# Parking Space Detection in the City of Granada

Luis Crespo Orti
e.luiscrespoorti@go.ugr.es

Isabel María Moreno Cuadrado
e.isa5456@go.ugr.es

Pablo Olivares Martínez
e.pablolivares@go.ugr.es

Ximo Sanz Tornero
e.ximo@go.ugr.es

## Abstract

*This paper addresses the challenge of parking space detection in urban areas, focusing on the city of Granada. Utilizing aerial imagery, we develop and apply semantic segmentation techniques to accurately identify parked cars, moving cars, and roads. A significant aspect of our research is the creation of a proprietary dataset specific to Granada, which is instrumental in training our neural network model. We employ Fully Convolutional Networks, Pyramid Networks and Dilated Convolutions, demonstrating their effectiveness in urban semantic segmentation. Our approach involves comparative analysis and optimization of various models, including Dynamic U-Net, PSPNet and DeepLabV3+, tailored for the segmentation of aerial images. The study includes a thorough experimentation phase, using datasets such as UDD5 and UAVid, alongside our custom Granada dataset. We evaluate our models using metrics like Foreground Accuracy, Dice Coefficient and Jaccard Index. Our results indicate that DeepLabV3+ offers the most promising performance. We conclude with future directions, emphasizing the need for a dedicated neural network for parked car detection and the potential for application in other urban environments. This work contributes to the fields of urban planning and traffic management, providing insights into efficient utilization of parking spaces through advanced image processing techniques.*

## 1. Introduction

Currently, there exists a problem related to the parking spaces in Granada. There is not enough availability leading to conflicts between drivers or spending too much time in order to park. Our project aims to correctly segment satellital images of the city. The objective is to identify the locations of parked cars, moving cars, and roads within a given image of Granada. This will be done through different approaches using a neural network model. Given a set of satellite images taken at a certain frequency, it would be possible to extract information about parking areas and their availability over time, providing drivers with tools to park efficiently in the city. For all the code, results and documentation reffer to the project on GitHub[1].

## 2. Background

*Fully Convolutional Networks* (FCNs) [25] represented a significant advancement in the field of deep learning for tasks that require understanding of spatial hierarchies, such as semantic segmentation. The architecture of FCNs typically consist of two parts: the downsampling (encoder) path and the upsampling (decoder) path. In these architectures, the encoder gradually reduces the spatial dimensions of the input image while increasing the depth to extract and learn features at multiple levels of abstraction. This process typically involves a series of convolutional and pooling layers. The decoder, on the other hand, gradually reconstructs the target output from the encoded features, often through a process known as up-sampling or transposed convolution. This architecture is essential for semantic segmentation tasks as it allows for detailed pixel-level predictions while retaining contextual information from the entire image.

*Pyramid networks* were first introduced with the Feature Pyramid Network (FPN) [23] and were designed to solve the problem of capturing objects at various scales. In semantic segmentation and object detection, objects can vary significantly in size, making their detection difficult. Pyramid networks tackle this by implementing a multi-scale approach. They create feature pyramids that maintain information at multiple resolutions, allowing the network to recognize objects and features across different scales effectively.

*Dilated convolutions* [41], also known as atrous convolutions, provide a solution to the problem of resolution loss in standard convolutional neural networks (CNNs). In semantic segmentation, maintaining high-resolution feature

[1] https://github.com/pab1s/granada-parking-segmentation

maps is crucial for accurate pixel-level classification. Dilated convolutions enable networks to expand the receptive field of filters without losing resolution or coverage. By adjusting the dilation rate, these networks can aggregate multi-scale contextual information without compromising the resolution.

## 3. Related Works

Deep learning advancements have notably propelled computer vision, particularly with models like VGG [34] and ResNet [16] exhibiting unparalleled feature extraction capabilities. These models often serve as the backbone for semantic segmentation models such as SegNet [2], Mask-RCNN [15], PSPNet [43], and DeepLab [4–6]. The primary challenge in deploying these models is their reliance on extensive, high-quality data [8, 14]. To mitigate this, strategies like synthetic data generation via generative adversarial networks [29], transfer learning [1], and data augmentation [13] have been effective in achieving high-quality results with smaller datasets, as outlined in [37].

The U-Net model [30] is particularly notable for its proficiency with limited datasets. This Fully Convolutional Network (FCN), originally developed for medical image segmentation, is based on an encoder-decoder structure. Its architecture, including innovations such as nested, dense skip pathways in U-Net++ [44] and adaptations for road extraction from satellite imagery [42], demonstrates its adaptability to various segmentation tasks.

In semantic segmentation for remote sensing, the evolution of analytical approaches has been significant. Initial methods focused on image descriptors [11, 12] and texture filters [33]. The field has since progressed to incorporating various machine learning techniques, as seen in the works of Verdie et al. [39] and Tokarczyk et al. [38]. Presently, the forefront of aerial image semantic segmentation leverages Convolutional Neural Networks (CNNs) and Transformers. FCNs, for instance, have demonstrated state-of-the-art results in segmenting low-quality aerial images from online maps [21, 27]. Recent developments like the Uncertainty Aware Network (UANet) by He et al. [17] have outperformed existing models like Buildformer [3], highlighting the ongoing advancements in this field.

## 4. Methods

In the execution of our project, we adopted a methodical and incremental approach.

Our initial objective was to identify the most effective model among three selected candidates for the semantic segmentation of aerial imagery. The preliminary phase involved training these models using a fusion of two datasets: UDD5 and UAVid, comprising images from China. This amalgamation was strategically chosen to enhance the models' understanding of aerial images, aiming for the network to capture the general characteristics of our problem.

In the second stage, we curated a dataset comprising segmented aerial images of Granada. Leveraging the pre-trained weights derived from the final execution of the top-performing model identified in the initial stage, we conducted retraining with our Granada dataset for several epochs, meticulously refining the model weights. This meticulous process yielded promising outcomes.

To achieve the primary goal of segmenting parked cars from the predicted images, we explored two distinct approaches. The first involved implementing a third stage of post-processing of data, employing a comprehensive heuristic computer vision algorithm method to detect parked cars depicted in Algorithm 1. In summary, the network "Unparked Car Model" processes parked and moving cars as a single class, and it is the heuristic algorithm that later discerns between them.

**Algorithm 1** Parked Car Detection

```
Define color codes for car, background, and road
for each image in the dataset do
    Create a mask for the car pixels
    Find the contours of the car pixels
    for each contour found do
        Create a mask for the current contour
        Draw the contour on the mask
        Create a dilation kernel
        Dilate the mask with the kernel
        Count the number of background and road pixels within the dilated mask
        if background pixels count is greater than road pixels count then
            Change color of the car pixels within the original mask
        end if
    end for
end for
```

The algorithm employs a dilation operation with a specified 15x15 kernel to enhance car contours, facilitating the connection of disjointed regions. This process effectively smoothens and extends the spatial coverage of the contours, contributing to a broader interpretation of the spatial characteristics.

Alternatively, the second approach to detect parked cars focused on treating parked and non-parked cars as two distinct classes. This modification aimed to empower the network to discern, in its output, which cars are parked and which are not, introducing a new class in the process. This model "Parked Car Model" autonomously learn this distinction.

# 5. Experiments

## 5.1. Datasets.

In the training and evaluation of our baseline model, we utilized UDD5 [7] and UAVid [26] datasets. Both datasets consist of images sharing similar characteristics and coherence. Furthermore, we have curated the images by retaining only the classes pertinent to our research, namely roads and cars, while categorizing the remaining elements as 'background'. Therefore, we will have three classes: 'background', 'road', and 'car'.

The project's primary goal was to accurately segment images from Granada. Consequently, it was necessary to apply specific transformations to our data to ensure generalization to this new dataset. Challenges arose from the suboptimal quality of the photos, characterized by low resolution and prevalent shadows, which adversely affected our final model's performance. For further discussion and examples about this topic see Appendix A. To mitigate this, a shadow transformation was integrated into the dataloader, resulting in improved model validation with the China images. In addition, we have incorporated various default transformations such as flipping, abrupt rotations, zooming, and adjustments in lighting.

In the final approach, the final model was trained for a few epochs starting from the weights of the baseline model, using a custom dataset named "GranadaAerial"[2], which consists of 90 labeled images for semantic segmentation of parked cars, moving cars, and roads from the city of Granada, Spain. This dataset includes 10 images designated for validation and another 10 for testing purposes. The images from Granada used for creating the dataset were obtained from [20] according to their policies for educational purposes and segmented using CVAT [31] and Adobe Photoshop by ourselves.

Regarding dataset partitioning, a manual approach was employed instead of a random split. Given the limited number of segmented images available, we deemed it fair to perform the partitioning manually, thereby ensuring a diverse selection of images in all sets (training, validation, and test) as in [9].

Constrained by a limited dataset, we employed data augmentation techniques over the train set, predominantly leveraging flipping, abrupt rotations, and lighting adjustments. Additionally, we introduced an aditional technique enabling image zooming up to 80% to ensure model consistency with scale among training and validation samples, addressing slight variations in the scales at which the images were captured in the dataset.

[2]The dataset can be found at https://drive.google.com/drive/folders/1rEgcZT_jyJ1zQ88i4epYUKs1QTDDiYya?usp=drive_link.

## 5.2. Baseline Models.

In our study, we selected three baseline models—Dynamic U-Net, PSPNet, and DeepLabV3+—each featuring a ResNet101 backbone. These models were chosen based on their architecture characteristics and proven performance, particularly on the Cityscapes dataset, which is closely related to our task of urban scene understanding through aerial imagery.

Dynamic U-Net is an adaptation of the U-Net architecture, known for its effectiveness in aerial image segmentation [19]. The 'dynamic' aspect of this model lies in its ability to adjust the architecture's encoder, making it highly adaptable substituting the traditional U-Net encoder by another backbone which can be more appropiate for the actual computer vision problem. This feature is particularly advantageous for our purpose, as it provides to such a powerful model as it is U-Net the capability to do transfer learning, a key aspect considering our computational resources.

PSPNet (Pyramid Scene Parsing Network) has demonstrated exceptional performance in scene parsing tasks, particularly evidenced by its results on the Cityscapes dataset. The model incorporates a pyramid pooling module that works at different scales, enabling it to capture global contextual information effectively. This ability is crucial for semantic segmentation in aerial imagery, where understanding the context is key to accurately classifying various urban elements.

DeepLabV3+, an advanced iteration in the DeepLab series, is renowned for its performance in semantic segmentation tasks, again proven on the Cityscapes dataset. This model introduces an improved atrous convolution strategy and includes an atrous spatial pyramid pooling (ASPP) module, which efficiently captures multi-scale contextual information.

The use of a ResNet101 backbone in the models provides deep feature extraction capabilities across multiple scales, enhancing the models' ability to discern intricate details essential for accurate semantic segmentation in complex urban landscapes.

## 5.3. Training Methodology.

The framework chosen to work with the models was Fastai [18]. For uniform training conditions, we employed focal loss. In scenarios where certain classes are underrepresented, standard cross-entropy loss may lead to the model being dominated by the majority class, resulting in suboptimal performance for the minority class. Focal loss mitigates this issue by down-weighting well-classified examples, allowing the model to focus more on difficult-to-classify instances [24]. In our first stages we trained PSPNet with cross-entropy obtaining worse results. We employ the Adam optimizer [22] for all models (Dynamic U-Net, PSPNet, DeepLabV3+), including their ResNet101 backbones

pretrained with ImageNet [10]. The chosen hyperparameters for Adam were $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The training duration was 50 epochs as there was no significant improvement observed across all three models, and the metric values stabilized around this point, guided by the one cycle policy for learning rate scheduling [36].

Adopting the one cycle policy as training strategy was based on substantial evidence from numerous studies attesting to its efficacy as widely seen in the literature. Despite its higher resource demands compared to fine-tuning or transfer learning, our hardware capabilities were sufficiently robust to support this approach. In alignment with this strategy, we utilized the learning rate finder method provided by Fastai to ascertain an optimal learning rate, thereby augmenting the overall efficacy of the training process

Each model was assigned a distinct image size, meticulously chosen to maximize the utilization of our available GPU VRAM memory and ensure an equitable comparison. Specifically, we endeavored to maintain a 3:2 image resize aspect ratio whenever feasible, in alignment with the aspect ratios of UDD5 and UAVid images (1.667 and 1.7, respectively). This approach was influenced by precedents set in other studies [40]. A batch size of 32 images was chosen, aligning with findings from various studies that suggest enhanced performance with larger batch sizes in this specific training strategy, as stated by Smith et [35].

After comprehensive evaluations, we ultimately chose DeepLabV3+ as our foundational model for the subsequent stage. Further experimentation was conducted with this selected model, and the final base model was trained over 60 epochs.

As mentioned earlier, we pursue two different approaches, leading to the training of two distinct models.

The first approach involves initializing our model with the weights obtained from the winning model and training it for twelve epochs using the proprietary dataset from Granada, using only three classes (road, car, and background), the model does not differentiate between a parked car and a car in motion. We executed a limited number of training cycles due to our initial use of pre-trained weights on a Chinese dataset, where the model exhibited satisfactory performance. However, it required adaptation to account for the distinct image scale and road characteristics present in Granada, which differ from those in China. And beyond this number of epochs, there is no substantial improvement in the metric values obtained trying to avoid overfitting.

We use the one-cycle policy. This strategy is designed to efficiently guide the training process, exploiting the cyclical learning rate schedule to enhance model generalization and convergence.

On the other hand, our second approach involves fine-tuning [32] the base model using the Granada dataset, distinguishing between moving cars and parked cars. This results in a total of four classes. Consequently, an additional layer is introduced into the base model to accommodate the four-class output

| Model | Valid loss | Foreground acc. | Dice coeff. | Jaccard coeff. |
|---|---|---|---|---|
| Dynamic U-Net | 0.0736 | 0.6954 | 0.7466 | 0.6269 |
| PSPNet | 0.0719 | 0.5964 | 0.7240 | 0.5994 |
| DeepLabV3+ | **0.05404** | **0.7726** | **0.7955** | **0.6836** |

Table 1. Performance evaluation of segmentation models at the 50th epoch.

In this case, we undergo a phase of freezing, encompassing the first 5 epochs, during which we focus on training only the model's head while keeping the underlying layers fixed. Subsequently, we will transition to the unfreezing phase, extending for the subsequent 10 epochs.

Furthermore, a batch size of 16 images was chosen in both cases due to, on the one hand, the considerations mentioned earlier about larger batch sizes, and on the other hand, because we did not have a very large dataset.

### 5.4. Metrics.

The effectiveness of semantic segmentation models in urban scene understanding, particularly from aerial imagery, is critically evaluated using specific metrics. In this study, we utilize three metrics: Foreground accuracy, Dice Coefficient, and Jaccard Index. Each of these metrics offers a unique perspective on model performance.

Foreground Accuracy is essential for emphasizing the model's performance in segmenting non-background classes. It can be obtained as follows

$$ACC_{foreground} = \frac{TP + TN}{TP + TN + FP + FN}$$

where $TP$ is the number of true positives, $TN$ the number of true negatives, $FP$ the number of false positives and $FN$ is the number of false negatives where ground truth background is not taken into account. This metric is particularly relevant in aerial urban imagery, where the focus is often on diverse urban elements rather than the background.

Dice Coefficient (DSC) measures the overlap between the predicted segmentation and the ground truth across multiple classes. It computes the Dice Coefficient for each class and averages these values, given as

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2TP}{2TP + FP + FN}$$

where $X$ is the number of segmented pixels, $Y$ is the number of pixels belonging to ground truth. $TP$, $FP$ and $FP$ are the same as for foreground accuracy. This metric is particularly useful for datasets with class imbalances, as it provides equal weight to each class.

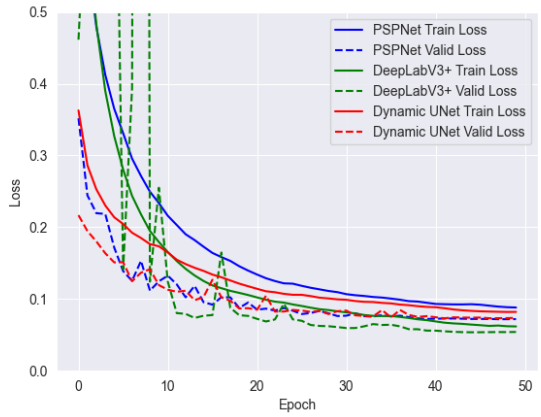


Figure 1. Training and validation loss comparison across epochs for Dynamic U-Net, PSPNet, and DeepLabV3+ models. The graph illustrates the trend of validation loss over 50 epochs, highlighting the stability of Dynamic U-Net and PSPNet and the occasional spikes in validation loss for DeepLabV3+, which recovers and maintains a leading performance in subsequent epochs.

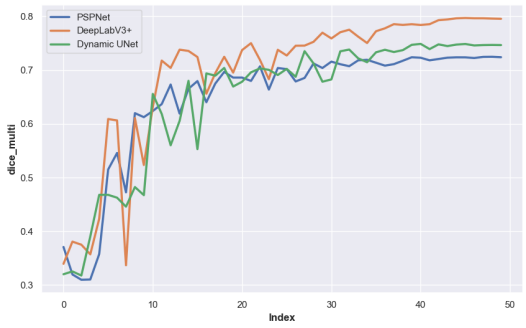


Figure 2. Comparison of Dice Coefficients for Image Segmentation during the first stage of the training.

Jaccard Index (JI) is an adaptation of the Intersection over Union (IoU) metric for multiclass scenarios. It calculates the IoU for each class and then averages these scores. It can be calculated from

$$JI = \frac{|X \cap Y|}{|X \cup Y|} = \frac{DSC}{2 - DSC}$$

where $X$ and $Y$ are the same as for the DSC. This metric is crucial as it is good handling class imbalance, well-suited for tasks where precise boundary detection and little sensitive to background.

### 5.5. Results

After training the three models, we observed a similar pattern of training and validation loss across them, as demonstrated in Figure 1. We extended the training duration despite the slow reduction in validation loss due to consistent improvements across performance metrics, which is particularly notable in the case of DeepLabV3+ as seen in Figure 2. As shown in Table 1, DeepLabV3+ achieved the best results among the models, with the lowest validation loss and the highest accuracy and similarity coefficients.

In our study, the validation curve generally stays below the training curve in the graphs. This observation prompts an investigation into the potential influence of the aggressive transformations applied during the data augmentation process on the model's performance.

An inspection of the validation loss graph for DeepLabV3+ reveals intermittent spikes, which we attribute to the model's interaction with more complex or diverse data samples within the batch, reflecting a temporary destabilization in the optimization process, probably due to the aggressive data augmentation accomplished during training. Another hypothesis we considered, as the records of the validation loss[3] show an isolated abrupt increment of this value in epochs 3 and 8, suggesting a possible overflow of the loss function during those epochs. Nevertheless, despite the feasibility of this hypothesis, we discarded this last option as the training loss behaved as expected. These aberrations underscore the necessity for a delicate balance in learning rate and suggest potential areas for refinement in hyperparameter optimization, despite using the proposed method by Fastai to find an appropiate learning rate for one cycle policy. The resilience of DeepLabV3+ is evident in its rapid recovery following these perturbations, ultimately leading to superior performance metrics.

We now present the outcomes of the two methodologies employed to tackle the Granada problem.

In the Figure 3 it is shown the train and valid loss of both final models during the second stage. We can assert that the validation loss of the UnParked Car model is reduced to the Parked Car model due to its lower number of classes, resulting in decreased classification complexity.

Furthermore, we present two visualizations of the performance of a particular image from the test for both models in Figure 4. The first image displays the mask predicted by the first approach, while the following image exhibits the heuristic algorithm for parked car detection applied to that mask. Similarly the third referees to the second approach. In this instance, as previously mentioned, the model directly predicts areas corresponding to parked cars in the image.

In both visualizations, white represents true positives, red denotes false negatives, and green signifies false positives. Green indicates areas present in the ground truth mask but not predicted, while red indicates areas predicted but not present in the original mask. All evaluations exclude

[3]The records of the metrics can be found in the results folder of the project.

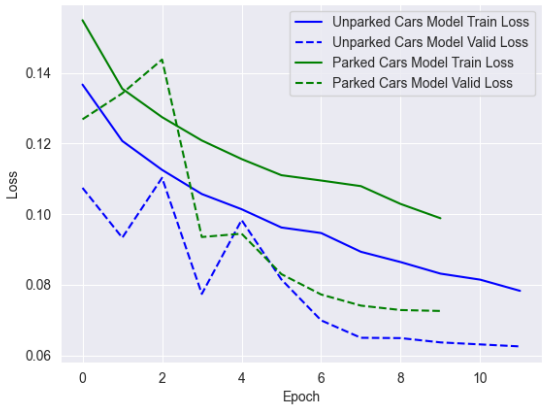


Figure 3. The figure displays the training and validation loss for both the Parked Car Model and the UnParked Car Model.

the background.This method was influenced by precedents set in other studies [28].

In this concrete example, each of the two approaches accurately represents the parked cars, as evidenced.

In general, both approaches yield favorable results. This reflects that in specific situations with small datasets, the heuristic approximation can be a viable alternative to a solution entirely based on neural networks.

## 6. Conclusions

In this paper, our primary objective was the detection of parked cars in Granada through semantic segmentation.We pursued this goal by exploring two distinct approaches, both of which have yielded favorable results.

A particular mention should be made regarding the carefully assembled custom dataset of Granada for this study due to the scarcity of datasets specifically designed for locales like Granada within the current research context.

In future work, leveraging daily aerial images of the city of Granada through a suitable tool could enable the use of our model for generating statistics on parking areas. This application extends beyond mere detection, allowing for the analysis of parking behaviors, routines, and citizen patterns. Such insights could prove valuable for urban planning and the optimization of parking infrastructure.

To further enhance the capabilities of our model, one avenue for exploration is the expansion of the dataset. By incorporating additional diverse images, including various weather conditions, different times of the day, and seasonal variations, we can improve the model's robustness and generalization. This expanded dataset would capture a broader range of parking scenarios, making the model more adept at handling real-world variations.

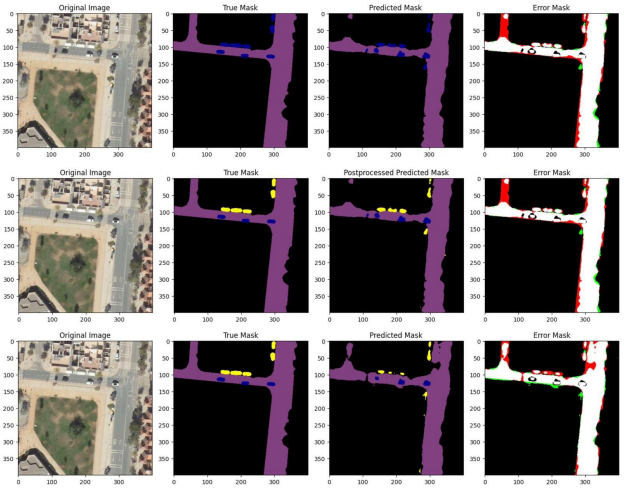


Figure 4. Analysis of Parked Car Detection Methodologies in Aerial Imagery. The analysis showcases three distinct approaches across rows: the baseline detection of cars and roads (Row 1), the implementation of a heuristic algorithm for parked car identification (Row 2), and a model specifically trained to detect parked cars (Row 3). For each method, the columns display the original image, the ground truth segmentation, the algorithm's predicted segmentation, and the error mask, respectively. The error mask uses red to signify missed detections (false negatives), green for incorrect detections (false positives), and white for correct detections (true positives).

## References

[1] Nouman Ahmed, Sudipan Saha, Muhammad Shahzad, Muhammad Moazam Fraz, and Xiao Xiang Zhu. Progressive unsupervised deep transfer learning for forest mapping in satellite image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 752–761, 2021. 2

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2

[3] Keyan Chen, Zhengxia Zou, and Zhenwei Shi. Building extraction from remote sensing images with sparse token transformers. *Remote. Sens.*, 13:4441, 2021. 2

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In

*Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[7] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 347–359. Springer, 2018. 3

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[11] Martin A. Fischler, Jay M. Tenenbaum, and Helen C. Wolf. Detection of roads and linear structures in low-resolution aerial imagery using a multisource knowledge integration technique. *Computer Graphics and Image Processing*, 15:201–223, 1981. 2

[12] Pascal Fua. Using generic geometric models for intelligent shape extraction. In *American Association for Artificial Intelligence*, number CONF, 1987. 2

[13] MAA Ghaffar, A McKinstry, T Maul, and TT Vu. Data augmentation approaches for satellite image super-resolution. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:47–54, 2019. 2

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[17] Wei He, Jiepan Li, Weinan Cao, Liangpei Zhang, and Hongyan Zhang. Building extraction from remote sensing images via an uncertainty-aware network. *arXiv preprint arXiv:2307.12309*, 2023. 2

[18] Jeremy Howard et al. fastai. https://github.com/fastai/fastai, 2018. 3

[19] Roberto Huerta, Fabiola Yépez, Diego Lozano-García, Víctor Hugo Cobián, Adrian Ferriño, Hector De Leon-Gomez, Ricardo González, and Adriana Vargas-Martínez. Mapping urban green spaces at the metropolitan level using very high resolution satellite imagery and deep learning techniques for semantic segmentation. *Remote Sensing*, 13, 05 2021. 3

[20] Instituto Geográfico Nacional (IGN) Spain. PNOA Image of Granada, Spain, 2022. PNOA Image - IGN Spain. 3

[21] Pascal Kaiser, Jan Dirk Wegner, Aurélien Lucchi, Martin Jaggi, Thomas Hofmann, and Konrad Schindler. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6054–6068, 2017. 2

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 3

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 3

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[26] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108 – 119, 2020. 3

[27] Dimitrios Marmanis, Jan D Wegner, Silvano Galliani, Konrad Schindler, Mihai Datcu, and Uwe Stilla. Semantic segmentation of aerial images with an ensemble of cnns. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:473–480, 2016. 2

[28] Pablo Mesejo, Andrea Valsecchi, Linda Marrakchi-Kacem, Stefano Cagnoni, and Sergio Damas. Biomedical image segmentation using geometric deformable models and metaheuristics. *Computerized Medical Imaging and Graphics*, 43:167–178, 2015. 6

[29] Caijun Ren, Xiangyu Wang, Jian Gao, Xiren Zhou, and Huanhuan Chen. Unsupervised change detection in satellite images with generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10047–10061, 2020. 2

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[31] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong,

Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, Aug. 2020. 3

[32] Zhiqiang Shen, Zechun Liu, Jie Qin, Marios Savvides, and Kwang-Ting Cheng. Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9594–9602, May 2021. 4

[33] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 1–15. Springer, 2006. 2

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[35] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018. 4

[36] Leslie N. Smith and Nicholay Topin. Super-convergence: very fast training of neural networks using large learning rates. In *Defense + Commercial Sensing*, 2018. 4

[37] Maofeng Tang, Konstantinos Georgiou, Hairong Qi, Cody Champion, and Marc Bosch. Semantic segmentation in aerial imagery using multi-level contrastive learning with local consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3798–3807, 2023. 2

[38] Piotr Tokarczyk, Jan Dirk Wegner, Stefan Walk, and Konrad Schindler. Features, color spaces, and boosting: New insights on semantic classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(1):280–295, 2014. 2

[39] Yannick Verdie and Florent Lafarge. Detecting parametric objects in large scenes by monte carlo sampling. *International Journal of Computer Vision*, 106:57 – 75, 2013. 2

[40] Hongming Xu and Tae Hyun Hwang. Automatic skin lesion segmentation using deep fully convolutional networks. *arXiv preprint arXiv:1807.06466*, 2018. 4

[41] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1

[42] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 2

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[44] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2

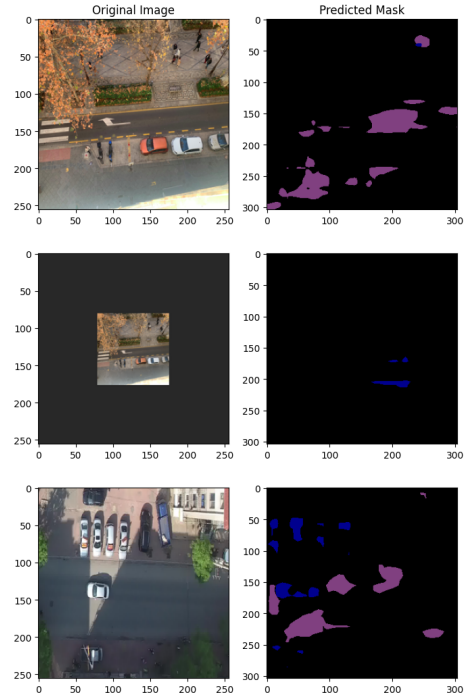


Figure 5. Performance evaluation of our model on images from Granada, exhibiting challenges in identifying cars and roads (left) and the corresponding predicted masks (right). Background is represented with black color, cars in blue color and roads in lilac color. The model's difficulty in detecting roads under shadows and identifying cars and roads at hte same scale suggests a domain shift and limited dataset diversity in the training phase.

## A. Preliminary Model Evaluation Visuals

In the initial phase of our research, we aimed to evaluate the performance of our model on images from Granada, shortly after the initial training phase. This decision stemmed from an awareness that our training datasets comprised primarily of aerial imagery sourced from countries markedly different from our own. This discrepancy is evident in Figure 5, where the model exhibits notable challenges in simultaneously identifying cars and roads within the same frame. Additionally, it struggles to accurately detect roads obscured by shadows. We hypothesize that these difficulties arise from a "domain shift" in the original dataset used for training. Furthermore, the dataset's limited diversity, characterized by wide roads and uniform perspectives, scales and weather conditions, likely contributed to the model's inadequate generalization capabilities.