KemptenCity - Semantic Segmentation of Urban Areas for Simulation

Peter Buckel[1,2], Stefan-Alexander Schneider[1], Jürgen Stübner[1], Florian Frank[3]

[1] Faculty of Electrical Engineering, University of Applied Science Kempten,
Kempten, Germany - (stefan-alexander.schneider, juergen.stuebner)@hs-kempten.de
[2] Baden-Wuerttemberg Cooperative State University (DHBW), Friedrichshafen, Germany - peter.buckel@dhbw.de
[3] Institute for Continuing Education, Knowledge and Technology Transfer (IWT),
Friedrichshafen, Germany – frank@iwt-bodensee.de

Keywords: Remote Sensing, Geoinformation Data, Deep Learning, Digital Twin, Simulation

Abstract

Autonomous driving and traffic flow simulation requires a realistic and accurate representation of the environment. Therefore, this research focuses on the semantic segmentation of aerial images for simulation purposes. Initially, a dataset was created based on true orthophotos from 2019 and Kempten's street cadaster, with true orthophotos being fully rectified aerial images. The chosen classes were oriented towards the subsequent conversion and usage in simulation. The proposed labeling workflow used cadaster data and demonstrated significant time efficiency compared to state-of-the-art datasets. Subsequently, a neural network was implemented that was trained and tested on the dataset. In addition, the network was also trained only on the lane markings to compare the network's performance. Both cases demonstrated excellent segmentation results. The generalizability was then tested on true orthophotos from 2021. The results indicated a solid generalizability, but still needs to be improved. Finally, the aerial information was converted into a 3D environment, that can be used in simulations. Our results confirm the usage of aerial imagery and street cadaster data as a basis for the simulations.

1. Introduction

Autonomous driving simulation has increased in popularity in recent years. Such simulations facilitate the examination of corner cases and reduce the need for driving extensive kilometers in real-world testing (Lemmer, 2019). Consequently, the market value of automotive simulation is estimated to reach a volume of approximately $2.9 billion USD in 2025 (Firm, 2018). Autonomous driving simulation requires a detailed representation of the environment and traffic areas. High-definition (HD) maps provide such information, including information about streets, the different types of lanes, and lane markings (HERE, 2024). Another field of simulation is the area of smart cities concepts designed to make cities more livable and sustainable (Batty et al., 2012). One significant component of a smart city is traffic-flow simulation, which also requires a representation of the environment and traffic areas. Driving simulation aims to identify the major congestion points and provide possible solutions (Lemmer, 2019).

Various data sources contribute to traffic analysis. For instance, HD maps are mainly recorded with mobile mapping systems and can be highly accurate depending on the sensors used. Conversely, open-source data such as OpenStreetMap (OSM) provide different information, however with varying precision (OpenStreetMap, 2024). Additionally, geographic information system (GIS) data and digital twins have recently been receiving increased attention. GIS involves acquiring, administrating, analyzing, and presenting spatial data (Lange, 2013), whereas a digital twin is a city's digital representation. For example, the town of Kempten has a digital twin, consisting of objects and shapes, such as street areas, walking paths, and buildings. The database also contains true orthophotos of Kempten (Fehr and Schneider, 2020). True orthophotos are fully rectified aerial images without hidden areas (Amhar et al., 1998). Furthermore, aerial images have recently been used for map creation due to the increasing development of remote sensing (Azimi et al., 2019b; Fischer et al., 2018; Azimi et al., 2019a). Remote sensing represents objects at or close to the earth's surface captured from a distance (Read and Torrado, 2009). The most significant advantage of remote sensing data is its scalability because of the vast area covered during recording. Through the combination of GIS data, open-source data, true orthophotos, and deep learning, further investigation of semantic segmentation of aerial images for simulation purposes can be made.

Therefore, a major contribution of this work lies in utilizing cadaster data for fast and efficient labeling of aerial imagery data. Subsequently, we present a dataset for generating a 3D traffic simulation environment. Building on this dataset, we built and evaluated a neural network for semantic segmentation. Finally, the transformation into a 3D environment is shown as an example.

2. The KemptenCity Dataset

Kempten has a 3D city model and an extensive cadaster, cataloging elements such as street areas, manhole covers, and walking paths. Furthermore, true orthophotos were recorded in 2019 and 2021. However, the data is irregularly maintained so that no reliable statement can be made about its accuracy. Additionally, Kempten's cadaster lacks information about lane markings. Therefore, we had the idea to use the orthophotos to generate the necessary information in combination with cadaster data.

2.1 Data Annotation

The combination of true orthophotos and cadaster data inspired us to develop a labeling workflow. The first step involved defining the necessary classes. Secondly, the proposed labeling workflow is explained.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-71-2024 | © Author(s) 2024. CC BY 4.0 License.

2.1.1 Classes An accurate representation of the environment is necessary for autonomous driving and traffic simulation. Researchers have developed a six-layer architecture for such simulations that structures objects to simulate scenarios (Pegasus, 2019). The most critical are the road and the road furniture layers. Layer one, the road layer, provides information about the topology and road geometry. Layer two, the road furniture layer, focuses on the infrastructure and objects, such as traffic signs, vegetation, guard rails, and manhole covers. Both layers are significant because traffic simulation would be impossible without a definition of the road and the infrastructure. Given the limitation of true orthophotos in providing enough information to extract traffic lights or signs, we focused on layer one. Therefore, the classes were oriented toward the Kempten Street cadaster to reduce the labeling time. Additionally, given the cadaster's limited information about parking lots, we decided to include only parking information about parking areas along the street and not parking lots. Moreover, manhole covers were annotated. This led to the following eight classes: street, traffic island, walking path, lane marking, parking, bus stop, manhole cover, and the background.

2.1.2 Labeling workflow Figure 1 illustrates the developed labeling workflow from top to bottom. The Kempten street cadaster and the true orthophotos served as input. The city of Kempten provided the street cadaster as a georeferenced shapefile, and the true orthophotos were provided in .tiff file format. First, the Kempten Street cadaster needed to be checked and adjusted. Therefore, the inputs were placed on top of each other. Next, each type of polygon (street, walking path, and more) was adjusted manually, meaning the polygon was manually adapted to the corresponding visual area from the true orthophoto. Afterwards, the classes were cross-checked against each other. For example, after straightening the street polygon, the walking path was adjusted and checked against the street. This process of adjusting and cross-checking with other labels was repeated for every class. Since the Kempten Street cadaster lacks information for the lane markings, a separate labeling process was needed. Therefore, we developed an automated labeling process to provide precise lane marking labels. This suits image-processing algorithms since lane markings stand out from the street in color and shape. Lane markings are located either on the street or on the walking path. Because of that, both were extracted from the orthophotos. This step reduced the influence of areas outside the traffic area. Subsequently, image-processing algorithms were applied. The computer-vision-based code uses the HSV color space as a solution to seek brightness-independent colors. The chosen method provides lane-marking labels based on the HSV color space. Finally, the lane-marking polygons were corrected manually.

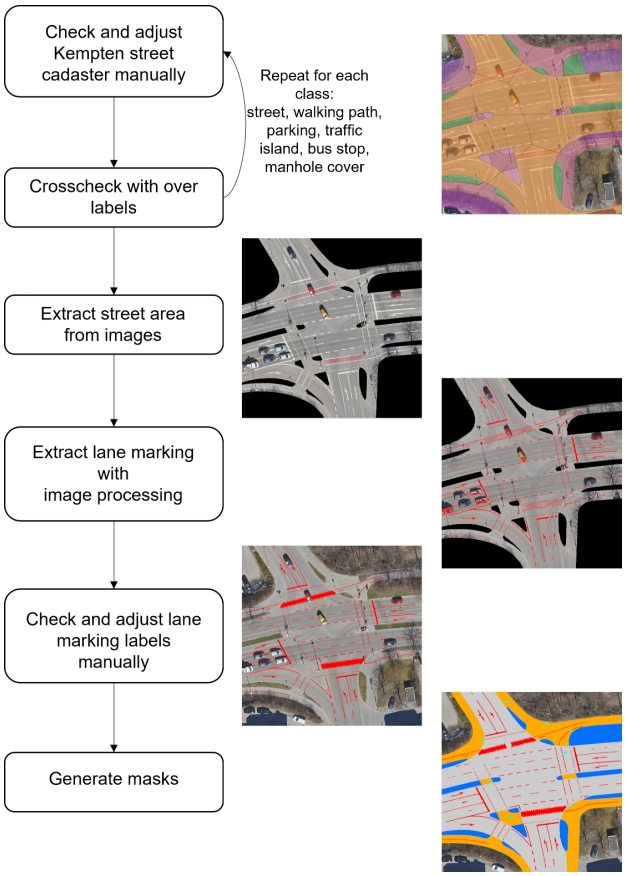

Figure 1. Labeling workflow for the KemptenCity dataset using Kempten's cadaster data. First, the cadaster data is adjusted to extract all classes except the lane markings. Next, the traffic area was extracted and the lane markings were detected. After the lane markings were manually corrected, the final mask was extracted.

2.2 Comparison with State-of-the-Art Datasets

There are several datasets, such as SkyScapes (Azimi et al., 2019b), Potsdam (ISPRS, 2022), Vahingen (ISPRS, 2022), and AerialLanes18 (Azimi et al., 2019a), that are comparable with the annotated KemptenCity dataset. Table 1 presents a comparison of these datasets. The datasets of Vahingen and Potsdam are open-source and provide labels for classes such as impervious surfaces, buildings, vegetation, and more. They distinguish the number of images, image size, ground sample distance (GSD), and aerial coverage. The datasets were manually labeled (ISPRS, 2022). The AerialLanes18 dataset consists of 20 RGB images of size 5616 x 3744 pixels with a GSD of 13 cm/pixel. Experts labeled the data with pixel accuracy and provided lane marking labels (Azimi et al., 2019a). In contrast, SkyScapes has pixel-accurate labels for 31 classes, including several classes for the different road types and lane markings, and it differentiates five tasks (Azimi et al., 2019b):

- SkyScapes-Dense: 20 classes with multiple (sub-) classes merged into a single category.
- SkyScapes-Lane: 12 different lane marking classes and a non-lane marking class.
- SkyScapes-Dense-Category: 11 classes, such as nature, driving area, parking, and road features.
- SkyScapes-Dense-Edge-Binary: Binary edge detection for fine-grained segmentation.
- SkyScapes-Dense-Edge-Multi: Multiclass edge detection for fine-grained segmentation.

All datasets, including SkyScapes and Vahingen, cover the street area extraction. The Potsdam and Vahingen datasets segment aerial images into only a few classes, such as buildings or vegetation, although neither provides classes for parking or lane markings. Conversely, the KemptenCity dataset covers an area of 7.5 km² and consists of 30 images with a size of 5000 x 5000 pixels. Furthermore, the datasets combine urban and

	KemptenCity	*SkyScapes (Azimi et al., 2019b)*	*AerialLanes18 (Azimi et al., 2019a)*	*Potsdam (ISPRS, 2022)*	*Vahingen (ISPRS, 2022)*
Classes	8	31	1	6	6
Images	30	16	20	38	33
Dimension [pixel]	5000x5000	5616x3744	5616x3744	6000x6000	2493x2063
GSD [cm/pixel]	10	13	13	5	9
Coverage [km²]	7.5	5.69	N/A	3.42	1.36

Table 1. Comparison of the KemptenCity dataset with current datasets. The table compares the number of classes and images, the dimension, GSD, and the coverage.

rural regions (highways). The true orthophotos have a GSD of 10 cm/pixel. We introduced two datasets: a multiclass dataset providing ground-truth labels for eight classes relevant for reconstructing the traffic area and a binary KemptenCity dataset that only provides pixel-accurate lane marking labels. For example, one simulation scenario involves a pedestrian crossing a street in front of a car. In this case, it is necessary to have information about the street, lane markings, traffic island, and walking path. The motivation behind the binary dataset was to produce better segmentation results. Compared to state-of-the-art datasets, the labeling time is the most significant difference. For instance, the whole labeling process for SkyScapes took approximately 3200 hours (Azimi et al., 2019b). In comparison, the KemptenCity dataset was completed in about 500 hours, providing pixel-accurate lane marking labels and approximately two-pixel accuracy for the other classes.

3. Semantic Segmentation

This section first introduces the metrics for evaluation and then discusses state-of-the-art segmentation networks.

3.1 Evaluation Metrics

The evaluation of semantic segmentation models requires specific metrics, with the most common metric being the Intersection over Union (IoU). The intersection between the target and the prediction is the number of correctly classified pixels. The union is the number of pixels in the prediction or the mask. The IoU describes the overlapping region between the ground truth and the predicted label. Precision and recall are additional metrics. Precision represents the number of predicted objects with corresponding ground truth, while recall compares the number of objects annotated in ground truth with the positive captured predictions. The general rule is that higher values of precision and recall indicate better model performance. Equal values of both metrics are ideal. If the precision exceeds the recall, the model predicted reasonably well, but the total number is small. In contrast, lower precision and higher recall suggest the model predominantly classifies correctly but also includes many wrong predictions.

3.2 State-of-the-Art Segmentation Networks

Semantic segmentation of aerial images is a common task for which there are various approaches. The most common ones are presented in the following section.

3.2.1 Semantic Segmentation Networks The U-Net architecture is the most widely used segmentation network introduced by (Ronneberger et al., 2015). It consists of an encoder (downsampling) and a symmetric decoder (upsampling) connected through skip connections. Among other things, the authors reduced the parameters and achieved reliable segmentation results. Thus, the U-Net architecture has been widely extended to other networks such as FPN (Lin et al., 2017), PSPNet (Zhao et al., 2017), and Deeplapv3+ (Chen et al., 2018). Long et al. (2015) introduced fully convolutional networks (FCNs) for end-to-end, pixel-to-pixel image segmentation. FCNs accept arbitrary-sized images and generate a correspondingly sized output. The authors were able to produce pixel-wise segmentation results. Chen et al. (2014) first introduced DeepLab. DeepLab adapts deep convolutional neural networks (DCNNs) with a fully connected conditional random field (CRF) to achieve better localization properties. It also utilizes atrous convolution instead of deconvolutional layers for upsampling. Atrous convolution uses the dilation rate for a wider field of view, thereby allowing effective upsampling without increasing the computation time or the number of parameters. According to Chen et al. (2014), DeepLab has three benefits: speed, accuracy, and simplicity. Chen et al. (2017a) further developed the basic DeepLab version and proposed a DeepLab architecture with an atrous spatial pyramid pooling (ASPP) layer for extracting multiscale features. DeepLabv3 enlarged the ASPP layer and integrated ResNet as a backbone (Chen et al., 2017b). The latest version, DeepLabv3+, uses atrous separable convolution, which combines depth-wise and point-wise convolution. It uses a modified Xception model as a backbone. DeepLabv3+ achieved a notable 89.0% mean intersection over union (IoU) score on the 2012 PASCAL VOC challenge (Chen et al., 2018).

3.2.2 Semantic Segmentation of Aerial Images Xin et al. (2019) introduced a binary road network segmentation method. Liu et al. (2019) developed RoadNet, a multitask CNN that solves three tasks to extract the road: Surface segmentation, edge detection, and centerline extraction. Kaiser et al. (2017) compared OSM data for labeling with highly accurate, manually labeled aerial images for road and building extractions. Their findings indicated that using automatically labeled images for training led to better segmentation performance and increased generalization if the data was used on a large scale. Additionally, they observed that OSM data could replace manually annotated data without a decrease in accuracy. Fischer et al. (2018) developed a computer vision-based algorithm for lane-marking extraction. They trained a random forest classifier based on the color and texture of lane markings. Since most aerial images contain vast areas of irrelevant background information, the authors applied a mask derived from OSM data. Azimi et al. (2019a) presented Aerial LaneNet, a novel network for binary lane-marking segmentation. Aerial LaneNet is based on a symmetric, fully convolutional neural network with wavelet decomposition. The LaneNet is an encoder-decoder architecture with a VGG 16 model as the encoder. The VGG16 was pretrained on the ImageNet dataset to overcome overfitting problems. Azimi et al. (2019b) also introduced SkyScapes, a dataset annotating 31 classes, including roads, buildings, and lane markings; it also consists of 12 (sub-) classes of lane markings. Finally, Azimi et al. also developed SkyScapesNet for feature extraction. SkyScapesNet is a multitask segmentation network with three branches: dense semantic segmentation, multi-edge detection, and binary-edge detection. FC-DenseNet is used as a primary baseline. Due to the significant variance in the size of the SkyScapes classes, the authors concentrated on feature

extraction. For example, they developed a CRASPP module inspired by atrous spatial pyramid pooling (ASPP). They reversed ASPP and concatenated it with the original ASPP to obtain receptive fields for small and large objects (Azimi et al., 2019b). In summary, these diverse architectures solve similar segmentation problems. The knowledge gained from state-of-the-art methods is combined and further developed to segment the traffic area.

4. Experiments

This section explains the process of finding the model. The dataset was split into three subsets: 50% for training, 30% for validation, and 20% for testing. The true orthophotos were cropped into 512 x 512-pixel patches. The model for the multiclass segmentation was trained for 30 epochs using a learning rate of 1×10^{-5} and a batch size of 1. Moreover, the Adam optimizer, ReLU (El-Amir and Hamdy, 2019), and the dice loss (Crum et al., 2006) function were used. For data augmentation, the images were cropped with a 50% overlap horizontally and vertically. Additionally, the patches were flipped, rotated, and sharpened, and the contrast-limited adaptive histogram equalization (CLAHE) (Pizer et al., 1987) algorithm was applied. During the training of the binary problem, the learning rate was changed to 1×10^{-4}, without data augmentation. The network was trained on a Tesla V100.

4.1 Model finding

The developed model is based on knowledge from state-of-the-art approaches and was further developed in this work, with the network structure based on U-Net (Ronneberger et al., 2015). U-Net's basic encoder-decoder architecture allows for easy implementation and adjustment, allowing the encoder to be changed using a pre-trained model. As shown by Azimi et al. (2019b,a), transfer learning improves the results and prevents overfitting in small and unbalanced datasets. Transfer learning could solve the biggest drawback of the KemptenCity dataset: its unbalance. For this reason, the encoder of the classic U-Net was replaced by a pre-trained model. Transfer learning uses pre-trained weights from existing datasets. Azimi et al. (2019a) used the ImageNet dataset for transfer learning for lane-marking segmentation. Therefore, we used pre-trained models on the ImageNet dataset and compared three pre-trained encoders: ResNet101, VGG19, and DenseNet201. Table 2 shows the results of these trained models on the test dataset. First, we compared the different pre-trained encoders with the U-Net. The combination of DenseNet201 and U-Net reached a mean IoU (mIoU) of 59.71%, significantly outperforming the other configurations. For this reason, it was used in the next step. The chosen pre-trained DenseNet-201 has 201 layers and was pre-trained on the ImageNet dataset (Huang et al., 2017). Further experimentation involved adjusting the depth of the DenseNet201-Unet combination (three, four, and five), which was defined in our case by the number of downsampling and upsampling steps. The DenseNet201-Unet, with a depth of four, was already the best and, more importantly, did not overfit. These results can be further improved by incorporating an ASPP layer as a local feature extractor, as shown in (Chen et al., 2017a). The network reached a mean IoU of 69.21%. The final network, DenseUnet ASPP, outperformed the original U-Net's performance by 17.03%.

4.2 Architecture of DenseUnet_ASPP

The final architecture of the DenseUnet_ASPP is shown in Figure 2. The true orthophoto is first cropped and fed into the network from the left to the right. The encoder is based on the pre-trained DenseNet201. The bridge, also called the bottleneck, consists of a dense block, the ASPP, and a dropout layer. In contrast, each decoder block consists of a transpose convolution, concatenation, dropout, and convolution block. Figure 2 illustrates the structure of the convolution block and the ASPP. The convolution block contains convolution, batch normalization, and ReLU activation.

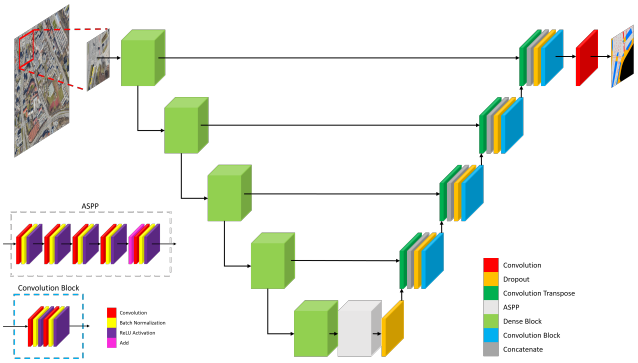

Figure 2. Architecture of the DenseUnet_ASPP.

5. Results

5.1 Multiclass Segmentation

The developed model was trained on the KemptenCity-Multiclass dataset. The results are analyzed more deeply here. Table 3 shows the IoU for each class of the DenseUnet_ASPP, illustrating that lane markings and the street were classified well. Moreover, the NN also differentiated between the walking path and the street. However, other classes, such as manhole covers and parking, need further improvement. The unbalanced dataset caused this unbalanced segmentation of the classes. Figure 3 shows the RGB image (left), the ground-truth mask (middle), and the predicted results (right). It can be seen that the network's results were quite close to the mask. The results of the streets and the lane markings were especially accurate.

Name	Depth	ResNet	VGG	DenseNet	ASPP	Mean IoU [%]	Accuracy [%]	Precision [%]	Recall [%]
U-Net	4	-	-	-	-	52.18	83.90	83.94	83.88
ResNet101_U-Net	4	X	-	-	-	38.38	67.94	68.08	67.80
VGG19_U-Net	4	-	X	-	-	48.30	82.33	82.70	82.12
DenseNet_U-Net	4	-	-	X	-	59.71	89.03	89.09	89.00
DenseNet_U-Net	3	-	-	X	-	43.01	52.35	52.38	52.33
DenseNet_U-Net	5	-	-	X	-	58.95	91.40	91.46	91.38
DenseUnet_ASPP	4	-	-	X	X	**69.21**	**93.58**	**93.59**	**93.57**

Table 2. Comparison of different models trained on the KemptenCity dataset.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-71-2024 | © Author(s) 2024. CC BY 4.0 License.

Name	IoU [%]									Accuracy	Precision	Recall
	Mean	Background	Lane Marking	Bus Stop	Traffic Island	Manhole Cover	Parking	Walking Path	Street	[%]	[%]	[%]
DenseUnet_ASPP	69.21	92.10	65.81	94.48	59.55	40.82	62.35	57.22	81.40	93.58	93.59	93.57

Table 3. Results of the final trained model, DenseUnet_ASPP, on the test dataset per class.

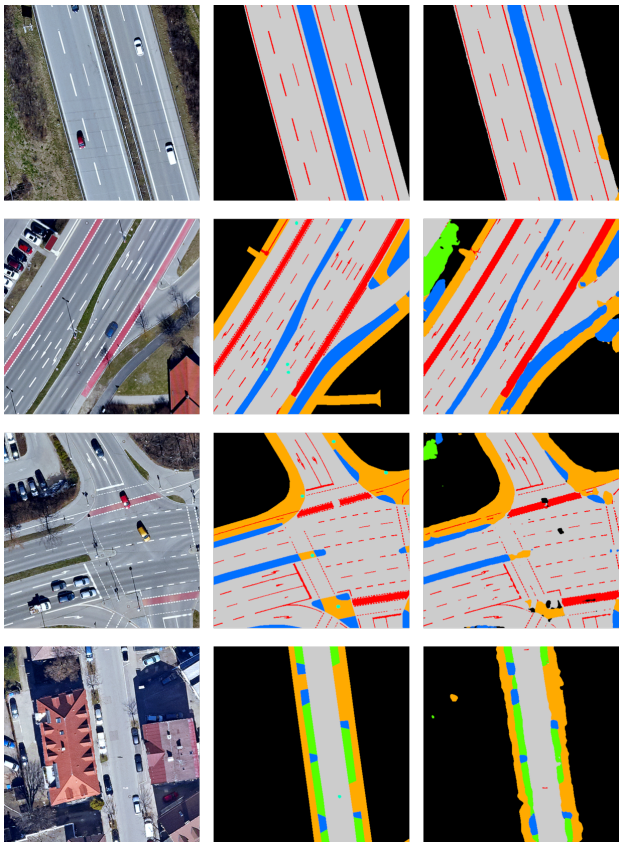

Figure 3. Results of the DenseUnet_ASPP applied to the KemptenCity-Multiclass images from the test dataset: RGB (left), GT mask (middle), predicted (right).

5.2 Binary Segmentation

Table 4 summarizes the results of the binary segmentation of the test dataset.

Name	Epochs	IoU [%]			Accuracy	Precision	Recall
		Mean	Background	Lane Marking	[%]	[%]	[%]
DenseUnet_ASPP	30	85.34	99.72	70.96	99.73	99.73	99.73
DenseUnet_ASPP	60	85.52	99.73	71.32	99.73	99.73	99.74

Table 4. Results of binary segmentation.

After 30 epochs, the DenseUnet_ASPP achieved an overall mean IoU of 85.34%, with an IoU of 70.96% for the lane-marking class. It slightly improved after 60 epochs to 85.52% mean IoU and 71.32% for lane markings. Figure 4 shows the results on the test dataset, with true positive lane markings in green, unsegmented lane markings, false negative (FN), in blue, and false positives (FP) in red.

6. Generalization

One network requirement is to generalize the results to other true orthophotos. Therefore, the network was applied to unseen Kempten images in the spring of 2021. These true orthophotos have a GSD of 7.5 cm/pixel and a 10000 x 10000 pixel resolution. In comparison, the true orthophotos from 2019 have

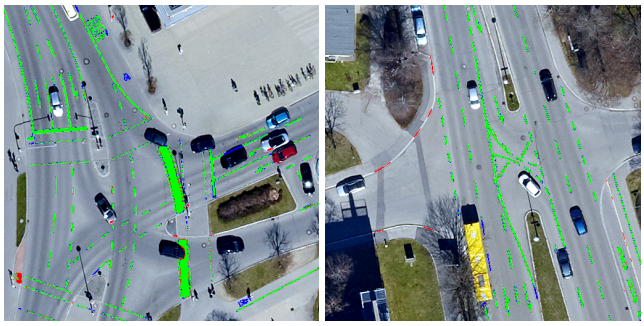

Figure 4. Results of the Binary segmentation. Green: TP, blue: FN, red: FP

a different brightness (the ones from 2021 are brighter). Figure 5 shows the results of the DenseUnet_ASPP applied to the orthophotos from 2021. Neither of the areas covered by the true orthophotos from 2021 was included in the KemptenCity dataset. The street and lane markings were well-segmented from the images. However, the segmentation of walking paths, manhole covers, parking, and traffic islands require improvement. While the binary predicted lane markings were good, the generalizability needs further improvements, especially for arrows, lane markings in shadows, and bicycle-way markings.

7. Reconstructed 3D Environment

The network's output could be further used in traffic simulations. Therefore, a conversion process was developed to convert the data into a 3D environment. The initial output, RGB images with labeled pixels, required conversion into shapefiles containing polygons, where each point has its own global coordinate. This conversion could be executed through several methods or programs. Lastly, the data must be converted into a 3D representation. Therefore, the shapefile was imported, extruded, and textured. The whole conversion process was partly automated.

However, the output of the multiclass segmentation network was flawed, making it insufficiently accurate for direct conversion. Consequently, the output must be post-processed, which can be performed in several ways with different levels of complexity. For example, a dashed lane marking with unclear borders must be classified and reshaped to its original shape, a step that can be very time-consuming. Thus, the ground truth data of the KemptenCity dataset was converted. Figure 6 shows the reconstructed intersection in front of the University of Applied Science Kempten. It consists of 3D buildings, true orthophotos, walking paths, manhole covers, parking areas, bus stops, street areas, and lane markings. This 3D environment can already be used for various simulations.

8. Conclusion

This work aimed to semantically segment the traffic areas from aerial images for simulation purposes. Therefore, this work investigated whether cadastral and open-source GIS data can be used to train and improve an NN for this purpose. We introduced a novel labeling workflow using cadaster data to decrease

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-71-2024 | © Author(s) 2024. CC BY 4.0 License.

Figure 5. DenseUnet_ASPP (multiclass and binary) network applied to the city of Kempten orthophotos data from 2021.

Figure 6. 3D representation of Kempten with the generate traffic area and the integrated LOD2 buildings placed on top of the orthophoto.

the labeling time significantly. The resulting KemptenCity dataset consists of 30 true orthophotos, each with a 5000 x 5000 pixel resolution, covering eight classes. The chosen classes include the road geometry and parts of the road infrastructure. In addition, we provided pixel-accurate labels for lane markings. Next, we presented a segmentation network to validate using the KemptenCity dataset. We trained and evaluated the network on all classes and only the lane markings. Overall, both binary and multiclass networks achieved good segmentation results. Finally, generalizability was demonstrated using the Kempten true orthophotos from 2021. While the generalizability was good for the street and dashed lane markings, the results of other classes could be improved. Finally, the neural network output must be converted into a 3D environment. However, the network output was fuzzy; hence, the conversion method was demonstrated based on the ground truth data from the dataset. The defined conversion method was relatively fast and consisted of a few partially automated steps. It was found that the generated environment could be imported into certain simulation tools. In contrast, specialized autonomous driving simulation tools require a separate semantic road description.

We plan to extend the dataset for future work to increase the results and generalizability. Furthermore, some simulation tools require specialized formats containing semantic descriptions of traffic areas. ASAM OpenDRIVE provides such information and uses an s-t coordinate reference system for the environment, allowing for the positioning of objects, such as traffic signs, to be placed at a right angle to the street (ASAM, 2021). Therefore, the development of a conversion method is the next step.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-71-2024 | © Author(s) 2024. CC BY 4.0 License.

9. Acknowledgements

We would like to thank the city of Kempten for providing the data used to generate the KemptenCity dataset.

References

Amhar, F., Jansa, J., Ries, C. et al., 1998. The generation of true orthophotos using a 3d building model in conjunction with a conventional dtm. 32, INTERNATIONAL SOCIETY FOR PHOTOGRAMMETRY & REMOTE, 16–22.

ASAM, 2021. ASAM OpenDRIVE. https://www.asam.net/standards/detail/opendrive/. Accessed: 2024-01-27.

Azimi, S. M., Fischer, P., Korner, M., Reinartz, P., 2019a. Aerial LaneNet: Lane-Marking Semantic Segmentation in Aerial Imagery Using Wavelet-Enhanced Cost-Sensitive Symmetric Fully Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57number 5, 2920–2938.

Azimi, S. M., Henry, C., Sommer, L., Schumann, A., Vig, E., 2019b. SkyScapes Fine-Grained Semantic Understanding of Aerial Scenes. *IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 7392–7402.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., Portugali, Y., 2012. Smart cities of the future. *The European Physical Journal Special Topics*, 214, 481–518.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40number 4, IEEE, 834–848.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Crum, W., Camara, O., Hill, D., 2006. Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis. *IEEE Transactions on Medical Imaging*, 25, 1451–1461.

El-Amir, H., Hamdy, M., 2019. *Deep learning pipeline: building a deep learning model with TensorFlow*. Apress.

Fehr, W., Schneider, S.-A., 2020. Digitale Zwillinge für automatisierte Fahrzeuge. *Digitale Transformation des ÖPNV*, 2020.

Firm, M. R., 2018. Automotive Simulation Market. https://www.marketsandmarkets.com/Market-Reports/automotive-simulation-market-242816468.html. Accessed: 2024-01-27.

Fischer, P., Azimi, S. M., Roschlaub, R., Krauß, T., 2018. Towards HD Maps from Aerial Imagery: Robust Lane Marking Segmentation Using Country-Scale Imagery.

HERE, 2024. HERE HD Live Map: Autonomous Driving System: Platform. https://www.here.com/platform/HD-live-map. Accessed: 2024-01-27.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q., 2017. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

ISPRS, 2022. 2D Semantic Labeling Contest. *2D Semantic Labeling*. https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx. Accessed: 2024-01-27.

Kaiser, P., Wegner, J. D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning Aerial Image Segmentation From Online Maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11), 6054–6068.

Lange, N., 2013. *Geoinformatik: in Theorie und Praxis*. 3., vollst. überarb. u. akt. aufl. 2013 edn, Springer Berlin Heidelberg Imprint Springer Spektrum, Berlin, Heidelberg.

Lemmer, K., 2019. *Neue autoMobilität II*. utzverlag GmbH.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI, 936–944.

Liu, Y., Yao, J., Lu, X., Xia, M., Wang, X., Liu, Y., 2019. RoadNet: Learning to Comprehensively Analyze Road Networks in Complex Urban Scenes From High-Resolution Remotely Sensed Images. *IEEE Transactions on Geoscience and Remote Sensing*, 57, 2043–2056.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

OpenStreetMap, 2024. OpenStreetMap. https://www.openstreetmap.org. Accessed: 2024-01-27.

Pegasus, 2019. Basics for Testing — Booth No. 15, Scenario Database. https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/15_Scenario-Database.pdf. Accessed: 2024-01-27.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., Zuiderveld, K., 1987. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39number 3, 355–368.

Read, J., Torrado, M., 2009. Remote sensing. *International Encyclopedia of Human Geography*, Elsevier, Oxford, 335–346.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Xin, J., Zhang, X., Zhang, Z., Fang, W., 2019. Road Extraction of High-Resolution Remote Sensing Images Derived from DenseUNet. *Remote Sensing*, 11number 21, 2499.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI, 6230–6239.