Deep learning residential building segmentation for evaluation of suburban areas development

Agnieszka Łysak[2][0000−0002−1439−6355] and Marcin Luckner[1][0000−0001−7015−2956]

[1] Jan Kochanowski University, Institute of Physics, ul. Uniwersytecka 7, 25-406 Kielce, Poland agnieszka.lysak@ujk.edu.pl
[2] Warsaw University of Technology, Faculty of Mathematics and Information Science, ul. Koszykowa 75, 00-662 Warsaw, Poland marcin.luckner@pw.edu.pl

Abstract. Deep neural network models are commonly used in computer vision problems, e.g., image segmentation. Convolutional neural networks have been state-of-the-art methods in image processing, but new architectures, such as Transformer-based approaches, have started outperforming previous techniques in many applications. However, those techniques are still not commonly used in urban analyses, mostly performed manually. This paper presents a framework for the residential building semantic segmentation architecture as a tool for automatic urban phenomena monitoring. The method could improve urban decision-making processes with automatic city analysis, which is predisposed to be faster and even more accurate than those made by human researchers. The study compares the application of new deep network architectures with state-of-the-art solutions. The analysed problem is urban functional zone segmentation for the urban sprawl evaluation using targeted land cover map construction. The proposed method monitors the expansion of the city, which, uncontrolled, can cause adverse effects. The method was tested on photos from three residential districts. The first district has been manually segmented by functional zones and used for model training and evaluation. The other two districts have been used for automated segmentation by models' inference to test the robustness of the methodology. The test resulted in 98.2% accuracy.

Keywords: Transformers, SegFormer, Deep learning, Semantic segmentation, Computer vision

1 Introduction

Suburbanisation, the process through which urban areas expand and evolve, is often seen as a positive outcome of population growth, rising incomes, advances in transportation technology, and the decentralisation of jobs. However, a byproduct of this process, known as urban sprawl, poses several challenges [6,22]. Urban sprawl, a term encompassing various phenomena such as the expansion of urban boundaries, land use practices, and their consequences, leads to issues like spatial mismatch between housing and employment zones, over-reliance on automobiles, fragmented local governance, and inefficient spatial planning [13]. These

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

(a) *Wietrznia*, Kielce (b) *Pod Telegrafem*, Kielce (c) *Marymont*, Warsaw

Fig. 1: Analysed residential zones in Kielce and Warsaw

issues impact various dimensions, including economic (higher infrastructure and public service cost, increasing unemployment), policy (unplanned growth, uncoordinated development), environmental (loss of vegetation, increased pollution) and social (car dependency, poverty, reduction in social interaction) [4].

Traditional methods of urban analysis, often manual and time-consuming, are increasingly supplemented by advanced technologies. There are some attempts, including urban sprawl monitoring using artificial neural networks [25] and convolutional neural networks [16,30]. There are also approaches based on SegFormer architecture, described in Section 3.1, for the problem of semantic segmentation of roads for sustainable mobility development [23], or buildings [12,21].

Our work examines the possibility of the application of deep-learning approaches to the problem of urban sprawl monitoring. We compared four deep learning architectures (with a particular interest in SegFormer architecture [28]) and analysed their generalisation ability. In this context, we examined the City of Kielce, experiencing significant urban sprawl [20]. The occurrence is growing [13] and is particularly evident in suburban residential districts like *Wietrznia* area (Fig. 1a) and *Pod Telegrafem* area (Fig. 1b). In contrast, the City of Warsaw, particularly the densely developed *Marymont* district (Fig. 1c), presents a different scenario, with lower levels of urban sprawl [20].

Our research and obtained results show that the Segformed architecture, trained initially for scene segmentation [34], can be applied – using manually labelled aerial photos and AdamW optimising algorithm [15] – for urban area segmentation. The proposed model achieved over 91 per cent accuracy on the testing data, which was not obtained by several other architectures.

After segmentation, a vector representation of the functional area was created, which can be converted to a segmentation map, substantively the same as the land use map, within the selected functional zones. This map can be used for further analysis, which is necessary in many fields, such as urban, economic, and spatial planning. If this methodology were standard, there would be no problem calculating such a map on the fly when new data is derived and monitoring the growth of the residential area with streamed new data.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

The rest of this paper is structured as follows. Section 2 summarises related works in the segmentation area and deep-learning techniques. Section 3 presents the workflow of the proposed solution and the deep-learning model – the SegFormer – used in the tests. Section 4 presents the obtained results on three different data sets and compares our framework with other state-of-the-art methods. Section 5 concludes the paper and presents possible further research.

2 Related works

A typical urban zone segmentation approach is image data for vector conversion, often with the addition of socioeconomic data and then pixel-wise classification with the use of algorithms like SVM [7], K-means [32], and XGBoost [2]. Also, convolutional neural networks (CNN) are introduced as models, which operate on pure image data only and extract features from them as a part of the learning process, like AlexNet and ResNet [9], custom CNN [35], DFCNN [1], SegNet [24], DeepLab family [27], and attention-based systems [17]. Though semantic segmentation of functional zones is complex and requires understanding crucial image features and capturing context dependencies, the transformer-based methodology seems the best intuition for the problem.

Vision Transformers (ViT) started from a concept of context dependencies extended on the whole image [8]. A disadvantage of the ViT architecture is the impossibility of solving more complex computer vision tasks, like detection or segmentation. The first attempt to overcome this drawback was Pyramid Vision Transformer (PVT) [26], a Transformer backbone for convolution block replacement, not a complete architecture. Thanks to the pyramid feature generation, it can work on image classification, detection and segmentation.

Transformer-based architectures have been invented for computer vision tasks with different purposes like Swin Transformer [14] to improve ViT, networks for image detection [5], or networks for image segmentation [5,33], especially SegFormer [28], the model chosen for this work.

Several works discussed urban segmentation – using deep learning – before. Pan et al. proposed using the U-net deep learning architecture to detect unplanned urban settlements [18]. The proposed method led to an overall accuracy of over 86% for the building segmentation, which can probably be improved using newer network architecture.

Zhang et al. proposed a deep learning-based framework called RFCNet [31]. The solution did not work on aerial photos but fused multiple views and generated plausible and complete structures. The solution could be used in urban sprawl observation because the authors presented that their solution can construct roof structures from photos

Finally, Yi et al. created UAVformer, a composite transformer network for urban scene segmentation of unmanned aerial vehicle (UAV) images [29]. The system works on various kinds of photos and performs more complex scenery segmentation than in our work. The obtained building recognition varies from 88.5% to 95.2% according to a data set. However, because the types of photos

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

and the set issues are slightly different in our case, it is hard to compare the results directly.

3 Methods

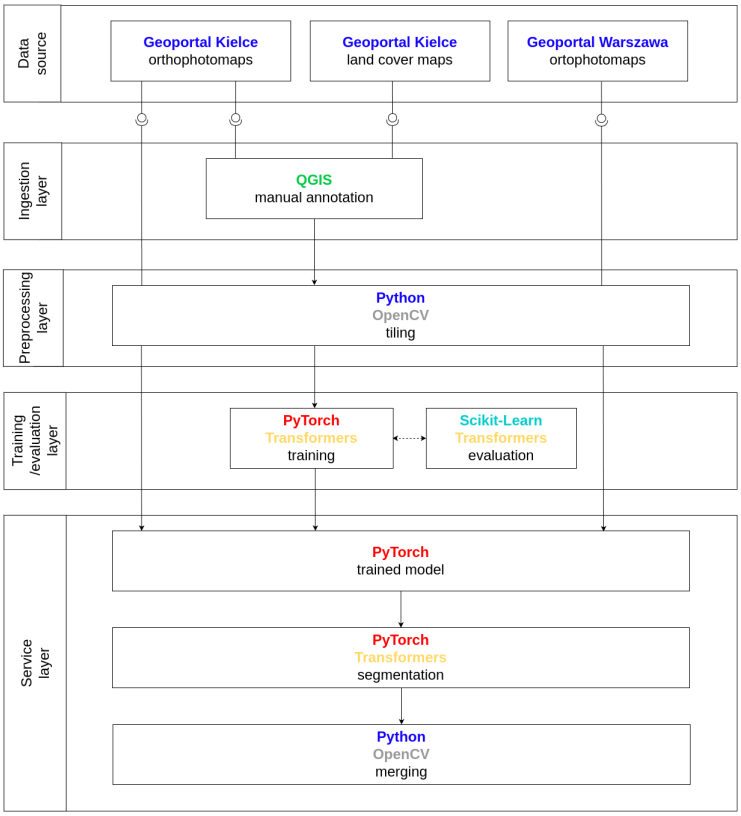


Fig. 2: Schematic workflow chart

In our work, we used maps from Kielce Geoportal (aerial photos and land cover maps) from 2019 and Warsaw Geoportal (aerial photos) from 2022. Photos from Kielce were used in the learning stage. Due to shifts and inconsistencies in combination with the photos, the land cover map has been rejected as a ground truth-functional zones segmentation map. Hence, data was labelled manually through geographic information system (GIS) software based on information about residential zone locations from the land cover maps. Then, the dataset was tiled and split into train and test parts.

Workflow (Fig. 2) takes a tiled image and produces a segmented image to classify a residential area. Although the dataset is limited to only three districts

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

from two Polish cities, it was selected based on availability and relevance to the specific urban areas. The intent was to focus on these regions to develop and validate our deep-learning framework under controlled and consistent conditions. This decision was also influenced by the constraints in data accessibility and the computational resources required for processing more extensive datasets.

Construction of the dataset Data for SegFormer training and evaluation was taken from Kielce's open-access maps. The photo was from April 2019, and it was constructed from a 5×5 centimetres from a plane flying 700 meters over ground height. It was downloaded from GIS Kielce open-access Geoportal via Web Map Service (WMS). The orthophoto was downloaded as a 1:500 scale map, converted to 1200 dpi Portable Graphics Format (PNG). The map was labelled in QGIS open-source software based on residential zone localisation information from the land cover map obtained from Kielce Geoportal for the same scope as the photo. Substantively, there were two classes: buildings and ground (building surroundings). Later, the image from the photo was converted to an RGB JPG image, and the annotated image with labels was converted to an 8-bit GRAY JPG image. Next, both images were tiled to 512×512 pixels tiles. The class information was coded according to the approach applied in the ADE20K semantic segmentation dataset, which contains over 20K images annotated with pixel-level objects [34]. Because we were implementing SegFormer architecture from the Huggingface Transformers library, where class info is inherited from the ADE20K dataset by default, we followed this behaviour and used the ADE20K *building* class for houses and *grass* class for building surroundings.

Dataset preprocessing Training $29,113 \times 15,938$ pixels image from Kielce *Wietrznia* residential district was tiled via OpenCV in Python programming language. The final training dataset contained three-channel (RGB) 512×512 tiles. The supporting segmentation masks were delivered as one channel (GRAY) 512×512 tiles. We removed empty tiles and tiles with 80% background and no building class existence. So, the training dataset finally counted 812 images with corresponding annotations.

Testing stages SegFormer was evaluated three times, presented in Fig 3. During the training, we checked the models' performance on one house image from *Wietrznia* district, which was from the same location as the training set but was not a part of the training data. The model has not seen this exact house. However, it was extracted from *Wietrznia* district, where the rest of the area was used for the models' training (Fig. 3a). The image was 1024×1024 RGB JPG, which was tiled into four 512×512 tiles.

After the experiment was done, the model was tested on the second *Pod Telegrafem* dataset (Fig. 3b), which was a district 4 kilometres away from the localisation of the training dataset and, of course, it also has not been seen by the model during the learning phase. Photo was also RBG JPG image, with $3072 \times$

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

(a) *Wietrznia* (b) *Pod Telegrafem* (c) *Marymont*

Fig. 3: Testing datasets from residential districts. The area contoured in blue was the training data for SegFormer. The yellow area (at the lower-left corner) was the local part of the testing data. The areas contoured in magenta were the remote parts of the testing data.

2560 resolution, divided into 30 512×512 tiles. The last test was performed on the residential district *Marymont* from Warsaw (Fig. 3c). Data for the SegFormer test was the Warsaw open-access map. The photo was from April and May 2022, and it was constructed from a 5×5 centimetres vertical photo from a plane flying at 1600 meters over ground height, taken with a camera for vertical photographs. It was downloaded from GIS Warszawa open-access Geoportal via WMS. The photo was 2048×2048 RGB JPG, divided into 16 512×512 tiles.

To sum up, the testing stage included the same area but a different house, from training data, data from the same city with similar building density and architecture (mostly detached houses), and data from a different city, from more densely built regions and a little more condensed architecture (more semi-detached and terrace houses).

3.1 Deep learning model

In the proposed framework, we used SegFormer b5 [28], pre-trained on the ADE20K dataset. We experimented with different versions of SegFormer, different epochs, and learning rates. The best combination was on the b5 version, after 28 training epochs and a 0.0006 learning rate.

SegFormer architecture The applied model combines Transformer and Multi-Layer Perceptron (MLP) architectures [26]. It consists of the encoder part for feature extraction and the decoder part for upsampling and segmentation mask prediction (Fig. 4).

The encoder's input image is divided into patches of 4×4 pixels. In contrast to ViT, smaller patches work better in detailed classification, like semantic or

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

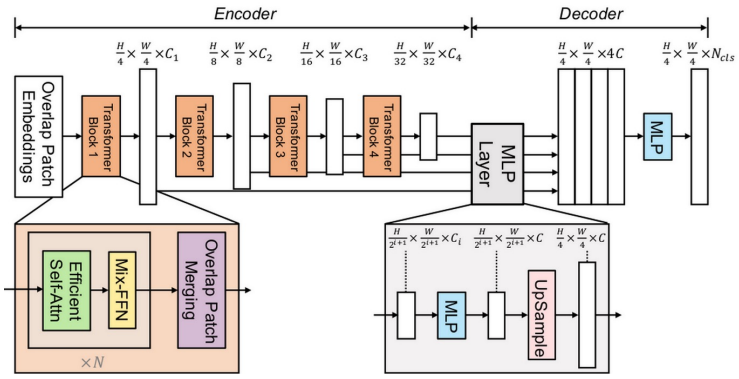


Fig. 4: SegFormer architecture [26]

instance segmentation. The transformer block – an equivalent to the convolution block in CNN – performs feature extraction. This Transformer block works without a positional encoding module. In semantic segmentation, the input image should be arbitrarily shaped. Therefore, classical rigid positional encoding is not possible to deploy. Interpolation for positional encoding in SegFormer is replaced by Mix-Feed Forward Network (Mix-FFN) with local positional information share.

The encoder consists of Efficient Self-Attention, Mix-FFN and Overlap Patch Merging. Self-Attention is being computed as standard, but with efficiency improvement, thanks to reducing the density of one of the attention mechanism formula components. Mix-FFN consists of a convolutional layer and Multi Layer Perceptron (MLP) for data-driven, flexible positional encoding, which is not fixed, like in a typical Transformer. Overlap Patch Merging enables feature size reduction. In the decoder, which is very lightweight, there is MLP, which takes features from the encoder and fuzzes them together to unify channel dimensions. Then, features are upsampled and concatenated together. Second, MLP fuses concatenated features and predicts a segmentation map.

Training The SegFormer model, by default, employs the cross-entropy loss function for optimisation. Our implementation utilised the PyTorch and Huggingface libraries, enabling us to choose from Huggingface-supported PyTorch optimisers. We opted for the AdamW optimising algorithm [15]. Our choice to utilise the AdamW optimiser was driven by its distinct advantages in enhancing model generalisation. A key feature of AdamW, an extension of the Adam optimizer [11], is its implementation of weight decay regularisation. This approach is particularly effective in minimising the loss function by selectively adjusting smaller weight values. AdamW minimises loss function by finding small weight values, which helps to overfit less and generalise better due to eliminating irrelevant components and suppressing static noise on the target.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

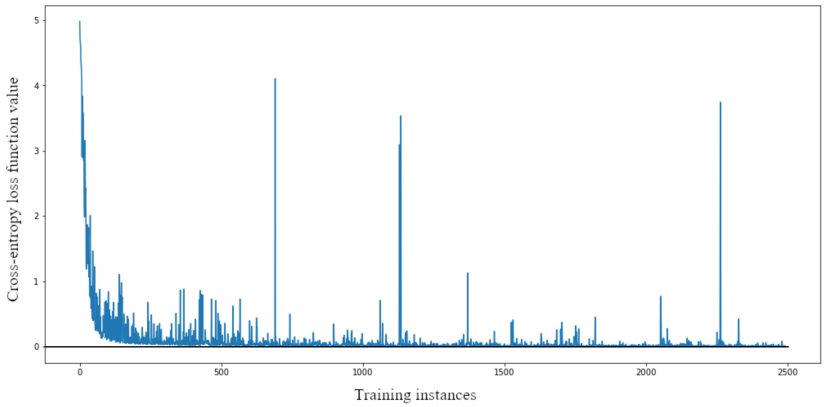


Fig. 5: Cross-entropy loss function during the learning phase plot

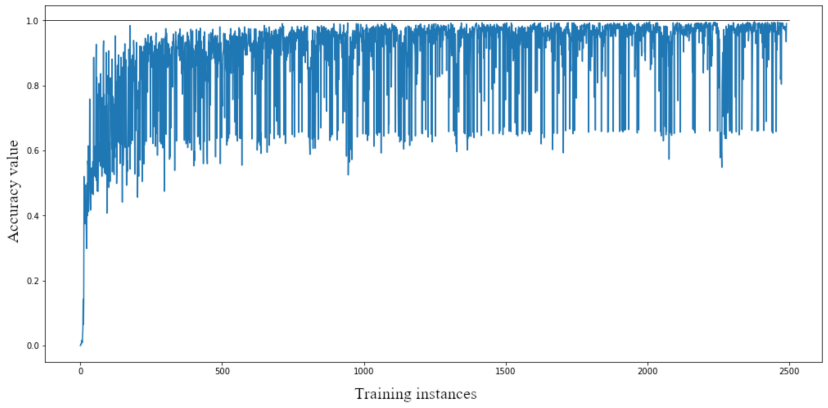
Our choice of a learning rate of 0.0006 was based on empirical testing, yielding the most favourable results regarding model performance. During training, images were batch-processed, each batch containing two images of dimensions 512×512 pixels in RGB colour mode. The optimisation process minimises the loss function by comparing the cross-entropy between the target and predicted pixel classes in logit mode.

Fig. 5 presents the loss function. The function decreases quickly and stabilises at a low level, which is a proper and expected behaviour. Sporadic peaks in its value could testify to outlying image tiles, in which models' predictions were inaccurate.

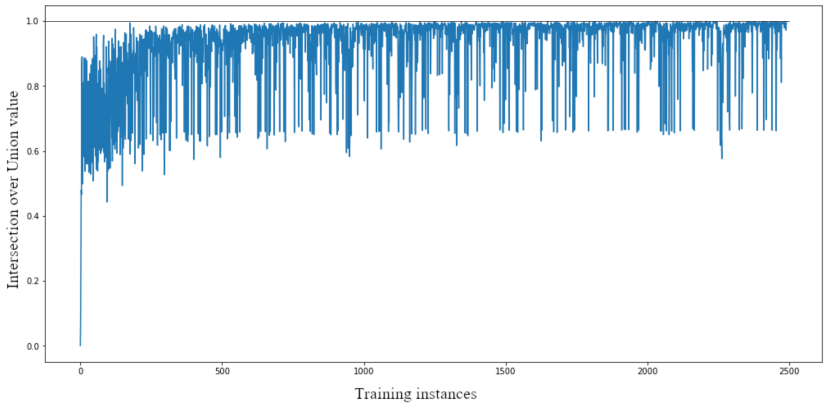
The model's parameters, which facilitated this minimisation, were adjusted following the selected learning rate. The training was concluded once the model performance metrics reached satisfactory levels, after which the model's parameters were saved for future inference purposes. Notably, an automatic hyperparameter tuning stage was not incorporated into this research.

Evaluation A two-step evaluation process assessed pixel-wise accuracy and the mean intersection over union metrics. Pixel-wise accuracy represents the percentage of pixels correctly classified concerning the target mask. A second metric – mean Intersection over Union (mIoU) – was introduced because accuracy could be misleading in cases where many pixels belong to the ground and fewer to residential buildings (imbalanced classes). mIoU calculates the overlap between the predicted and target masks, divided by the combined area of both masks. Fig. 6 presents metrics in the training stage using *Wietrznia* dataset. The trained model achieved an average pixel-wise accuracy of 0.93 (Fig. 6a) and 0.86 mIoU (Fig. 6b).

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9



(a) Accuracy



(b) IoU

Fig. 6: Metric plots for the learning phase

4 Results

4.1 SegFormer tests on various data sets

The model was evaluated on data marked with yellow and magenta in Fig. 3. Fig. 7 presents the qualitative results for these data sets.

In the testing phase, the model's inferencing capability was initially tested on a single house from the *Wietrznia* dataset. The model's proficiency in processing images of varying dimensions was evaluated using a 2048×2048 pixels map tile. In this test, the model successfully segmented the building area, achieving a 96.2 % mIoU and 99.6 % accuracy. The qualitative results (Fig. 7a) were highly encouraging regarding qualitative performance.

Subsequently, we extended our evaluation to include two additional datasets: *Pod Telegrafem* and *Marymont*. For these tests, the input tiles maintained the

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

(a) *Wietrznia*

(b) *Pod Telegrafem*

(c) *Marymont*

Fig. 7: SegFormer results

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

exact dimensions as those used in the training phase (512 × 512 pixels). The SegFormer model demonstrated robust segmentation capabilities in urban areas, achieving an 86.7% mIoU and 98.2% accuracy for *Pod Telegrafem* district (the results are shown in Fig. 7b) and 50.3% mIoU and 91.4% accuracy for *Marymont* district (the results are shown in Fig. 7c). These results validate the model's effectiveness in diverse urban environments.

4.2 SegFormer vs. other methods

The SegFormer results were compared against several pre-trained state-of-the-art algorithms to evaluate their effectiveness. That includes the vision transformers-based solution DPT [19] published in 2021, the mask transformer-based method Mask2Former [3] published in 2022, and the latest version of the acclaimed real-time object detection and image segmentation model YOLOV8 [10] published in 2023.

All algorithms were subjected to the same testing procedure to ensure a fair and consistent comparison. The *Wietrznia* dataset served as the basis for this evaluation. Each algorithm was applied to segment a single building within this district and then extended the testing to include the *Pod Telegrafem* and *Marymont* districts. The outcomes of these comparative tests have been presented in Tab. 1.

Table 1: Comparison of the results of our framework and state-of-the-art methods for image semantic segmentation

Algorithm	Dataset	IoU [%]	Accuracy [%]
DPT	*Wietrznia*	84.0	97.9
Mask2Former	*Wietrznia*	76.2	96.5
SegFormer	*Wietrznia*	96.1	99.5
YOLOV8	*Wietrznia*	96.5	99.6
DPT	*Pod Telegrafem*	20.1	87.0
Mask2Former	*Pod Telegrafem*	38.7	82.2
SegFormer	*Pod Telegrafem*	86.7	98.2
YOLOV8	*Pod Telegrafem*	34.9	87.4
DPT	*Marymont*	37.3	71.6
Mask2Former	*Marymont*	56.6	87.4
SegFormer	*Marymont*	50.3	91.4
YOLOV8	*Marymont*	62.2	91.6

The proposed SegFormer-based approach outperformed most of the other methods in the tests. However, it is notable that the YOLOV8 algorithm achieved a higher mIoU on *Wietrznia* and *Marymont* test sets. Additionally, YOLOV8 demonstrated slightly better accuracy in these datasets.

It is worth noticing that any reference method could not obtain the IoU close to the proposed framework on a separate *Pod Telegrafem* data set. This

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

finding underscores the impressive generalisation capabilities of the SegFormer framework, particularly in diverse urban settings.

5 Conclusions

Our study demonstrates the effective application of the SegFormer model, which was fine-tuned rather than pre-trained, for semantic segmentation in residential zones. This approach yielded promising results, both qualitatively and quantitatively.

Comparison with the state-of-the-art methods, in their pre-trained versions, showed that in many cases, SegFormer and the workflow proposed in this research achieved the best results on the test sets from *Wietrznia*. In the second and third testing districts, SegFormer architecture performed very well, being outperformed by YOLOV8 in two testing sets.

Notably, the SegFormer model exhibited strong generalisation capabilities, particularly in the additional tests using data from another city. Despite a decrease in performance compared to the original city (Kielce), which was anticipated due to the model not being trained on data from other cities, the results were still robust and satisfactory. The model had not been trained on data from any other city; therefore, urban aesthetics from another region can cause the outcome to deteriorate. A solution for problems like that is additional training of the actual model on data from other metropolises of interest.

The main contribution is the novel application of the known SegFormer architecture to urban sprawl monitoring, a relatively underexplored area. Also, analysis of the custom dataset from Kielce and Warsaw provides new insights specific to these urban areas.

We acknowledge certain limitations in our research, primarily related to the dataset's size and diversity. These limitations could affect the model's generalizability across different urban settings. To address this, we suggest expanding the dataset to include various photos from diverse areas. Such an approach would likely enhance the model's ability to account for variations in urban characteristics, including different residential area locations and roofing materials, thereby bolstering the overall robustness and generality of the method.

Likewise, more functional zone types and data from a more comprehensive time range could be included to broaden the methodology. Also, exploring the integration of historical data to track and predict urban sprawl over time presents a promising direction. Additionally, fine-tuning the SegFormer configuration's hyperparameters could yield improved results for specific applications, such as photo segmentation. Finally, augmenting our dataset is another strategy that could further refine our outcomes.

In summary, while our study demonstrates the potential of using advanced deep learning models like SegFormer in urban sprawl monitoring, it also opens up several avenues for further exploration and improvement. The insights gained from this research contribute to deep learning, computer vision, urban planning, and sustainable development. The practical implications of this research

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

for urban planners and policymakers are significant. The ability to accurately and efficiently monitor urban sprawl can inform more sustainable urban development practices, help resource allocation and support environmental conservation efforts.

References

1. Bao, H., Ming, D., Guo, Y., Zhang, K., Zhou, K., Du, S.: DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data. Remote Sensing **12**(7), 1088 (mar 2020). https://doi.org/10.3390/rs12071088, https://doi.org/10.3390%2Frs12071088
2. Chen, S., Zhang, H., Yang, H.: Urban functional zone recognition integrating multisource geographic data. Remote Sensing **13**(23) (2021). https://doi.org/10.3390/rs13234732
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2022-June**, 1280–1289 (2022). https://doi.org/10.1109/CVPR52688.2022.00135
4. Chiguvi, D., Kgathi-Thite, D.: Analysis of The Positive and Negative Effects of Urban Sprawl and Dwelling Transformation in Urban Cities: Case Study of Tati Siding Village in Botswana. Journal of Legal, Ethical and Regulatory Issues **25**(S2), 1–13 (2022)
5. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the Design of Spatial Attention in Vision Transformers. Advances in Neural Information Processing Systems **12**(NeurIPS), 9355–9366 (2021)
6. Cocheci, R.M., Petrisor, A.I.: Assessing the Negative Effects of Suburbanization: The Urban Sprawl Restrictiveness Index in Romania's Metropolitan Areas. Land **12**(5) (2023). https://doi.org/10.3390/land12050966
7. Deng, Y., He, R.: Refined Urban Functional Zone Mapping by Integrating Open-Source Data. ISPRS International Journal of Geo-Information **11**(8) (2022). https://doi.org/10.3390/ijgi11080421
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
9. Izzo, S., Prezioso, E., Giampaolo, F., Mele, V., Di Somma, V., Mei, G.: Classification of urban functional zones through deep learning. Neural Computing and Applications **34**(9), 6973–6990 (2022). https://doi.org/10.1007/s00521-021-06822-w, https://doi.org/10.1007/s00521-021-06822-w
10. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (Jan 2023), https://github.com/ultralytics/ultralytics
11. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings pp. 1–15 (2015)
12. Li, M., Rui, J., Yang, S., Liu, Z., Ren, L., Ma, L., Li, Q., Su, X., Zuo, X.: Method of Building Detection in Optical Remote Sensing Images Based on SegFormer. Sensors **23**(3) (2023). https://doi.org/10.3390/s23031258

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

13. Lityński, P.: The intensity of urban sprawl in Poland. ISPRS International Journal of Geo-Information **10**(2) (2021). https://doi.org/10.3390/ijgi10020095
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE International Conference on Computer Vision pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. 7th International Conference on Learning Representations, ICLR 2019 (2019)
16. Mansour, D., Souiah, S.A., El Amin Larabi, M.: Built-up area extraction through deep learning. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 6805–6808 (2021). https://doi.org/10.1109/IGARSS47720.2021.9554694
17. Niu, R., Sun, X., Tian, Y., Diao, W., Chen, K., Fu, K.: Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–18 (2022). https://doi.org/10.1109/TGRS.2021.3065112
18. Pan, Z., Xu, J., Guo, Y., Hu, Y., Wang, G.: Deep learning segmentation and classification for urban village using a worldview satellite image based on U-net. Remote Sensing **12**(10), 1–17 (2020). https://doi.org/10.3390/rs12101574
19. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision Transformers for Dense Prediction. Proceedings of the IEEE International Conference on Computer Vision pp. 12159–12168 (2021). https://doi.org/10.1109/ICCV48922.2021.01196
20. Renata, R.C., Barbara, C., Andrzej, S.: Which polish cities sprawl the most. Land **10**(12) (2021). https://doi.org/10.3390/land10121291
21. Song, J., Zhu, A.X., Zhu, Y.: Transformer-Based Semantic Segmentation for Extraction of Building Footprints from Very-High-Resolution Images. Sensors **23**(11) (2023). https://doi.org/10.3390/s23115166
22. Spirkova, D., Adamuscin, A., Golej, J., Panik, M.: Negative effects of urban sprawl. In: Charytonowicz, J. (ed.) Advances in Human Factors in Architecture, Sustainable Urban Planning and Infrastructure. pp. 222–228. Springer International Publishing, Cham (2020)
23. Tao, J., Chen, Z., Sun, Z., Guo, H., Leng, B., Yu, Z., Wang, Y., He, Z., Lei, X., Yang, J.: Seg-Road: A Segmentation Network for Road Extraction Based on Transformer and CNN with Connectivity Structures. Remote Sensing **15**(6) (2023). https://doi.org/10.3390/rs15061602
24. Tian, T., Chu, Z., Hu, Q., Ma, L.: Class-wise fully convolutional network for semantic segmentation of remote sensing images. Remote Sensing **13**(16), 200–215 (2021). https://doi.org/10.3390/rs13163211
25. Tsagkis, P., Bakogiannis, E., Nikitas, A.: Analysing urban growth using machine learning and open data: An artificial neural network modelled case study of five Greek cities. Sustainable Cities and Society **89**, 104337 (2023). https://doi.org/10.1016/j.scs.2022.104337
26. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. Proceedings of the IEEE International Conference on Computer Vision pp. 548–558 (2021). https://doi.org/10.1109/ICCV48922.2021.00061
27. Wang, Y., Gao, L., Hong, D., Sha, J., Liu, L., Zhang, B., Rong, X., Zhang, Y.: Mask DeepLab: End-to-end image segmentation for change detection in high-resolution remote sensing images. International Journal of Applied Earth Observation and Geoinformation **104**, 102582 (2021). https://doi.org/10.1016/j.jag.2021.102582, `https://doi.org/10.1016/j.jag.2021.102582`

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9

28. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. Advances in Neural Information Processing Systems **15**(NeurIPS), 12077–12090 (2021)
29. Yi, S., Liu, X., Li, J., Chen, L.: UAVformer: A Composite Transformer Network for Urban Scene Segmentation of UAV Images. Pattern Recognition **133** (2023). https://doi.org/10.1016/j.patcog.2022.109019
30. Yin, B., Guan, D., Zhang, Y., Xiao, H., Cheng, L., Cao, J., Su, X.: How to accurately extract large-scale urban land? Establishment of an improved fully convolutional neural network model. Frontiers of Earth Science **16**(4) (2022). https://doi.org/10.1007/s11707-022-0985-2
31. Zhang, X., Aliaga, D.: RFCNet: Enhancing urban segmentation using regularization, fusion, and completion. Computer Vision and Image Understanding **220**(April), 103435 (2022). https://doi.org/10.1016/j.cviu.2022.103435, `https://doi.org/10.1016/j.cviu.2022.103435`
32. Zhang, X., Li, W., Zhang, F., Liu, R., Du, Z.: Identifying urban functional zones using public bicycle rental records and point-of-interest data. ISPRS International Journal of Geo-Information **7**(12) (2018). https://doi.org/10.3390/ijgi7120459
33. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L.: Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 6877–6886 (2021). https://doi.org/10.1109/CVPR46437.2021.00681
34. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic Understanding of Scenes Through the ADE20K Dataset. International Journal of Computer Vision **127**(3), 302–321 (2019). https://doi.org/10.1007/s11263-018-1140-0
35. Zhou, W., Ming, D., Lv, X., Zhou, K., Bao, H., Hong, Z.: SO–CNN based urban functional zone fine division with VHR remote sensing image. Remote Sensing of Environment **236**(November 2019), 111458 (2020). https://doi.org/10.1016/j.rse.2019.111458, `https://doi.org/10.1016/j.rse.2019.111458`

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63783-4_9