International Journal of Artificial Intelligence,
ISSN 0974-0635; Int. J. Artif. Intell.
2025, Volume 23, Issue Number 2

Machine Learning Methods for Analysis of Photo/Video Files from Cameras

O. Akylbekov[1], **Sh. Alshynov**[2], **A. Tulegenova**[3], **L. Ramazanova**[4] And **Zh. Baimukanova**[5]

[1]Department of Software Engineering, Satbayev University, Almaty,
Republic of Kazakhstan
Email: akylbekovolzhas7@gmail.com

[2]Department of Computer Engineering, Astana It University,
Astana, Republic of Kazakhstan
Email: s.alshynov@hotmail.com

[3, 5]School of Engineering and Digital Sciences,
Nazarbayev University, Almaty, Republic of Kazakhstan
Email: a-tulegenova@outlook.com, zbaimukanova@outlook.com

[4]Department of Pedagogy, Psychology and Primary Education,
K. Zhubanov Aktobe Regional University, Aktobe, Republic Of Kazakhstan
Email: lramazanova2@hotmail.com

ABSTRACT

The research relevance is determined by the need to improve the accuracy and speed of image segmentation in complex urban environments for automated monitoring systems. The study aimed to develop and evaluate the effectiveness of deep neural networks for solving the problem of semantic segmentation of photo and video data obtained from surveillance cameras. The study tested DeepLabv3+ and U-Net architectures adapted for real-time image processing. Data augmentation methods, including adaptive contrast enhancement and illumination normalisation, have been implemented to improve the algorithms' resistance to adverse lighting conditions and weather factors. Experimental results on the Cityscapes and ADE20K datasets showed that the DeepLabv3+ model achieved an average IoU of 0.73 on the test data, and the use of optimised post-processing mechanisms reduced the accuracy drop to IoU = 0.65 in difficult shooting conditions. The image processing speed was 40 ms per frame, making the model suitable for use in real-time systems. The obtained results confirm the effectiveness of the proposed architectures and emphasise the need for further optimisation of the algorithms to improve the segmentation accuracy in low-light conditions. The practical significance of the study is to increase the reliability of traffic monitoring and public safety in the urban environment.

Keywords: Deep neural networks, semantic image segmentation, U-Net architecture, DeepLabv3+, Vision Transformers, dataset, PyTorch, video analytics, automated surveillance systems, urban environment monitoring.
Mathematics Subject Classification: 68T45, 68U10

1. INTRODUCTION

CCTV cameras are widely used in public safety, traffic monitoring and urban infrastructure management. At the same time, the amount of data generated by such systems significantly exceeds the possibilities of manual analysis, which necessitates the introduction of automated approaches. In this context, machine learning methods are effective for automating the processing of visual information, ensuring high accuracy and speed of analysis of large arrays of images and video data.

One of the key areas of visual data analysis is semantic segmentation, which allows not only to identify objects in images but also to classify each pixel, forming a complete picture of the scene. This approach is crucial for solving problems related to the interpretation of complex urban scenes, where it is necessary to account for the interaction of different types of objects: vehicles, pedestrians, urban infrastructure elements, and other road users.

Modern research confirms the high efficiency of deep learning methods for processing large amounts of graphic data and recognising visual elements in complex shooting conditions. In particular, the use of convolutional neural networks (CNNs) and transformers allows for achieving high accuracy in semantic segmentation and classification tasks (Chornyy et al., 2022). Several studies have proposed multi-level architectures that include image preprocessing, segmentation, and subsequent analysis to improve the reliability of the results.

Rakhmetulayeva et al. (2023) created a computer model capable of automatically analysing photographs to identify key elements of tourist behaviour and perception in specific tourist areas. To solve these problems, the authors used deep learning methods, in particular CNN, which can automatically extract features from images, which greatly facilitates the process of detecting objects and behavioural patterns in photos.

Du et al. (2020) presented a semantic segmentation method using DeepLabv3+ in combination with object-oriented image analysis for processing high-resolution satellite images. The proposed approach involves the integration of deep neural networks with traditional image analysis methods, which allows for increased accuracy and detail of segmentation. The study noted that this technique is particularly effective for recognising complex objects in satellite images due to the combination of contextual information and deep learning.

Pemasiri et al. (2020) developed a multimodal approach to semantic image segmentation that combines different data sources to improve the accuracy of analysis. The researchers used deep neural networks to combine information from different modalities, which contributes to a deeper understanding of the scene and improved object recognition. The study emphasised that the method reduced the impact of noise and improved the generalisability of the models, which is especially important in analysing complex visual data.

Zaitoun and Aqel (2015) provided a theoretical overview of various image segmentation techniques, including both traditional and modern methods, including those based on machine learning. At the same time, the authors presented a comprehensive analysis of segmentation approaches in the context of their application in various fields, such as medical diagnostics, environmental monitoring, and industrial processes. In turn, Minaee et al. (2021) noted that due to the high level of adaptability, the semantic segmentation method is effective even in complex environments, such as urban environments, where the interaction between different types of objects occurs in a dynamic space, or natural landscapes, where there are various natural elements with fuzzy boundaries.

The issue of analysing photos and videos from surveillance cameras in the urban environment was studied by Muhammad et al. (2022). The study considered the problem of using semantic segmentation to detect and recognise various objects in difficult road conditions, as well as the prospects for the development of this technology in the future. In turn, Cao and Bao (2020) analysed the use of U-Net model ensembles to improve semantic segmentation in geospatial data processing. The study proposed fundamentally new approaches to improving segmentation methods for more efficient monitoring of the environment and natural resources using satellite images. Pohudina et al. (2020) and Sevak et al. (2017) presented a method for identifying and counting objects based on the integration of computer vision and machine learning algorithms. The study analysed the possibility of automating image analysis in technical systems, which made it possible to increase the efficiency of visual data processing in various engineering applications. Khan et al. (2011) considered deep neural network architectures focused on the semantic segmentation of medical images. The study substantiated the effectiveness of using convolutional neural networks and transformers in the tasks of automatic analysis of medical data, emphasising their ability to improve diagnostic accuracy.

Despite significant progress in the field of semantic segmentation, there are still unresolved issues that limit the effectiveness of these methods in urban environments, namely, the decrease in accuracy when the shooting conditions change. Most models demonstrate positive results in controlled environments, but their effectiveness decreases in low light, precipitation, or complex scenes with high object density. Most studies are focused on either satellite imagery or medical imaging, while in urban environments, high object dynamics, complex background scenes, and sudden changes in lighting need to be addressed.

The research novelty is determined by the development of advanced data processing methods, including adaptive contrast correction, illumination normalisation, and enhanced local object boundary detection, which increases the stability of models in difficult shooting conditions. DeepLabv3+ and U-Net architectures were optimised in the study to address the specifics of the urban environment, which improved the quality of segmentation in images with high object density and increased the accuracy of identifying key elements of the scene. In addition, a balance between accuracy and processing speed has been achieved, making the proposed methods suitable for automated systems and enabling their application for real-time analysis of video streams.

The study aimed to develop and analyse the effectiveness of deep neural networks for the task of semantic segmentation of photo and video data coming from surveillance cameras in an urban environment. The research is aimed at improving the accuracy of segmentation under difficult shooting conditions and optimising algorithms for use in real-time in the city of Almaty.

2. MATERIALS AND METHODS

Type of research and timeframe. The present study is empirical research aimed at developing and validating a deep learning model for analysing photo and video data from surveillance cameras located in the urban environment of Almaty. The study was conducted from May 2023 to December 2024.

Data collection. The data for the study was obtained from surveillance cameras located in urban areas of Almaty on major transport routes, intersections and public places, such as Abay Avenue, Dostyk Avenue, Al-Farabi Street, as well as in the areas of Republic Square, Pervomaisky Park and Mega Alma-Ata and Dostyk Plaza shopping centres. In addition, open video surveillance datasets were used to test the model in various scenarios and ensure its versatility. These datasets include AI City Challenge (2025) (for traffic and incident analysis), Open Images Dataset 17 (2025) (with annotated images and videos), KITTI Vision Benchmark Suite (2025) (for traffic), and Waymo Open Dataset (2025) (for autonomous transport systems). Each of these datasets was adapted to the specifics of the task under study, in particular, additionally annotated videos were processed to test the quality of classification, segmentation and tracking. Given the specifics of the task, an important aspect was also the use of data augmentation techniques, such as rotations, changes in illumination, and the addition of noise, which enabled the model to work in different conditions.

Methods used. The choice of methods was determined by the specifics of the tasks, including object detection, classification, segmentation, and tracking of moving elements in a sequence of video frames.

Segmentation methods. To solve image and video segmentation tasks, CNN was used as the main architecture to separate the image into objects and backgrounds for accurate scene interpretation. DeepLabv3+ was chosen for traffic monitoring because it demonstrates high robustness to changing lighting conditions and weather factors. This is especially relevant in urban environments where shadows, headlight glare and other external influences can be present. This model uses deep convolutional networks with attentional mechanisms, which improves the accuracy of object segmentation in complex scenes, such as dense traffic or pedestrians. The U-Net model was used for accurate object detection, as it is effective even with limited training data. Its symmetric architecture with layer concatenation improved object localisation, improving the accuracy of boundary identification between scene elements. This is especially relevant in tasks where a clear separation between the road, sidewalks, cars and people is important. Segment Anything Model (SAM) was chosen for automatic data markup because it can adapt to new images without additional training. This significantly reduced the labour costs for annotation and made the system more flexible when shooting conditions change. In addition, SAM effectively separates complex objects, even if their annotations are incomplete or partially missing.

Vision Transformers (ViTs) were used to improve segmentation in urban environments, as they are better able to handle the high-density images typical of urban areas. Thanks to the self-attenuation mechanism, ViTs incorporate the global context of the scene, which is especially critical when analysing complex intersections, multi-lane roads and crowds. Vision Transformers were chosen because of the ability to efficiently process images using self-attention mechanisms, which improves processing of the spatial dependencies between objects in the image. In contrast to CNNs, which operate on local features, ViTs analyse the image, which is especially useful in complex scenes with many overlapping objects. To train the models, Cityscapes, ADE20K, Mapillary Vistas, and KITTI datasets containing annotated images and videos adapted for segmentation tasks in urban environments were used.

Classification methods were used to detect and identify objects in the video footage, which was used to recognise different categories of urban scene elements, such as cars, pedestrians, cyclists and road signs. YOLOv4 was chosen as the main classification model as it optimally combines accuracy and processing speed, which is critical for real-time video analysis. YOLOv4 uses a single-stage architecture that significantly speeds up data processing compared to two-stage methods such as Faster R-CNN. An additional advantage of YOLOv4 is its pre-training on the COCO dataset, which reduced model adaptation time and improved its accuracy in urban environments without the need to collect huge amounts of proprietary data. This makes it an effective solution for video surveillance systems that require fast and accurate object detection.

Tracking methods. Tracking objects is substantial in analysing video streams, as it can be used to track movements in a sequence of frames. This is especially relevant for monitoring pedestrians, vehicles, and other moving elements in an urban environment. For this task, the Deep SORT model was used, which has shown high efficiency in tracking objects in dense scenes. The main advantage of Deep SORT is the use of the Kalman filter, which accurately predicts the trajectory of objects, even if they are partially overlapped with other elements of the scene. In addition, Deep SORT uses deep neural networks to identify objects, which reduces the probability of errors if the position, lighting, or angle changes. Thanks to these features, the model works efficiently in real-time, making it the best choice for analysing urban video streams.

To implement the semantic segmentation model, modern frameworks and high-performance hardware were used to ensure efficient training and data processing. PyTorch was used to implement DeepLabv3+, as it provides flexibility in architecture configuration and allows for efficient training of deep models on large data sets. TensorFlow 2.0 was chosen to work with YOLOv4 and U-Net because it supports GPU accelerators, which significantly improves performance, and includes optimised Keras and OpenCV libraries for working with images and video. The models were trained on NVIDIA Tesla V100 GPUs, which enabled parallel processing of large amounts of data at high speed, ensuring high accuracy and stability of the algorithms.

Interpretation of the results. Both quantitative and qualitative approaches were used to interpret the results of the study. The quantitative analysis included metrics for evaluating the performance of the models, which were used to test the ability to adapt to new images and to check the robustness of various variations that may occur in practical applications. The metrics used to evaluate the results included accuracy, recall, precision, and F1-measure for classification, as well as metrics for segmentation, such as Intersection over Union (IoU) and Dice Coefficient. The study evaluated the model's performance using the pixel-level accuracy (PA) metric. This metric was used to determine the accuracy with which the model classifies each pixel in the processed images, which is critical for tasks requiring high detail, such as segmentation. These metrics were used to evaluate the performance of the models in object recognition, segmentation, and classification tasks.

The qualitative approach involved visualising the results of the models, including overlaying the segmented objects on the original images and video, which was used to assess the accuracy and consistency of the segmentation with real data. Furthermore, comparisons were made with the results of other models and data from open data sets to determine the versatility and adaptability of the developed model to the conditions of the real urban environment.

3. RESULTS

Development and validation of a neural network model. The model was developed using modern approaches in deep learning and computer vision, applying methods of classification, segmentation and object tracking. The Mask R-CNN model was used to detect individual objects in the frame and perform their segmentation at the level of instances (Figure 1).

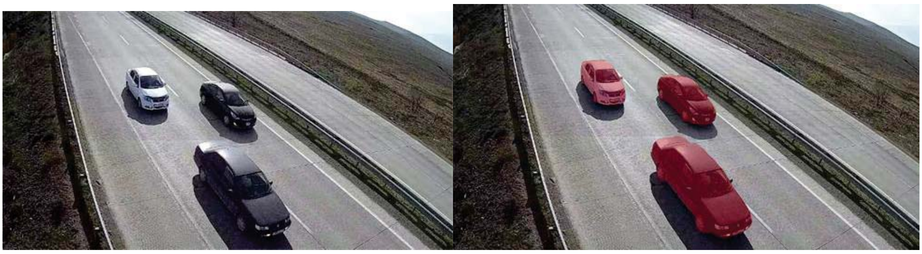

Figure 1. Example of semantic segmentation overlay using the Mask R-CNN model.

Evaluation of the test results and subsequent optimisation of the model showed its ability to work effectively in practical applications such as urban traffic, different weather conditions, and complex lighting situations.

Model testing. The purpose of the testing was to verify the accuracy of object classification, segmentation, and tracking on different datasets and in conditions close to real-world applications.

One of the most important stages of testing is to validate the model on a test data set, which can be used to determine its ability to generalise the results to new data that is not available in the training process. Testing new, unseen data, such as video and images from real-world surveillance cameras, was used to assess the model's robustness to variability in external conditions. Evaluation of the model's performance on the test set showed similar results to the validation data set (Figure 2).

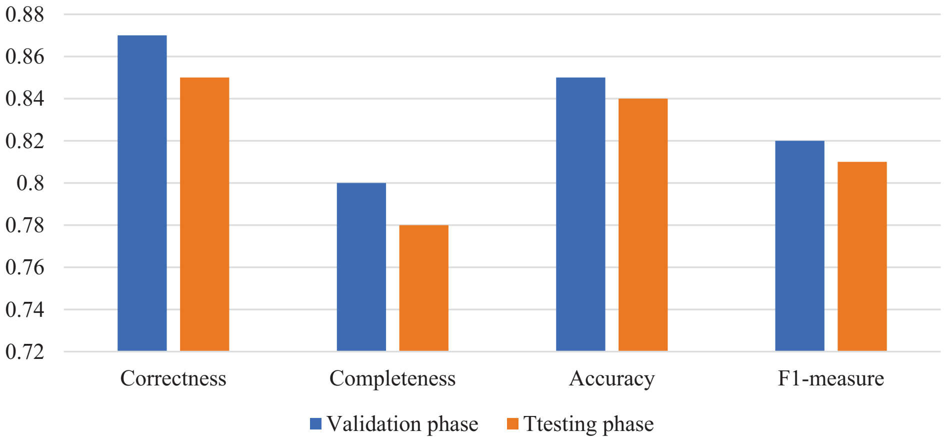

Figure 2. Evaluation of model efficiency at the validation and test stages.

As a result, model validation on a test dataset confirmed the high accuracy and reliability of the model, even under changing conditions. The image processing speed meets the requirements of real-world applications, making the model suitable for use in video surveillance systems.

An additional aspect of the testing was stress testing of the model, which was used to assess its stability and efficiency under extreme conditions such as poor lighting, rain, or fog. The key results of the model's testing and stress testing are summarised in Table 1.

Table 1: Results of testing and stress testing of the model.

Key test criteria	Result
Classification accuracy	85%
Average IoU value	0.73
Frame processing time	45 ms
Additional stress testing criteria	**Result**
IoU for pedestrians	0.65
IoU for vehicles	0.7

After testing on the validation and test datasets, additional validation was performed on various public datasets. This was used to assess the model's versatility and ability to adapt to new conditions. Testing on such datasets as the AI City Challenge Dataset, KITTI, and Mapillary Vistas Dataset showed stable results, which confirms the high level of generalisation (Table 2).

Table 2: Testing on different datasets.

Data set	Criteria	Result
AI City Challenge	classification accuracy	86%
Average IoU value	IoU	0.78
Frame processing time	F1 measure	0.80

As part of the study, the effectiveness of the model was also evaluated using pixel-level accuracy metrics (PA). The research process involved collecting many images for training and testing the model. These included various scenarios where pixel classification was critical. The images were labelled manually, which was used to create a test data set for further validation of the model. After the training phase was completed, the model was tested on a test data set that was not used in the training process. Pixel accuracy (PA) was calculated for each image to evaluate the model's performance in practical applications. The results of the study showed that the accuracy varied depending on the complexity of the images and the nature of the objects being segmented (Table 3).

Table 3: Average PA accuracy values for different image types.

Image type	Pixel-level accuracy (PA)
Simple image (clear contours)	97%
Complex image (soft transitions)	88%
Objects with high variability (noise)	85%

During the evaluation process, the validation dataset achieved a classification accuracy of 87%, which is a high result for this type of task, especially when various variables such as lighting, weather conditions or scene complexity are present. The completeness for the main classes, such as cars and pedestrians, exceeded 80%, which demonstrates the model's ability to detect objects effectively even in difficult conditions. The accuracy of 85% confirms the high level of object detection accuracy among all objects classified as a particular class.

The study confirmed the high accuracy of the model in pixel classification, especially for clearly defined objects. The model demonstrates resilience to changing environmental conditions, effectively detecting key objects even in complex scenes, making it suitable for practical applications.

Segmentation evaluation shows a high level of accuracy in detecting object boundaries in images (Table 4).

Table 4: Segmentation results.

Key test criteria	Result
Average IoU value for the main classes	0.75
Mean IoU for all classes	0.72
Dice Coefficient for the main classes	0.76

Segmentation evaluation confirmed the high accuracy of the model in detecting object boundaries, even in complex urban environments with multiple elements. Segmentation remains stable and accurate, including scenes with partial overlap of objects, which confirms the reliability of the model.

These results highlight the high performance of the model, its ability to generalise to new data and its robustness to complex environmental conditions, which is critical to ensuring effective traffic monitoring in practical applications.

After initial testing of the model and evaluation of its performance in practical conditions, several shortcomings were identified, including problems with the detection of small objects such as bicycles or motorcycles. This limitation can be attributed to the specific architecture of the model, which proved to be less effective at detecting small objects due to insufficient contextual information, which provides more powerful models for larger objects such as cars or pedestrians. This is a common problem for models built on deep CNNs, which can struggle to process objects with few pixels, affecting their ability to detect them accurately.

The main reason for the problem with detecting small objects is insufficient resolution at the level of the lower layers of the network, where the model often fails to preserve important details for accurate representation of small objects. This became evident during the analysis of the results at the testing stage, where images with small objects had low classification and segmentation accuracy, which led to frequent misses of classes such as bicycles and motorcycles in difficult real-time conditions.

To address the identified issues, several optimisation approaches were selected, aimed at improving the detection accuracy of small objects. One of the most effective methods is to modify the neural network architecture to improve its ability to recognise small details. This is achieved by integrating additional layers that allow the model to process high-detail images more efficiently at different levels. Additional convolution layers were added to preserve more spatial information about small objects, and the pooling layer was modified to provide greater accuracy for smaller objects.

The idea of adding additional layers to the model is based on the principles of multi-level neural network architectures, where each subsequent layer allows the network to learn increasingly complex features of images. In traditional CNN architectures, reduction of image size through pooling layers can cause loss of important detailed information, which has a particularly negative effect on small objects (Ahmadov, 2024). Increasing the number of convolution layers significantly improves the model's ability to retain context in many small image elements, which is important for the effective segmentation of small objects.

In parallel with the addition of layers, the activation function was optimised to provide better non-linearity between layers, allowing the network to better account for more complex pixel-to-pixel relationships. This contributed to improved segmentation quality for small objects such as bicycles or motorcycles, as well as improved overall model performance on the test set.

To improve the segmentation of small objects, the architecture was also adapted by using a multi-scale convolution approach. This method involves the use of different convolution filters at different levels of image size, which improves the efficiency of processing both large and small objects within the same image. As a result, more contextual information is preserved, which helps to detect objects at different scales more accurately.

More sophisticated pooling strategies have also been used to preserve more spatial information even after image reduction. This includes adaptive pooling methods that dynamically select the most effective filters based on the characteristics of the input image.

After making changes to the architecture and model settings, the model was retested, which showed a significant improvement in the accuracy of small object segmentation. The results showed an increase in accuracy of 5% for classes such as bicycles and motorcycles. This demonstrates the effectiveness of the chosen optimisation strategy and the increased ability of the model to accurately detect objects even in difficult conditions such as poor lighting or complex backgrounds.

In addition, by introducing additional layers and changes to the activation function, the model has also become more stable when processing complex scenes that include both small and large objects, reducing the number of errors associated with missing small classes.

In general, the optimisation process has significantly improved the results of small object detection and segmentation, which is critical for the practical application of the model in urban environments and automated video surveillance systems.

Final testing and integration. At the final stage of the research, final testing and integration of the model into a real video surveillance system were carried out. This was an important step in verifying its ability to operate stably in real-time, as well as its readiness for implementation in large-scale urban infrastructures for automated monitoring of traffic and public spaces. The testing aimed to evaluate the model's ability to work effectively when integrated with real surveillance cameras, as well as to determine its performance in complex, unpredictable real-time conditions.

The integration of the model into a real-world video surveillance system required considering several factors, such as compatibility with existing infrastructure, data processing time requirements and the need for stable operation with large video streams. One of the key aspects was an adaptive approach to real-time, which meant that it was necessary to maintain minimal latency in processing frames without losing accuracy. Therefore, as part of the integration, the model was optimised to run on high-performance servers using parallel data processing, which significantly reduced the latency of classification and segmentation tasks.

To ensure the stable operation of the model in real-time, a mechanism for continuous monitoring of its performance was integrated, as well as debugging the system concerning the available hardware resources. This included conducting tests with real video streams from surveillance cameras, which determined the level of classification accuracy, video stream processing speed and error rate in real-world use of the system.

After integration into a real system, the model demonstrated stable performance in continuous real-time video processing. This is confirmed by the high average IoU value (0.73) on the test data, as well as the consistent classification accuracy (85%) under different lighting conditions and weather changes. Additionally, the F1-measure was 0.82, which indicates a good balance between accuracy and completeness of object detection. These indicators confirm the high level of efficiency of the model in terms of constant adaptation to changes in the video stream and difficult lighting conditions.

One of the most significant indicators assessed as a result of integration is IoU, which reached a level of 0.7 for the main classes. This result is a good indicator of segmentation efficiency, as it indicates the high accuracy of the model in separating objects and background environments in different images, which in turn improves its ability to accurately detect and segment important elements in practical applications.

Integrating the model into a real-world video surveillance system showed it has a lot of potential for automated traffic and public space monitoring. It successfully spots traffic violations, recognises pedestrians and vehicles even in tough weather conditions, and adapts to different urban scenarios. Stable results on test datasets confirm its suitability for integration into video surveillance systems, opening opportunities for improving the safety and efficiency of public space monitoring. An important result of this work is the possibility of using the model to improve road safety, for detecting traffic violations such as pedestrians crossing on a red light, vehicles driving against the direction of traffic, or other potentially dangerous situations.

4. DISCUSSION AND CONCLUSION

The model demonstrated high efficiency in object recognition and semantic segmentation in an urban environment, achieving average IoU values of 0.73 for the test data. Despite the high classification and segmentation accuracy, the method has several weaknesses that require attention. One of the main disadvantages is a decrease in accuracy (up to 0.65 IoU) when working in low light conditions when processing complex scenes with high noise levels, low-contrast objects, uneven lighting, or when identifying small objects such as bicycles. In such conditions, the model may have difficulty accurately detecting object boundaries, which leads to increased classification errors. To improve the processing of low-contrast objects, pre-trained neural networks, as well as image enhancement methods, such as contrast enhancement and noise reduction algorithms, can be used (Bisenovna et al., 2024; Smailov et al., 2025). In addition, the use of ensemble methods, which combine multiple models to obtain more robust predictions, can improve segmentation accuracy in complex environments. Schneider (2020) considers the use of pre-trained networks to improve the quality of processing complex scenes. The study addressed methods of contrast enhancement and noise reduction, as well as the use of pre-trained models to improve segmentation in difficult conditions. S. Schneider suggests using transfer learning to adapt models to new conditions and data types, which reduces the dependence on many manually labelled images.

In addition, the method depends on a large amount of manually labelled data for training. Errors in labelling or inconsistencies in different datasets can reduce the generalisability of the model and reduce its effectiveness in practical applications (Makhazhanova et al., 2024; Li et al., 2024). One of the ways to reduce the dependence on manually labelled data is to use unsupervised and unsupervised learning methods. This will reduce the need for large amounts of labelling and increase the overall generalisation ability of the model. In addition, the use of generative models (e.g., GAN) to create synthetic data will help diversify the training set and improve its representativeness.

Another issue is image processing time, which can be critical in applications requiring fast response times, such as real-time video surveillance systems. Although current processing speeds meet most requirements, complex scenes with many objects can increase computation time, which must be considered when implementing the system. To improve image processing speed, hardware accelerators such as graphics processing units (GPUs) or tensor processing units (TPUs) can be used, which will significantly speed up calculations (Costanzo et al., 2024). It is also necessary to optimise the model architecture by reducing the number of parameters without compromising segmentation quality. The implementation of quantisation and model compression methods will help improve performance without compromising accuracy (Tkachenko et al., 2025; Rubino et al., 2021). Fülöp and Horváth (2022) described methods for quantising and compressing neural networks to optimise processing speed without significant loss of accuracy. This accelerates the model for use in automated systems where processing speed is important, such as video surveillance systems.

The present study analysed photo and video data of the urban environment and modified neural network architectures were used to achieve high segmentation accuracy, which is consistent with the findings of Statkevich (2024) on the effectiveness of U-Net modifications. Pakhomov et al. (2021) proposed a method for automatic segmentation of images into semantically significant areas without supervision using pre-trained StyleGAN2 and CLIP models. The study demonstrated that clustering in the feature space of the generative model can be used to identify semantic classes without manual labelling. The present study differs in that it uses supervised methods, but the idea of using generative models for segmentation may be promising for future research. In turn, Ho et al. (2020) proposed diffusion models that form a new approach to image generation by iterative noise removal. The method was highly praised for its ability to generate complex images with a high level of detail. Furthermore, the study by J. Ho et al. showed that diffusion models can be used to generate high-quality video and other types of data.

Csurka et al. (2022) analysed the twenty-year development of semantic image segmentation, from early methods to modern approaches, including the use of transformers. The study also discussed low-supervision and self-learning methods to improve segmentation. This study also follows current trends by using deep neural networks to achieve high segmentation accuracy. Kar et al. (2021) highlighted that traditional segmentation methods, such as clustering or thresholding, show limited effectiveness for processing large sets of complex images, especially in urban environments. Instead, modern neural network architectures, such as U-Net or DeepLab, show significantly better results. The present study confirmed these findings: DeepLabv3+ provided a segmentation accuracy of 0.73 IoU, which confirms its effectiveness in complex urban environments.

Giunchiglia et al. (2021) theoretically investigated the problem of forming concepts representing visually perceived objects and proposed a theory and algorithm for constructing such concepts. The present study addressed the practical aspects of segmentation but considered theoretical foundations. Brust and Denzler (2019) investigated the relationship between visual and semantic similarity, emphasising that semantic information can complement or replace missing visual data. The study demonstrated that semantic similarity correlates with visual similarity, which is important for semantic segmentation tasks. The present study supports this idea by using semantic features as a basis for improving segmentation accuracy, demonstrating the relevance of this approach.

Kritskiy et al. (2023) considered the development of software for data segmentation, including images, using merging and splitting methods. The study emphasised the importance of homogeneity of each image region for segmentation accuracy. In the present study, segment homogeneity was also ensured, but deep neural networks were used, which allowed for higher accuracy in tasks with heterogeneous input data. Sakshi and Kukreja (2022) developed a decision support system for urban environment monitoring, which achieved a gain in building segmentation accuracy on digital satellite and aerial images of up to 3%, which is consistent with the approaches used in this study. The study obtained a similar accuracy result (up to 0.78 IoU), but vision transformers (ViT) were used to improve performance.

Butko et al. (2024) examined the use of convolutional neural networks for the semantic segmentation of plant species, which made it possible to estimate vegetation areas. In the study, the use of the Cityscapes and ADE20K datasets enabled high performance in urban environments. Although the present study addressed urban environments, the use of convolutional neural networks for semantic segmentation is a common approach, confirming the versatility of this method for different types of images. In this study, similar datasets were used to train the presented models, and they provided stable performance, especially in object recognition tasks in a transportation environment.

Yang and Yu (2021) highlighted the successful application of CNNs for object detection and semantic segmentation in medical images. In this study, CNNs, the DeepLabv3+ and U-Net models, were used to segment images and video from surveillance cameras in an urban environment, which confirmed the effectiveness of this architecture for solving problems requiring accurate object extraction in images. Qureshi et al. (2022) addressed deep segmentation methods for medical images, including the use of a semantic approach. The correlation between the results in the medical field and the urban environment emphasises the potential of deep learning in various fields. The correlation between medical and urban data indicates the versatility of neural networks in recognising and segmenting objects regardless of the type of image.

Jiang et al. (2022) considered the application of deep learning technologies for analysing remote sensing images. The present study has common aspects, as it also uses deep methods for image analysis for the segmentation of urban scenes. The similarity in approaches to analysing images from different sources, such as satellite and surveillance cameras, demonstrates the versatility of deep learning methods for image analysis in different contexts. Guo et al. (2017) considered various methods of semantic segmentation using deep neural networks. In turn, Lateef and Ruichek (2019) emphasised the importance of segmentation accuracy for various applications, which is also an important aspect of the present study. The present study, using advanced segmentation methods, confirms their effectiveness for image processing in complex urban environments, which correlates with the author's conclusions about the importance of technology development to improve the accuracy and efficiency of image analysis.

Patel et al. (2022) presented a modified UNet architecture for automatic extraction of the road network from satellite images, which improved the accuracy of road detection. The present study also employed deep neural networks to segment urban scenes, but with a focus on detecting not only roads but also other urban objects such as buildings, vehicles, and pedestrians. The results demonstrate that the use of more flexible architectures, such as DeepLabv3+, can improve segmentation accuracy in complex urban landscapes. Cordts et al. (2016) developed the Cityscapes dataset, which has become a standard for analysing urban scenes and testing segmentation algorithms. The study addressed the creation of high-quality image markup for model training, which significantly improved the accuracy of segmentation algorithms in urban environments. In the present study, Cityscapes are also used as one of the basic data sources, but models were additionally tested on more heterogeneous images with different lighting conditions and weather factors. This evaluated the stability of the segmentation models in real-world scenarios, confirming their effectiveness even in difficult conditions. In turn, Kampffmeyer et al. (2016) studied the segmentation of small objects on urban satellite images using convolutional neural networks and uncertainty estimation models. The present study also uses deep learning methods to analyse urban scenes, but focuses on improving the accuracy of the segmentation of complex dynamic objects. This confirms the versatility of deep neural networks for various semantic segmentation tasks and their effectiveness in complex urban environments.

Ulku and Akagunduz (2022) analysed state-of-the-art deep learning architectures for the semantic segmentation of 2D images. The study classified the approaches into three stages: the pre-and early deep learning era, the fully convolutional network (FCN) era, and the post-FCN era. The present study, which uses modern deep learning methods, follows the trends described in the post-FCN era, with a focus on improving localisation and scale invariance. Zhu et al. (2015), Akylbekov et al. (2022) and Sultan et al. (2023) reviewed and described in detail a wide range of image segmentation methods, from classical approaches to modern semantic segmentation and co-segmentation methods. The study emphasised the importance of combining low-level and high-level features to achieve accurate results.

The present study supports this idea by using deep neural networks to combine different levels of features to improve segmentation accuracy.

The analysis demonstrated that the present research correlates with current scientific trends in the use of modern deep learning architectures, such as DeepLabv3+, PyTorch, and Vision Transformers. Other studies confirm the effectiveness of these approaches in complex urban environments. Thus, this study harmoniously integrates into the modern scientific context and contributes to the improvement of semantic image segmentation, offering effective solutions for analysing the urban environment. At the same time, the comparison emphasises the importance of adapting new ideas and technologies for the further development of this field.

The study demonstrated the high efficiency of using deep neural networks for semantic image segmentation in urban environments, for automated traffic monitoring and security. The system, based on the DeepLabv3+ and U-Net architectures, demonstrated segmentation accuracy with an IoU of 0.73 on the Cityscapes and ADE20K test datasets. This confirms the models' ability to effectively segment objects in urban environments with a high level of detail. However, in low light and difficult weather conditions, the model's accuracy dropped to 0.65 IoU, which highlights the need for further research to improve the model's robustness to such environmental changes.

Practical tasks that can be solved using the proposed method include automated traffic monitoring, where models can identify vehicles, pedestrians, and road signs, helping to analyse traffic flows and prevent congestion. Ensuring safety in urban environments is also important, where segmentation helps identify potentially dangerous situations, such as pedestrians on the road or abnormal vehicle behaviour (Piera et al., 2016; Oklander et al., 2019). The system can also be useful for intelligent transport systems, such as adaptive traffic light control and traffic management based on video stream analysis. Other important applications include environmental monitoring to identify green areas and assess vegetation health, as well as infrastructure analysis, including the condition of roads, pavements, and buildings, which will assist in urban planning and repairs.

One of the important results of the study is the identification of the impact of complex conditions on segmentation accuracy and model performance. To achieve high segmentation accuracy with minimal time (40 ms per frame), it was important to optimise the model architecture for fast real-time processing. This highlights the importance of developing lightweight models that can run on peripheral devices or with limited resources, which creates new horizons for integrating the technology into real-world monitoring systems. In addition, the study results prove that the use of architectures based on deep neural networks enables efficient processing of complex images in real-time, maintaining high segmentation accuracy even with a large variety of objects.

The innovation of this study is the adaptation of existing segmentation methods to the specific conditions of an urban environment, where many objects, diverse weather conditions, and rapid changes in illumination significantly affect the accuracy of the models. The use of architectures such as DeepLabv3+ can achieve high performance even in such challenging conditions, which is significantly different from traditional methods used for image classification. At the same time, the challenges posed by lighting and weather conditions open prospects for further research aimed at improving the stability and adaptability of models to such changes.

Recommendations for further optimisation of the model include improving the algorithms for low-light conditions and difficult weather situations, which can be achieved using the Vision Transformers (ViT) method. This method can significantly improve the accuracy of object segmentation in complex scenes with overlapping objects, improving the detection of small objects in changing conditions. It is also recommended to use synthetic data to improve training, for example, using generative models (GANs), which can create additional training sets, expand the range of practical situations, and improve the model's adaptation to various conditions. In addition, it is important to focus on developing lightweight models that will run efficiently on devices with limited computing resources, which will ensure fast and accurate data processing, including on mobile platforms.

Future research should aim to develop methods that will improve segmentation performance in environments with limited computing resources, such as mobile devices. It is also important to develop models that can adapt to changes in the urban environment, which can be achieved by using synthetic or augmented datasets, solving problems related to data markup and minimising training costs, improving the adaptability and versatility of models in practical applications.

5. REFERENCES

Ahmadov, S., 2024, Data encryption as a method of protecting personal data in a cloud environment. *Bull. Cherkasy State Tech. Univ.* **29**(3), 31-41.

Akylbekov, O., Said, N.A., Martínez-García, R., Gura, D., 2022, ML models and neural networks for analyzing 3D data spatial planning tasks: Example of Khasansky urban district of the Russian Federation. *Adv. Eng. Softw.* **173**, 103251.

Bisenovna, K.A., Ashatuly, S.A., Beibutovna, L.Z., Yesilbayuly, K.S., Zagievna, A.A., Galymbekovna, M.Z., Oralkhanuly, O.B., 2024, Improving the efficiency of food supplies for a trading company based on an artificial neural network. *Int. J. Elect. Comput. Eng.* **14**(4), 4407-4417.

Brust, C.-A., Denzler, J., 2019, Not just a matter of semantics: The relationship between visual and semantic similarity. In: G.A. Fink, S. Frintrop, X. Jiang (Eds.), *Proc. 41st DAGM German Conf. Pattern Recognition*, 414-427. Springer, Cham.

Butko, I.M., Golubenko, O.I., Makoveichuk, O.M., 2024, Semantic segmentation in multispectral imagery. *ITSynergy* **1**, 16-29.

Cao, F., Bao, Q., 2020, A survey on image semantic segmentation methods with convolutional neural network. In: *2020 Int. Conf. Commun. Inf. Syst. Comput. Eng. (CISCE)*, 458-462. IEEE, Kuala Lumpur.

Chornyy, S., Brendel, O., Gratiashvili, D., 2022, Authenticate images based on their semantic segmentation in deep learning neural networks with their pre-processing with use of filtering methods. *Theory Pract. Forensic Sci. Criminalistics* **26**(1), 125-137.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016, The cityscapes dataset for semantic urban scene understanding. In: *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 3213-3223. IEEE, Las Vegas.

Costanzo, L., Rubino, G., Rubino, L., Vitelli, M., 2024, PFC Control Signal Driven MPPT Technique for Grid-Connected PV Systems. *IEEE Transact. Power Elect.* **39**(8), 10368-10379. 10.1109/TPEL.2024.3393294

Csurka, G., Volpi, R., Chidlovskii, B., 2022, Semantic image segmentation: Two decades of research. *Found. Trends Comput. Graph. Vis.* **14**(1-2), 1-162.

Du, S., Du, S., Liu, B., Zhang, X., 2020, Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. *Int. J. Digit. Earth* **14**(3), 357-378.

Fülöp, A., Horváth, A., 2022, End-to-end training of deep neural networks in the fourier domain. *Mathematics* **10**(12), 2132.

Giunchiglia, F., Erculiani, L., Passerini, A., 2021, Towards visual semantics. *SN Comput. Sci.* 2, 446.

Guo, Y., Liu, Y., Georgiou, T., Lew, M.S., 2017, A review of semantic segmentation using deep neural networks. *Int. J. Multimed. Inf. Retr.* **7**(2), 87-93.

Ho, J., Jain, A., Abbeel, P., 2020, Denoising diffusion probabilistic models. In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), *34th Conf. Neural Inf. Process. Syst. (NeurIPS 2020)*, 1-12. Vancouver Convention Centre, Vancouver.

Jiang, B., An, X., Xu, S., Chen, Z., 2022, Intelligent image semantic segmentation: A review through deep learning techniques for remote sensing image analysis. *J. Indian Soc. Remote Sens.* **51**(9), 1865-1878.

Kampffmeyer, M., Salberg, A.B., Jenssen, R., 2016, Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: *2016 IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 1-9. IEEE, Las Vegas.

Kar, M.K., Nath, M.K., Neog, D.R., 2021, A review on progress in semantic image segmentation and its application to medical images. *SN Comput. Sci.* **2**, 297.

Khan, M.Z., Gajendran, M.K., Lee, Y., Khan, M.A., 2021, Deep neural architectures for medical image semantic segmentation: Review. *IEEE Access* **9**, 83002-83024.

Kritskiy, D., Shkurenko, N., Popov, O., Kravtsova, O., 2023, Development of software for data segmentation by photo and video information. *Aerosp. Tech. Technol.* **3**, 61-75.

Lateef, F., Ruichek, Y., 2019, Survey on semantic segmentation using deep learning techniques. *Neurocomputing* **338**, 321-348.

Li, X., Anukul, T., Ying, F., 2024, Cross-platform adaptation of algorithmic editing techniques. *Bull. Cherkasy State Tech. Univ*. **29**(2), 45-56.

Makhazhanova, U., Omurtayeva, A., Kerimkhulle, S., Tokhmetov, A., Adalbek, A., Taberkhan, R., 2024, Assessment of Investment Attractiveness of Small Enterprises in Agriculture Based on Fuzzy Logic. *Lect. Not. Networks Syst.* **935**, 411-419.

Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D., 2021, Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(7), 3523-3542.

Muhammad, K., Hussain, T., Ullah, H., Del Ser, J., Rezaei, M., Kumar, N., Hijji, M., Bellavista, P., De Albuquerque, V.H.C., 2022, Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Trans. Intell. Transp. Syst.* **23**(12), 22694-22715.

Oklander, M., Yashkina, O., Yashkin, D., 2019, Minimization of transportation risks in logistics by choosing a cargo delivery route with the minimal projected number of road accidents. *East. Eur. J. Enter. Tech.* **5**(3-101), 57-69.

Pakhomov, D., Hira, S., Wagle, N., Green, K.E., Navab, N., 2021, Segmentation in style: Unsupervised semantic image segmentation with Stylegan and CLIP. arXiv:2107.12518.

Patel, M.J., Kothari, A.M., Koringa, H.P., 2022, A novel approach for semantic segmentation of automatic road network extractions from remote sensing images by modified UNet. *Radioelectron. Comput. Syst.* **3**, 161-173.

Pemasiri, A., Nguyen, K., Sridharan, S., Fookes, C., 2020, Multi-modal semantic image segmentation. *Comput. Vis. Image Underst.* **202**, 103085.

Piera, M.A., Buil, R., Ginters, E., 2016, State space analysis for model plausibility validation in multi-agent system simulation of urban policies. *J. Simul.* **10**(3), 216-226.

Pohudina, O., Kritskiy, D., Bykov, A.N., Szalay, T., 2020, Method for identifying and counting objects. In: M. Nechyporuk, V. Pavlikov, D. Kritskiy (Eds.), *Integrated Computer Technologies in Mechanical Engineering: Synergetic Engineering*, 161-172. Springer, Cham.

Qureshi, I., Yan, J., Abbas, Q., Shaheed, K., Riaz, A.B., Wahid, A., Khan, M.W.J., Szczuko, P., 2022, Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Inf. Fusion* **90**, 316-352.

Rakhmetulayeva, S., Marat, A., Iliev, T., Mukasheva, A., 2023, Analysis of the impact of video quality on feature extraction from a video stream using convolutional neural networks. *Bull. Almaty Univ. Energy Commun.* **1**, 99-110.

Rubino, L., Rubino, G., Conti, P., 2021, Design of a power system supervisory control with linear optimization for electrical load management in an aircraft on-board dc microgrid. *Sustainab Switz.* **13**(15), 8580.

Sakshi, N., Kukreja, V., 2022, Image segmentation techniques: Statistical, comprehensive, semi-automated analysis and an application perspective analysis of mathematical expressions. *Arch. Comput. Methods Eng.* **30**(1), 457-495.

Schneider, S., 2020, *Deep learning based computer vision for animal re-identification*. University of Guelph, Guelph.

Sevak, J.S., Kapadia, A.D., Chavda, J.B., Shah, A., Rahevar, M., 2017, Survey on semantic image segmentation techniques. In: *2017 Int. Conf. Intell. Sustain. Syst. (ICISS)*, 306-313. IEEE, Palladam.

Smailov, N., Kadyrova, R., Abdulina, K., Uralova, F., Kubanova, N., Sabibolda, A., 2025, Application of facial recognition technologies for enhancing control in information security systems. *Inf. Automat. Pom. Gosp. Ochron. Srod.* **15**(3), 55-58.

Statkevich, R.V., 2024, *Image segmentation method using deep neural networks*. Igor Sikorsky Kyiv Polytechnic Institute, Kyiv.

Sultan, D., Mendes, M., Kassenkhan, A., Akylbekov, O., 2023, Hybrid CNN-LSTM network for cyberbullying detection on social networks using textual contents. *Int. J. Adv. Comput. Sci. Appl.* **14**(9), 748-756.

Tkachenko, O., Chechet, A., Chernykh, M., Bunas, S., Jatkiewicz, P., 2025, Scalable Front-End Architecture: Building for Growth and Sustainability. *Inf Slov*. **49**(1), 137-150.

Ulku, I., Akagündüz, E., 2022, A survey on deep learning-based architectures for semantic segmentation on 2D images. *Appl. Artif. Intell.* **36**(1), 2032924.

Yang, R., Yu, Y., 2021, Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front. Oncol.* **11**, 638182.

Zaitoun, N.M., Aqel, M.J., 2015, Survey on image segmentation techniques. *Procedia Comput. Sci.* **65**, 797-806.

Zhu, H., Meng, F., Cai, J., Lu, S., 2015, Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *J. Vis. Commun. Image Represent.* **34**, 12-27.