigned a 'road' or 'background' label. In formulating our solution for the road segmentation task, we are motivated to develop a holistic approach that considers not only architectural variation, but also techniques covering different stages and aspects of the Machine Learning (ML) development process, in order to explore additional opportunities for improvement to prediction accuracy and generalisation. Significantly, the paper contributes and combines three strategies to 'blend' the predictions obtained from both individual, and group of models–encompassing both discriminative and generative approaches.

Overall, the following techniques were explored:

- ***Data Augmentation***: the effectiveness of various image augmentation operations were tested and assessed
- ***Model Architectures***: two interrelated models based on encoder and decoders with skip connections informed by U-net, which encompass both discriminative and generative approaches, were explored and evaluated.
- ***Blending Predictions***: the three strategies explored include the blending of predictions from a model for different test-time augmentation of the inputs(§3.4), predictions from a model obtained at various locally optimum states during the training process, or *Snapshot Ensemble*(§3.8), and predictions from the two modelling approaches explored (§3.9).
- Finally, the project implements techniques such as *Block Erasure* to promote better generalisation (§3.3), *Dilation Convolution* to support richer learning of spatial information (§3.7), and post-processing techniques based on *Thresholding* to provide additional image correction (§3.5)

2 Related Work

State-of-the art segmentation generally builds on *Fully Convolutional Neural Network* (FCN) with a two-part structure comprising of an encoder and a decoder (Fig. 1.1)

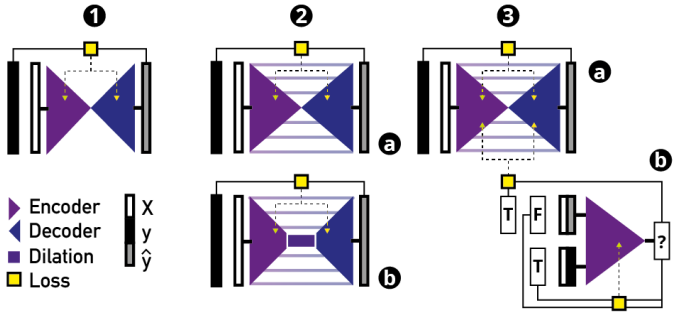


Fig. 1. Comparison on the architecture of the primary classes of models explored in this project

Necessary for translation invariance, the pooling layers cause detail loss, which is problematic for image segmentation whereby the exact alignment of the output segmentation and the input images are important. To address this limitation, researchers introduced skip connections to transfer information from the encoder's hidden layers to the decoder's (Fig. 1.2-a). Examples like *SegNet* [1], which transfers pooling indices, and *U-nets* [11], which transfers entire feature maps, have excelled in range of segmentation tasks. Additional innovations includes Yu et al. [14], which replaced pooling operations with a dilated convolution module that has the main benefit of maintaining high-resolution images even with deep networks (Fig. 1.2-b). The image-to-image translation task has also been explored through generative modelling, as explored by *Pix2Pix*, which is based on *Conditional Generative Adversarial Networks* (cGAN) [5] (Fig.1.3). cGANs have demonstrated impressive versatility, and are increasingly adapted in segmentation tasks [10, 3].

3 Methods

3.1 Dataset and Pre-Processing

The course provided two datasets for training and testing purposes, which respectively include 100 pairs of aerial images and their corresponding segmentation maps at a dimension of 400×400, and 94 aerial images at a dimension 608×608. Pixels labelled as road in the segmentation map have values set to 1.

3.2 Data Augmentation

Augmentation provided a mechanism for addressing a number of challenges presented by the dataset: these included (1) the small number of training samples, (2) the inconsistency in the dimension of the images between the training and submission test set, and (3) the high dimension of all images overall, which would be computationally expensive to model when used at their given dimensions.

To increase the diversity and quantity of training samples, which addresses challenge 1, a staged augmentation procedure

was implemented which consists of cropping, rotating, flipping and image adjustments (Fig. 2).

Notably, the use of randomised cropping provides an integrated remedy to challenges 2 and 3: the model can be trained on images from the training dataset that are cropped consistently at a dimension of 256×256, and be used to predict over patches that are similarly cropped from the testing dataset at the same dimension–patches that would then be stitched together to create the final submission output, as §3.4 details. The crop-and-patch approach takes advantage of the observation that both the training and testing dataset appear to share a similar scale, or map-to-ground ratio, so that the features contained in a 256×256 block learnt in the former will likely have application for the latter at the same scale. Significantly, challenge 2, i.e. the differences in dimension between the two datasets, can be addressed without the need of rescaling, which would have led to a reduction of quality and may limit the portability of the model, which is trained on the training dataset, for the testing dataset.

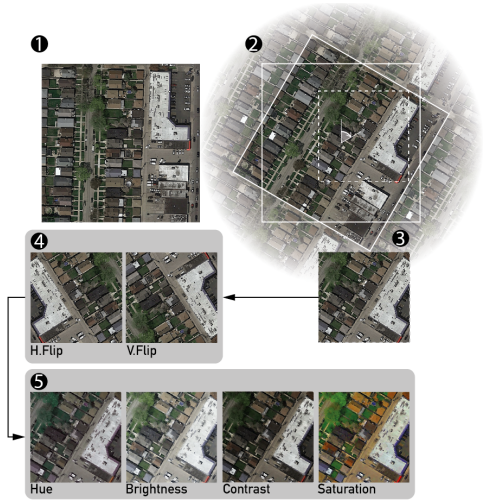


Fig. 2. Diagram visualising the data augmentation methods applied in the pre-processing step. (1) Once loaded, a pair of aerial and segmentation images are (2) randomly rotated at a multiple of 15 degrees. Regions outside the boundary edge of the image following rotation are filled by mirroring the image along the boundaries. (3) The rotated images are then cropped at random at a dimension of 256× 256. Finally, (4) the images may be randomly flipped and (4) randomly adjusted for their brightness, contrast, hue and saturation in a randomised order.

3.3 Gaussian Blocking

After the augmentation procedure, rectangular blocks of random dimension filled with Gaussian noise are inserted onto the aerial image. The method, called block erasure, acts as a regulariser that aims to promote better generalisation[15]. Visual inspection of the predicted output provided some suggestions that the technique may improve the model's ability in ensuring street continuity in the predicted segmentation map, as Fig. **??** shows.

3.4 Blending I: Patching Test-time Augmented Predictions

For each test image, the implemented models predict on 25 256×256 crops with equidistant spacing between them. The final prediction is then computed by the *stitching* of all 25 patches. For each pixel, the final value is computed by a weighted sum of the corresponding pixels in the patches, where all weights sum up to 1. This process is visualized in

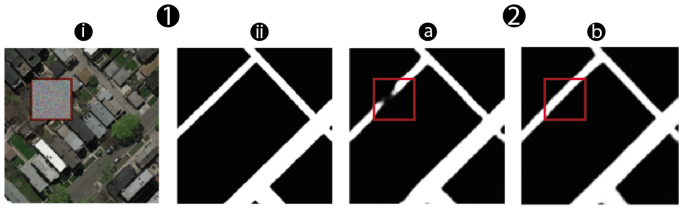


Fig. 3. By adding random gaussian blocks during training phase (1-i), a network might predict street continuities better (2b) even when the input image is blocked, in comparison to models trained without it (2a). (1-ii) shows the groundtruth.

(Fig. 4.1). The implemented weighting scheme, as shown in (Fig. 4.2), additionally decreases the weights of pixels near borders while increasing the weights in the middle–motivated by the observation that the U-net models tended to struggle with predictions near image borders.

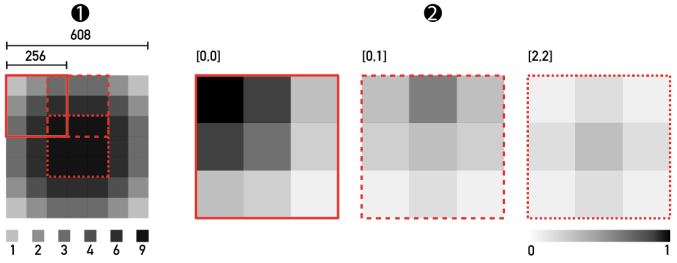


Fig. 4. Diagram visualising the number of patches a pixel is covered by (1) and how the pixels of a patch are weighted (2)

Fundamentally, the patching incorporates a technique used in computer vision called *test-time augmentation* (TTA) techniques, which has been shown to increase the robustness of a model predictions [13]. Informed by their success, this project similarly applies TTA to all 25 patches before they are stitched, so that each patch is the blended prediction obtained from 8 transformation of the input aerial images, as Fig. 5.1 shows.

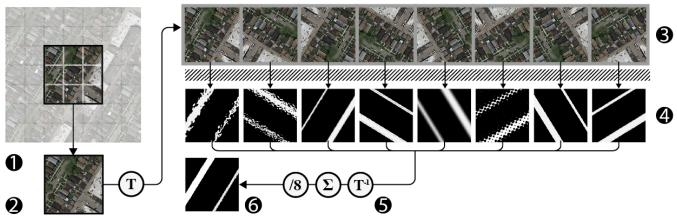


Fig. 5. Diagram visualising test time augmentation. Our implementation makes use of the eight symmetry transformations of the square. For each patch, seven additional transformations are applied (2-3) to create eight images per patch. The images are fed to the segmentation model–outputting 8 predictions (4). The corresponding inverse transformations are then applied to each image (5) to achieve reorientation. Finally, the eight predictions are averaged to create the final blended output (Fig. 5.6).

3.5 Post-processing predictions

In some studies, *thresholding* was used to convert the probability-based output of the segmentation model into the binary for submission. Both manual and automatic thresholding methods were tested, such as *Otsu's method* [9] and *Li's iterative Minimum Cross Entropy method* [7], and various morphological operators such as morphological closing, erosion and dilation [12] were tested to correct undesirable artefacts, or

incorrect classification, in the output, such as small opening in roads segments (Fig. A8).

3.6 Model Selection & Metrics

Both pixel-wise accuracy and *mean intersection-over-union* (MIoU) were measured throughout the training process to facilitate evaluation; however, owing to their poor differentiability, they are not used as training losses. The evaluation and final selection of the loss functions relied on both a initial benchmark study, as §4 describes, and observations during the prototyping and development process, as §4 documents.

3.7 Models

U-net. This project adopts a number of adjustment to the architecture of the U-net in it seminal formulation[11]. For the encoder, input images of size 256×256 are progressively encoded to feature maps with dimension of 16×16 in the bottleneck. Each encoder block consists of the structure of (Conv → BatchNorm → Dropout → Conv → BatchNorm → MaxPool). The number of filters used in each convolution layer are doubled with each down-sampling step, starting with 8 filters in the first encoder. Using transpose convolution, the otherwise similar decoder blocks then upsample the feature maps from the bottle neck to 256×256 outputs. Moreover, skip-connections feed the feature maps of every encoder block to the corresponding decoder block were they are concatenated with the upsampled feature maps. Using six dilated convolutions in the bottleneck helped to improve accuracy scores on the public test set.

Because class imbalance is unavoidable in road segmentation, whereby negative pixels often vastly outnumber positive pixels labelled as roads, this project extensively tested the *Balanced cross entropy*, *Soft dice loss* and the *Lovász-Hinge loss* [2], which measure the relative overlap between samples and are scale invariant and better suited for imbalanced datasets [8]. For our final models, $loss = 0.5 * loss_{BalancedCE} + 0.5 * loss_{softdice}$ was used.

The U-net models were trained with randomly initialised weights on the provided training data using the *Adam optimiser* [6] with a learning rate of 0.001 and batch size of 32. For every model, the number of training epochs was manually chosen to be large enough such that the model can train until there is no more notable improvement in the validation metrics.

Pix2Pix. As in the case for GAN-based models, the Pix2Pix multi-model architecture is composed of a *Generator* and a *Discriminator*. In this implementation, the Generator is based on a modified U-Net. Each block in the encoder follows the structure of (Conv → BatchNorm → Leaky ReLU) and each block in the decoder is (Transposed Conv → BatchNorm → Dropout(applied to the first 3 blocks) → ReLU). Additionally there are UNet fashioned skip connections between the encoder and decoder. Because the final layer has a $\tanh$ activation function the data was normalized into the range $[-1, 1]$. Compared to the original implementation the number of filters were halved to reduce both computational intensity, and complexity, which can promote better generalisation. The generator loss is composed of two components that are combined via a parameter λ to $loss_{generator} = loss_{gan} + \lambda * loss_{L1}$. While the $loss_{gan}$ is the sigmoid cross entropy of generated images and array of ones, the $loss_{L1}$ is the mean absolute error between the generated image and the target, which helps to receive structurally similar images. Experimental evidence suggests $\lambda = 200$ works well for our problem set. The discriminator is an adapted *PatchGAN* [5], where each block is (Conv → BatchNorm → Leaky ReLU). Similar to the generator we halved the the filter sizes, such that the discriminator doesn't overpower the generator. For both the real and fake loss the discriminator uses the sigmoid cross entropy, which are then summed up for the total loss $loss_{discriminator} = loss_{real} + loss_{fake}$. For the optimization of the generator as well as the discriminator Adam was used. The learning rate was set to $2e{-}6$ with $beta1 = 0.5$. Because the generator has to solve a harder problem than the discriminator, it was constantly overpowered by the discriminator during our initial experiments. To solve this issue, we apply the gradient updates to the weights only every 4th training step for the discriminator.

The final Pix2Pix baseline was trained for 600 epochs with a batch size of 1.

Flat Baselines. Apart from a baseline that predicts all pixels as background, we further studied the distribution of RGB values of roads in the training images and define a range of RGB values that in most cases correspond to a road. For each image a threshold method is used to mask all pixels that have an RGB value within this predefined range as road, and all other pixels as background. The original training data without augmentation was used.

3.8 Blending II: Snapshot Ensemble

Since the training process of neural networks are not deterministic, and the optimisation of neural networks is highly non-linear, multiple local optimal minima are typically reached within a single training run. However, even though the models may have similar validation score at their instances of local minimia, the predictions they offer on the validation images will generally be different. For instance the model may be better at predicting for certain features in one instance, but differ in another. This was noticeably the case when the predictions on the images from the testing dataset were compared. The observation motivated the use of Snapshot Ensemble method to increase robustness and accuracy. Snapshot Ensemble [4] uses cyclic learning rate schedules, such as cosine annealing, when training a neural network to find multiple local minima. The models corresponding to the multiple local minima can then be used in ensemble. In the simplified approach adopted in this project, the predictions of the model at their multiple optimal states are directly recorded, and averaged to obtain a final prediction.

3.9 Blending III: Blending Models' Predictions

Just as the quality of prediction may differ for model with identical architecture both within, and between multiple runs of training for a given level of performance according to their validation score, our experiments also appear to suggest that the U-net and Pix2Pix models exhibit different strengths and weaknesses. For instance, Fig. 6 shows that the U-net is better at discrening the road layout within parking lots whereas Pix2Pix tended to fail in this area in our experiments. By contrast, the Pix2Pix exceled with general street layout. In order to leverage the strength of both of them, the predictions of both models were blended by averaging their pixel-wise predictions. Evidence that the blended approach have contributed to better gernalisation can be found in Table Table I, whererby the blended model

scored 2% higher than either single model, when used in isolation, achieved.

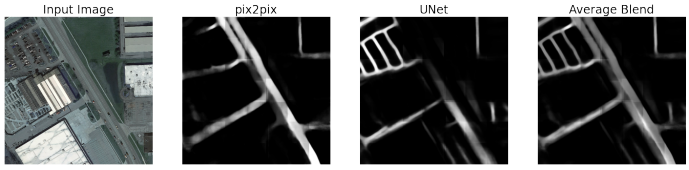


Fig. 6. Diagram visualising the predictions of the separate and the combined models.

4 Results

The final results of our models are presented in Table I.

In general, we achieved high accuracy with our models when examining the score achieved for the training dataset: the models can correctly identify roads in most cases, even when they are covered by trees or other obstructions. When an image is made up entirely of pavements however, accuracy is rather low: it may be that the lack of anchor objects, such as trees or cars, can reduce the ability of the model to correctly delineate the road layout (Fig. 4-222).

Furthermore, the results show that there is not a clear correlation between the public scores achieved by the baseline models and their corresponding validation scores that supported their training. This suggests that the ability for the model to reconstruct images accurately for the training dataset, as measured by validation score, may not be an effective indicator on the model's ability to generalise for the testing dataset, and is therefore insufficient as the lone driver of models' training or selection. Indeed, the best public score obtained in this project blended results from models that did not necessarily obtain the highest validation accuracy, namely Pix2Pix (Table 1). This provides support for the team's inclusion of a generative approach, which may have had greater successes, compared to the discriminative models explored, in learning a more structured notion of the segmentation task–one that did not necessarily translate to a higher accuracy over the training dataset, but nonetheless complemented the discriminative approach to achieve better generalisation for the testing dataset.

5 Discussion

Although the results are encouraging, it is clear that there is still room for further improvement.

6 Conclusion

The presented work provided a promising demonstration on the value in developing ML applications with a comprehensive 'ensembling' approach. Three techniques were investigated in particular: the blending of results within a model for different transformation of the inputs, the results from the optimal weights obtained from multiple runs of training, and the results from different modelling approaches that may complement one another. Finally, our experience has demonstrated the importance of developing a holistic ML approach whereby both investigations on model architecture, and techniques and extension for the setup may complement each other and contribute to better predictive capacity.

Fig. 7. Overlaying the patched prediction on the satellite images.

		Public Score	Validation Score
Baselines			
	All background	0.84743	-
	RGB thresholding	0.77142	-
	U-net	0.83978	-
Extensions			
	U-net with cropping	0.89113	0.9392
	U-net with augmentation	0.90092	0.9756
	U-net with cropping and dilation	0.88228	0.9344
	U-net with augmentation and dilation	0.91159	**0.9832**
	Pix2Pix with augmentation	0.89922	0.9642
Ensembling			
	Pix2Pix with snapshot	0.90053	-
	U-net with snapshot	0.90402	-
	U-net blended with Pix2Pix	**0.92313**	-

TABLE I
PUBLIC KAGGLE SCORES AND VALIDATION ACCURACY OF EXPLORED MODELS. BOLD MARKS OUR BEST RESULT.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[2] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. *The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks*. 2017. arXiv: 1705.08790 [cs.CV].

[3] Dragos Costea et al. "Creating roadmaps in aerial images with generative adversarial networks and smoothing-based optimization". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 2100–2109.

[4] Gao Huang et al. *Snapshot Ensembles: Train 1, get M for free*. 2017. arXiv: 1704.00109.

[5] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[6] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. arXiv: 1412.6980.

[7] CH Li and Peter Kwong-Shun Tam. "An iterative algorithm for minimum cross entropy thresholding". In: *Pattern recognition letters* 19.8 (1998), pp. 771–776.

[8] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571.

[9] Nobuyuki Otsu. "A threshold selection method from gray-level histograms". In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66.

[10] Mina Rezaei et al. "A conditional adversarial network for semantic segmentation of brain tumor". In: *International MICCAI Brainlesion Workshop*. Springer. 2017, pp. 241–252.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[12] Jean Serra. *Image analysis and mathematical morphology*. Academic Press, Inc., 1983.

[13] Connor Shorten and Taghi M. Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. In: *Journal of Big Data* 6.1 (2019), p. 60.

[14] Fisher Yu and Vladlen Koltun. “Multi-scale context aggregation by dilated convolutions”. In: *arXiv preprint arXiv:1511.07122* (2015).

[15] Zhun Zhong et al. “Random Erasing Data Augmentation.” In: *AAAI*. 2020, pp. 13001–13008.

7 Appendix

Fig. 8. From left to right: The satellite image, the prediction with our model, the binary images after using a manual threshold of 0.3 and Li’s method, respectively, and the resulting image after applying morphological operators to it.

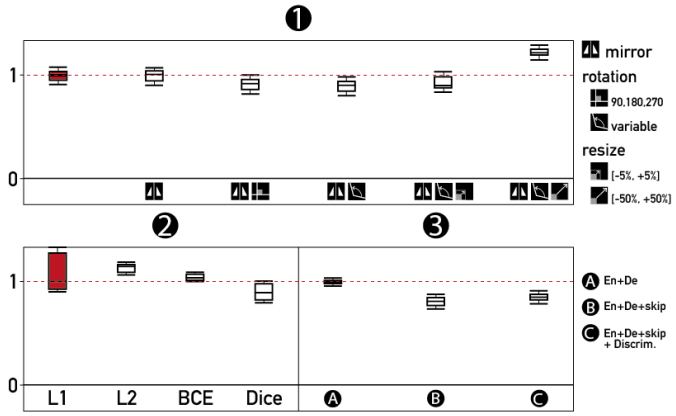


Fig. 9. Box plot showing results

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

Title of work

Blending Generative and Discriminative U-Nets for Road Segmentation of Aerial Images

Authored by

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

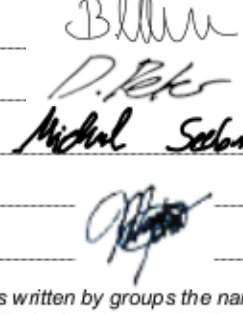
Name(s):	First name(s):
Morath	Bastian
Peter	Daniel
Seeber	Michael
Tam	Kam-Ming Mark

- have committed none of the forms of plagiarism described in the '_____' information
-
-
-

Place, date

Zürich, 31. July 2020

Signature(s)



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