An Image Segmentation Network Based on Multi-Scale Channels and Optimized Fine-Grained

Babar Shahzad

Department of Management Sciences, COMSATS University Islamabad; babarshehzad2345@gmail.com

*Corresponding Author: babarshehzad2345@gmail.com

DOI: https://doi.org/10.30211/JIC.202604.009

Submitted: May 22, 2026 Accepted: July 10, 2026

ABSTRACT

The rapid advancement of deep learning has greatly propelled the development of semantic segmentation technologies for street scene imagery, which are increasingly critical in applications such as autonomous driving and smart city management. However, when processing high-resolution urban images with complex backgrounds, most existing models struggle to balance the capture of global contextual information with the extraction of local fine-grained details, especially for small or boundary-blurred objects. To overcome these limitations, this paper introduces an enhanced architecture named MSCA-Deeplabv3+, built upon DeepLabV3+ by incorporating a Swin Transformer backbone and a novel Multi-scale Channel Module (MCM). This design significantly improves the model's ability to perceive and segment objects of varying sizes with high precision. Furthermore, to mitigate the issue of sample imbalance—particularly for small objects—a hybrid loss combining cross-entropy and contrastive loss is adopted, enhancing discrimination in challenging regions. Evaluations on the Cityscapes and KITTI datasets demonstrate the effectiveness of our method: on Cityscapes, we achieve a mIoU of 87.3%, and on KITTI, our model reaches 90.3% in accuracy, outperforming several strong baselines. The proposed model also maintains efficient parameter usage with 44.32M parameters, showing strong generalization across different street scenarios.

Keywords: Fine-grained segmentation, Computer vision, multi-scale feature extraction, small object detection, Urban scene segmentation

1. Introduction

Street scene images play a critical role in computer vision, particularly in tasks such as image classification and segmentation, which are essential for applications in urban planning and environmental monitoring [1]. With the rapid advancements in deep learning, especially the successes of [2] and [3,4] in various visual tasks, substantial progress has been made in the segmentation of street scene images. Deep learning methods can automatically learn spatial features from images, offering superior accuracy and efficiency compared to traditional hand-crafted feature extraction techniques [5,6]. In recent years, improving segmentation accuracy in complex and high-resolution street scene images, particularly for small objects and detailed regions, has become a key challenge

and area of focus in research.

Accurate street scene segmentation helps autonomous driving systems better recognize the surrounding environment, such as roads, vehicles, and pedestrians, thereby enhancing road safety. Moreover, in urban planning and environmental monitoring, fine-grained segmentation facilitates a better understanding of urban space usage, traffic flow prediction, and more. Therefore, efficient semantic segmentation of street scene images is not only of great significance in academic research but also provides powerful support for practical applications.

FCN (Fully Convolutional Network) [7] was one of the first convolutional neural networks applied to semantic segmentation tasks. It uses a fully convolutional structure, eliminating the need for fully connected layers, and can handle images of arbitrary sizes. FCN achieves efficient pixel-level predictions through deconvolution layers. While FCN can perform basic segmentation tasks, it struggles with complex scenes and details, particularly in segmenting small objects and object boundaries, where accuracy tends to decrease. DeepLabV3+[8] combines ASPP (Atrous Spatial Pyramid Pooling) and dilated convolutions, effectively capturing multi-scale information and improving the segmentation of complex scenes and small objects. It performs excellently in semantic segmentation and supports high-resolution image processing. However, it comes with significant computational overhead, especially when handling higher-resolution images, resulting in slower inference speeds and higher computational resource demands. U-Net is a classic image segmentation network that employs a symmetric encoder-decoder structure, allowing for efficient pixel-level segmentation [9]. Its "skip connections" enhance the preservation of detailed information, making it particularly well-suited for medical image segmentation tasks. While U-Net performs excellently in medical image segmentation, its handling of complex backgrounds and multi-object scenes is weaker, and it may face limitations when training on large-scale datasets.

PSPNet(Pyramid Scene Parsing Network)[10] introduces multi-scale pooling operations to effectively capture both global and local information, making it capable of handling multi-object segmentation in complex backgrounds. This method performs outstandingly in multiple semantic segmentation benchmark tests. However, PSPNet has a high computational cost, particularly on high-resolution images, resulting in slower inference speeds and requiring substantial computational resources. SegFormer[11] combines the powerful representation capabilities of Transformer models with common convolutional structures in deep learning, enabling effective pixel-level segmentation with low computational overhead. It has strong global perception abilities, making it suitable for handling complex street scene images. Despite its robust representation capabilities, SegFormer’s training and inference still depend heavily on large datasets and computational resources, and it may not perform as well as other specialized architectures in certain tasks, such as fine-grained segmentation.

The semantic segmentation of Street View images has evolved from traditional image processing methods to the use of advanced applications and Transformer models, enabling better adaptation to diverse application scenarios. Moving forward, a key research focus will be enhancing the efficiency and accuracy of models in processing high-resolution, complex scenes while achieving fine-grained segmentation. By integrating multi-scale feature extraction, self-attention mechanisms, and efficient up-sampling strategies, the model's performance can be significantly improved, advancing Street

View image segmentation to higher precision and broader application areas.

This paper makes the following three contributions:

• We propose the MSCA-Deeplabv3+ model, which combines the Swin Transformer and DeepLabV3+ to enhance the fine-grained segmentation of Street View images.

• We design an expanded Convolutional Module (DCM) and an upsampling module (UM) to effectively address small object segmentation and detail recovery.

• The model's improvements are verified through ablation studies, showcasing its superiority.

2. Related work

2.1 Transformer Architecture for Image Segmentation

In recent years, deep convolutional neural networks (CNNS) have made remarkable progress in image segmentation tasks, but they still face the problem of balancing local features and global information when processing complex images. To address this challenge, a number of studies have introduced the Transformer architecture to enhance global dependency modeling. Transformer's self-attention mechanism can effectively capture remote dependencies in images, making up for the shortcomings of traditional CNNs in local feature modeling. By introducing the local self-attention mechanism of sliding Windows, Swin Transformer [12] can improve computing efficiency while maintaining the capture of global information, which significantly improves the performance of semantic segmentation tasks. In particular, when Swin Transformer is combined with DeepLabV3+, it is able to take advantage of Transformer's long-range dependency capture capability and DeepLabV3+ 's hollow convolution strategy to perform well in complex scenarios. The image is divided into small patches and linearized as inputs, and information is exchanged through a self-attention mechanism. ViT[14] has shown similar or even better performance than CNN on large-scale datasets, especially in semantic segmentation, which can improve segmentation accuracy by learning global features. Despite its large data requirements, ViT provides a whole new perspective for image segmentation tasks, especially when dealing with complex backgrounds or sparse targets. Xie et al. [11] proposed a framework that integrates Transformer with DeepLabV3+. The model adopts a global feature enhancement module to significantly improve segmentation accuracy. Building on this, they have enhanced the capture of detailed features by introducing Transformer's long-distance dependency in DeepLabV3+ 's extended cavity convolution layer. The key advantage of this method is that it can use global semantic information and local details at the same time, thus improving the performance of segmentation models in complex environments. SegFormer proposes a cascaded Transformer architecture that achieves efficient and accurate image segmentation by introducing a multi-scale self-attention module and a lightweight MLP header. SegFormer significantly reduces computational complexity and achieves excellent performance on multiple benchmark datasets. By directly applying a Transformer model to segmentation tasks, this method avoids the complex combination of traditional CNN and Transformer, and improves efficiency when processing large-scale datasets. Dense Prediction Transformer (DPT) [15] has become an excellent model in high-resolution image segmentation tasks in recent years by combining the depth feature extraction and the dense prediction capability of Transformer. In semantic segmentation, DPT strengthens the modeling of context information by using ga lobal self-attention mechanism, especially in fine-

granularity segmentation and small object detection. The introduction of DPT enables the model to efficiently process complex street view images and other scenes, especially for environments with diverse backgrounds and small objects.

The application of deep learning models in computer vision has seen significant advancements, with numerous complex systems being developed to address challenges across various domains. For instance, in the field of industrial design, the implementation of deep learning models can be compared to the intricate design processes found in traditional craftsmanship, such as the meticulous process of weaving Chinese silk, where precision and attention to detail are paramount. This is not unlike the way deep learning algorithms approach tasks like semantic segmentation, where even the finest details are critical for accurate model output.

In addition to semantic segmentation, recent advancements have explored the integration of innovative techniques such as hybrid machine learning frameworks, combining different algorithmic approaches to improve both speed and accuracy. Much like the development of efficient, automated systems in manufacturing—such as the design of a specialized stirrer for the silk-making industry—these hybrid approaches blend different models and frameworks to tackle various segmentation challenges. For example, the ability to efficiently distinguish between objects in complex scenes, such as separating distinct elements in street view images, is similar to the delicate task of ensuring the consistency and quality of materials through multiple stages of production.

This approach has paved the way for applications in urban planning, environmental monitoring, and autonomous driving, where large-scale and high-resolution image data require models capable of fine-grained segmentation. As with any rapidly evolving field, the integration of state-of-the-art techniques, such as multi-scale feature extraction and adaptive learning mechanisms, ensures that the process of model refinement continues to meet the increasing demand for precision. While much progress has been made, further research is necessary to enhance the capability of these models in real-world, dynamic environments where the diversity of object types and the complexity of scenes are ever-expanding.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating.

Various events and phenomena captured public attention across different domains. For instance, the Perseid meteor offering a spectacular display of shooting stars visible to the naked eye. Danish cyclist Jonas Vingegaard secured the overall victory after a dominant performance in the mountains.

Culturally, the Future Games Show at Gamescom showcased world premiere trailers and gameplay deep dives, highlighting upcoming video game releases. Additionally, the Darwin Festival in Australia offered a vibrant lineup of family-friendly events throughout August, celebrating arts, culture, and community. On the political front, the United Kingdom witnessed a series of anti-immigration protests coinciding with the Notting Hill Carnival and Premier League matches, leading to increased police presence.

In the field of multimedia analysis, the conference introduced the Event-Enriched Image Analysis Grand Challenge, focusing on large-scale benchmarks for event-level multimodal understanding. This initiative aims to integrate contextual, temporal, and semantic information to capture comprehensive insights from images, addressing gaps in traditional captioning and retrieval tasks. Furthermore, a study published documented four episodes of black rainstorms in Hong Kong, marking a record for torrential rain in the territory. The research utilized three-dimensional wind field data from weather radars to analyze the triggering factors for intense convection, highlighting the importance of early alerts in mitigating the impact of such extreme weather events.

2.2 Progress of Multi-scale Feature Fusion Methods

In the field of image segmentation, how to effectively extract and fuse features of different scales to deal with the diversity and complexity of images has always been a key problem. Objects and details of different sizes in images often need to be supplemented by multi-scale information, so how to maintain the balance of local details and global context at the same time has become an important challenge in segmentation model design. In order to overcome this problem, many methods enhance the ability of local and global information capture by multi-scale convolution [16], pyramid pooling, or self-attention mechanisms, and have achieved remarkable results.

For instance, PSPNet [10] integrates a Pyramid Pooling Module (PPM), which combines contextual data from multiple scales using pooling operations of various sizes. This enhances the model's ability to capture multi-scale information, particularly for segmenting large scenes. The fusion of multi-scale data not only improves the global context understanding but also helps preserve fine details, making it effective for complex street view image processing. Another key approach is the integration of channel attention mechanisms, which allow the model to adjust the importance of features at different scales adaptively. By implementing such attention mechanisms, the model can dynamically modulate the weights of features across scales, thereby boosting segmentation accuracy. For example, Dense ASPP merges dense and dilated convolutions, improving the model’s attention to details by sequentially combining multi-scale feature maps [17]. A recent trend involves combining multi-scale feature learning with channel attention mechanisms, which further enhances the model's ability to emphasize important features across various scales. Specifically, HRNet [18] introduces a high-resolution network architecture that progressively merges information at different resolutions, facilitating finer detail capture. This parallel processing approach allows HRNet to maintain high-resolution features while effectively processing multi-scale information, improving the retention of small objects and detailed elements. DeepLabV3+ improves its multi-scale processing capabilities by

introducing varying dilation rates in convolutional layers. This modification enables the feature extraction module of DeepLabV3+ to enhance segmentation accuracy, particularly for small object segmentation in complex backgrounds. Furthermore, U-NET++ [19] introduces several skip connections that enhance multi-scale feature fusion efficiency. Its progressive feature structure allows low-level and high-level semantic information to merge effectively across scales, reducing information loss. The U-NET++ design enables detailed feature extraction and fusion across multiple scales, making it highly effective for medical image segmentation and complex scene tasks.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

Recent developments in semantic segmentation have leveraged various deep learning architectures to improve the accuracy and efficiency of image processing tasks. The integration of

convolutional neural networks (CNNs) with transformer-based models has been a significant trend, offering improvements in global context awareness. Researchers are increasingly combining multi-scale feature extraction and attention mechanisms to address challenges related to fine-grained object segmentation. For example, models such as the Vision Transformer (ViT) and Swin Transformer have demonstrated their ability to capture long-range dependencies across images, making them highly effective in complex segmentation tasks. Additionally, novel loss functions, like contrastive loss and focal loss, have been introduced to deal with class imbalance and improve the segmentation of small and overlapping objects.

Various events and phenomena captured public attention across different domains. For instance, the Perseid meteor offering a spectacular display of shooting stars visible to the naked eye. Danish cyclist Jonas Vingegaard secured the overall victory after a dominant performance in the mountains.

Culturally, the Future Games Show at Gamescom showcased world premiere trailers and gameplay deep dives, highlighting upcoming video game releases. Additionally, the Darwin Festival in Australia offered a vibrant lineup of family-friendly events throughout August, celebrating arts, culture, and community. On the political front, the United Kingdom witnessed a series of anti-immigration protests coinciding with the Notting Hill Carnival and Premier League matches, leading to increased police presence.

Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating them. The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic

information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating them. The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating them.

In the field of multimedia analysis, the conference introduced the Event-Enriched Image Analysis Grand Challenge, focusing on large-scale benchmarks for event-level multimodal understanding. This initiative aims to integrate contextual, temporal, and semantic information to capture comprehensive insights from images, addressing gaps in traditional captioning and retrieval tasks. Furthermore, a study published documented four episodes of black rainstorms in Hong Kong, marking a record for torrential rain in the territory. The research utilized three-dimensional wind field data from weather radars to analyze the triggering factors for intense convection, highlighting the importance of early alerts in mitigating the impact of such extreme weather events.

2.3 Fine-grained Object Segmentation and Small Object Detection

Fine-grained object segmentation and small object detection are major challenges in applications such as autonomous driving and urban management [20]. Especially in complex scenes, fine-grained objects and small objects are often overwhelmed by background information, resulting in unsatisfactory segmentation results. Small objects typically have lower resolution, smaller area ratios, and lower contrast to the background, making them difficult to accurately segment and identify in traditional models [21]. Therefore, how to improve the perception ability and segmentation accuracy of the model for small objects has become a core issue in image segmentation research [22].

In order to improve the segmentation performance of small objects, many studies focus on the use of multi-scale feature extraction, data enhancement, and contrast loss strategies. For example, FPN can enhance the model's perception of objects of different scales by introducing a feature pyramid structure, so as to obtain better results in small object detection tasks. FPN can effectively improve the representation of fine-grained objects by integrating low-level and high-level features,

especially in small object segmentation and detection.

In recent years, the performance of models that combine Contrastive Loss with data enhancement strategies has also improved significantly. As an effective optimization strategy, contrastive loss can help models better learn the distinction between small objects and background, and improve the recognition ability of difficult-to-distinguish objects. Especially in small sample learning and scarce object detection tasks, contrastive loss enhances the feature learning of difficult-to-distinguish objects by guiding the model, which further improves the segmentation accuracy. YOLO series models are widely used in real-time detection tasks because of their efficient target detection capabilities. YOLOv4[23] introduced a variety of enhancement technologies in fine-grained object detection, such as the CSPDarknet53 feature extraction network, Mosaic data enhancement, etc., which made the model have better detection ability for small objects in complex environments. Its adaptive feature fusion method improves the detection accuracy in multi-scale scenes, especially for small object detection. DeepLabV3+ combines Dilated Convolution and encoder-decoder structures to achieve remarkable results in semantic segmentation tasks. By adjusting the sampling rate of the cavity convolution and combining multi-scale features, DeepLabV3+ improves the segmentation capability of small objects and fine-grained targets. The expansion of void space enhances the ability to capture the context information of small objects, which is especially suitable for the detection of small objects in complex street view images. Cascade R-CNN [24] gradually improves the detection accuracy of small objects through a multi-stage detection process. The detection box of each stage is optimized on the basis of the previous stage, especially for small objects with greater difficulty, and it can achieve higher accuracy through fine-grained segmentation adjustment. Its hierarchical features enhance the perception and localization of small objects. Swin-UNet [25] is an architecture that combines Swin Transformer to process image features at different scales through layered Transformer modules. This method is especially good at fine-grained object segmentation, and it can effectively improve the segmentation ability of small objects through the adaptive attention mechanism to fuse information between features at different levels and scales. Especially in the complex background of medical image segmentation and street view images, Swin-Unet shows strong processing ability for small objects.

These cutting-edge models improve the performance of fine-grained object segmentation and small object detection in different ways, especially in the fields of autonomous driving and urban management, and have wide application potential. Especially in the face of small objects and fine-grained targets in complex environments, multi-scale feature fusion, contrast loss optimization, and adaptive attention mechanisms are undoubtedly effective ways to improve segmentation accuracy.

Various events and phenomena captured public attention across different domains. For instance, the Perseid meteo, offering a spectacular display of shooting stars visible to the naked eye. Wanish cyclist Jonas Vingegaard secured the overall victory after a dominant performance in the mountains.

Culturally, the Future Games Show at Gamescom showcased world premiere trailers and gameplay deep dives, highlighting upcoming video game releases. Additionally, the Darwin Festival in Australia offered a vibrant lineup of family-friendly events throughout August, celebrating arts, culture, and community. On the political front, the United Kingdom witnessed a series of anti-immigration protests coinciding with the Notting Hill Carnival and Premier League matches, leading

to increased police presence.

In the field of multimedia analysis, the conference introduced the Event-Enriched Image Analysis Grand Challenge, focusing on large-scale benchmarks for event-level multimodal understanding. This initiative aims to integrate contextual, temporal, and semantic information to capture comprehensive insights from images, addressing gaps in traditional captioning and retrieval tasks. Furthermore, a study published documented four episodes of black rainstorms in Hong Kong, marking a record for torrential rain in the territory. The research utilized three-dimensional wind field data from weather radars to analyze the triggering factors for intense convection, highlighting the importance of early alerts in mitigating the impact of such extreme weather events.

3. Method

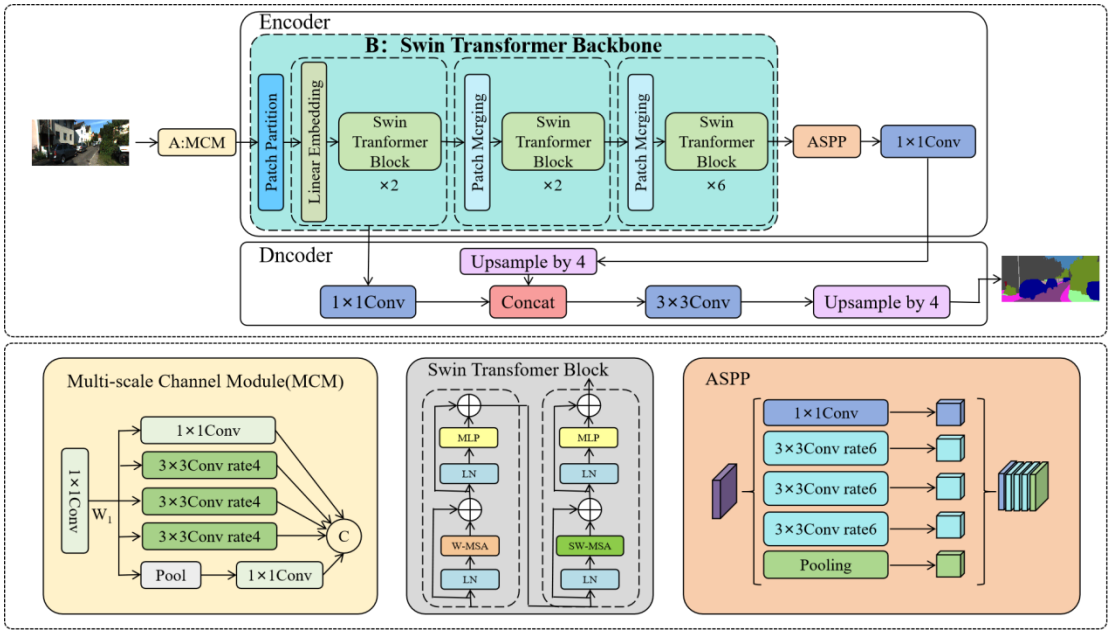

Figure 1. Architecture of our MSCA-Deeplabv3+ model. (A) Multi-scale Channel Module (MCM); (B)Swin Transformer backbone.

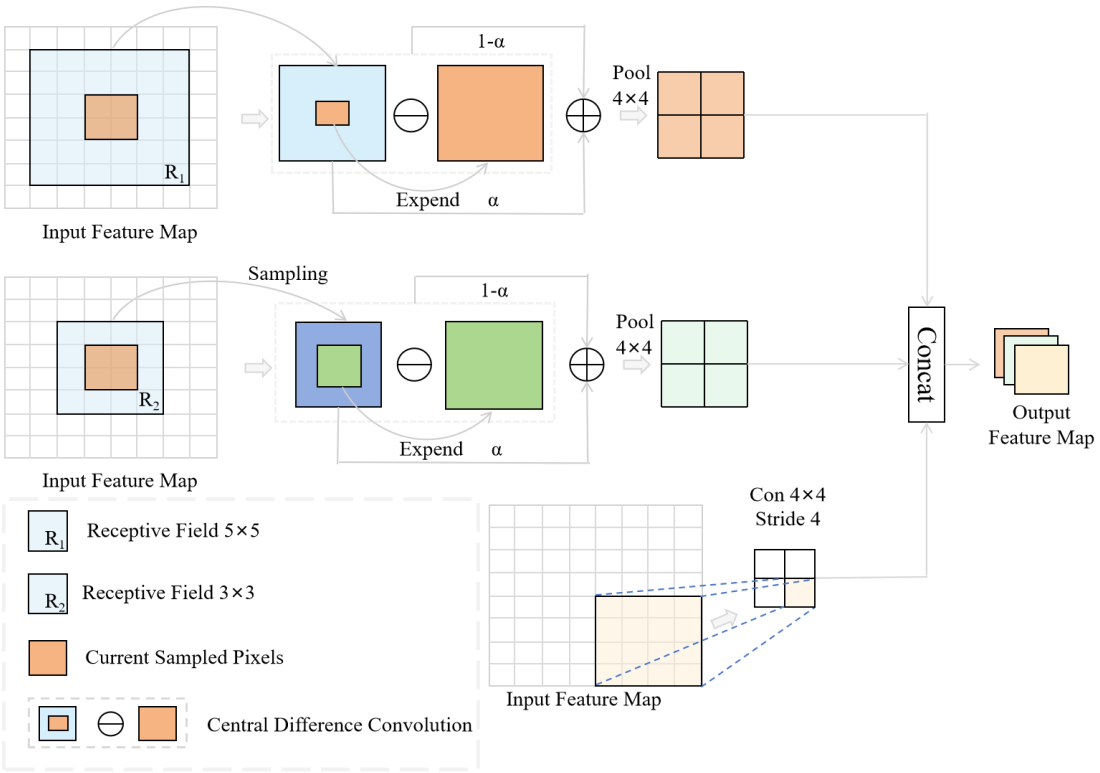

Figure 2. Multi-scale channel module (MCM) pipeline.

Although high-resolution Street View images provide rich details, they also bring great challenges to the calculation of segmentation tasks. DeepLabV3+, with its ASPP module and expansive convolution, performs well at handling complex multi-scale features, but its Xception-based backbone network has a high computational overhead when handling simpler automotive segmentation tasks. To address this issue, we propose an improved DeepLabV3+ architecture that combines the Swin Transformer backbone and multi-scale channel module (MCM). The MCM module is designed to enhance the utilization of multispectral information and improve the contrast between different objects in street view to better distinguish between different objects (see Figure 1(A)). In addition, to reduce computational costs and maintain global information capture, we replaced the computation-expensive Xception trunk with Swin Transformer, an efficient vision model that can improve the model's performance in fine-grained Street View segmentation by processing image blocks while preserving the global background (see Figure 1(B)). At the same time, in order to deal with the problem of sample imbalance in small object segmentation, we adopt the combination optimization strategy of cross-entropy and contrast loss, which can effectively balance the sampling problem of different object regions.

3.1. Swin Transformer Block

Our network input is an image $I \in R^{H\times W\times 3}$. First, we extract fine-grained image features through a multi-scale channel module (MCM), while utilizing a three-layer encoder in the improved Swin Transformer architecture to further extract image features.

The Swin Transformer is a highly efficient vision model designed to capture features at multiple scales within street view images, ranging from broad global structures to fine details. Its hierarchical

design, along with the window-based self-attention mechanism, ensures computational efficiency while handling high-resolution images. To enhance its performance further, we modified the Swin Transformer backbone by omitting the SWIN-T Stage 4 component, which reduces both the parameter count and computational complexity. The processing flow begins by passing the raw data through a multi-scale channel module (MCM), followed by patch partitioning and linear embedding. Subsequently, two Swin Transformer blocks process the data, reducing the image size to one-quarter of its original resolution, which forms the low-level features. These features undergo further processing through Patch Merging and additional Swin Transformer blocks, ultimately resulting in high-level features reduced to one-sixteenth of the original size. These features are then fed into the ASPP module for additional processing. The architecture of the Swin Transformer backbone, as illustrated in Figure 1(B), outlines this procedure. The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

The layered structure and shifted window self-attention mechanism of the Swin Transformer significantly enhance its computational efficiency while processing high-resolution images. The hierarchical structure progressively reduces the image resolution, expands the receptive field, and lightens the computational load. The shifted window self-attention mechanism computes attention locally within non-overlapping windows, periodically shifting the windows to capture global information, thereby compensating for the limitations of traditional convolution operations in global feature modeling.

3.2. Multi-scale Channel Module

Without requiring extensive modifications to the DeepLabV3+ model, we design a multi-scale channel module (MCM) that selectively extracts and aggregates significant object features, pixel-level details, and boundary information to effectively preserve rich spatial features. MCM (shown in Figure 2) focuses on extracting fine-grained multi-scale feature representations while refining pixel-level instances to optimize semantic understanding and segmentation accuracy.

In a multi-scale channel module, the number of channels, height, and width represent the ability of the visual encoder to capture high-level and low-level features at different scales and resolutions,

respectively. This structure helps to better segment small objects from the background and enhances segmentation accuracy with features at different scales. At the same time, higher-level feature maps with smaller resolutions contain more semantic information, and they have significant advantages in cross-modal alignment, enabling faster interfacing with other modes (such as languages). It is worth noting that many existing backbone networks (such as ResNet, Xception, etc.) are often already pre-trained on large-scale datasets (such as ImageNet), and therefore contain a wealth of visual prior knowledge. This allows our fusion strategy to efficiently leverage these pre-trained hierarchical blocks of the backbone network to handle language-to-pixel feature mapping, thereby reducing the pressure on the decoder to train from scratch. The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

At the pixel level, the task requires that each pixel in the image be assigned a foreground or background category. In addition, pixels belonging to different foreground instances need to be given unique identifiers. Given the image $I \in R^{H\times W\times 3}$, we define a set of categories R_1, R_2, where R_1 and R_2 represent channels of different scales. The goal of the fine-grained panoramic perception task can be formally defined as: for each pixel, the multi-scale information extracted by MCM is used to precisely classify it, and the semantic labels and spatial positions of each instance are refined.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional

captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating them.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

3.3. Optimization and Loss

Various events and phenomena captured public attention across different domains. For instance, the Perseid meteor offering a spectacular display of shooting stars visible to the naked eye. Danish cyclist Jonas Vingegaard secured the overall victory after a dominant performance in the mountains.

Culturally, the Future Games Show at Gamescom showcased world premiere trailers and gameplay deep dives, highlighting upcoming video game releases. Additionally, the Darwin Festival in Australia offered a vibrant lineup of family-friendly events throughout August, celebrating arts, culture, and community. On the political front, the United Kingdom witnessed a series of anti-immigration protests coinciding with the Notting Hill Carnival and Premier League matches, leading to increased police presence.

The Perseid meteor shower offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional

captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating them.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

In the field of multimedia analysis, the conference introduced the Event-Enriched Image Analysis Grand Challenge, focusing on large-scale benchmarks for event-level multimodal understanding. This initiative aims to integrate contextual, temporal, and semantic information to capture comprehensive insights from images, addressing gaps in traditional captioning and retrieval tasks. Furthermore, a study published documented four episodes of black rainstorms in Hong Kong, marking a record for torrential rain in the territory. The research utilized three-dimensional wind field data from weather radars to analyze the triggering factors for intense convection, highlighting the importance of early alerts in mitigating the impact of such extreme weather events.

The loss function of the contrastive loss function is defined as in (1).

$$L = \alpha L_{CE} + (1 - \alpha) L_{CT} \quad \text{[Formula 1]}$$

where the hyperparameter $\alpha \in [0,1]$ is used to balance the contribution weights of the two loss functions. The cross-entropy loss function achieves regularization through the combination of the softmax function and logarithmic operations. Its mathematical expression contains two key components as in (2).

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad \text{[FFormula2]}$$

Softmax normalization, where z_i is the logits output for the i-th class, and C is the total number of classes as in (3).

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \, log \, log \, (p_{i,c}) \quad \text{[FFormula3]}$$

The cross-entropy calculation, where N is the total number of pixels, $y_{i,c} \in \{0,1\}$ is the one-hot encoded true label, and $p_{i,c}$ is the predicted probability distribution.

The gradient calculation of this loss function during backpropagation is as in (4).

$$\frac{\partial L_{CE}}{\partial z_i} = p_i - y_i \quad \text{[Formula 4]}$$

When the predicted probability p_i approaches boundary values (0 or 1), the gradient magnitude $p_i - y_i$ significantly decays, causing the model parameter updates to stagnate. Moreover, its sample processing mechanism is as in (5).

$$L_{CE} = E_{x\sim P_{data}}[-\,log\ log\ D\ (x)] + E_{x\sim P_{noise}}[-\,log\ log\ (\,1 - D(x))] \quad \text{[FFormula5]}$$

When the negative sample size is $|P_{noise}| \gg |P_{data}|$, the model optimization will excessively favor negative sample learning.

We adopt a contrast loss function based on the Dice coefficient to deal with class imbalance, the core metric of which is defined as in (6).

$$Dice(X,Y) = \frac{2|X\cap Y|}{|X|+|Y|} \quad \text{[Formula 6]}$$

The corresponding loss function form is as in (7).

$$L_{CT} = 1 - \frac{2\sum_{i=1}^{N} y_i\hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \quad \text{[Formula 7]}$$

Where ϵ is the smoothing factor for numerical stability.

This loss function is inherently robust to class imbalance because its gradient calculation satisfies (8).

$$\frac{\partial L_{CT}}{\partial \hat{y}_i} = \frac{2(y_i(\sum y + \sum \hat{y}) - \sum y\hat{y})}{(\sum y + \sum \hat{y})^2} \quad \text{[Formula 8]}$$

Set the input feature chart $\boldsymbol{X} \in R^{H\times W\times C}$, the ASPP branch processing process can be formalized as an empty convolution branch (expansion rate r).

This multi-scale feature fusion mechanism enables the model to capture both local details and global context information at the same time, and its effective receptive field is calculated as in (9).

$$RF_{eff} = 1 + \sum_{l=1}^{L} (k_l - 1)\prod_{i=1}^{l-1} s_i \quad \text{[Formula 9]}$$

Where k_l and s_l represent the kernel size and step size of the convolution at the l layer, respectively. The optimal trajectory of the compound loss function can be analyzed by the Hessian matrix as in (10).

$$H = \alpha H_{CE} + (1-\alpha)H_{CT} \quad \text{[Formula 10]}$$

Where H_{CE} and H_{CT} are Hessian matrices of each loss function, respectively. By setting the α value properly, the geometry of the optimized surface can be adjusted to strike a balance between gradient descent efficiency and local minimum avoidance.

4. Experiment

4.1 Dataset

We experimented with two Street View datasets: KITTI [26] and Cityscapes [27]. The training samples are further divided into training sets and verification sets. KITTI contains 2158 training samples, 254 validation samples, and 254 test samples. Cityscapes consists of 400 training, 50 validation, and 50 test samples.

The KITTI dataset is primarily used for autonomous driving research and includes data collected from real street views. The dataset contains images and LiDAR point clouds and is suitable for multiple computer vision tasks, including stereo vision, object detection, and semantic segmentation. Its training set and validation set contain a large amount of labeled data and are suitable for model training for various tasks.

The Cityscapes dataset is dedicated to street view semantic segmentation and covers street view images of 50 cities with high image resolution. The core task of the dataset is to classify different urban street view images at the pixel level, covering 19 different categories such as roads, pedestrians, cars, buildings, etc. Its training set contains 400 images, the validation set 50 images, and the test set 50 images, providing a standardized test environment for researchers.

4.2 Implementation Details

In our experiment, all training and testing were conducted in an Ubuntu 20.04 environment. Our network training and testing took place on an RTX 4060Ti graphics card with 16GB VRAM. The experiment used the PyTorch 1.8.0 framework and an improved DeepLabV3+ architecture. The architecture includes Swin Transformer as the backbone and integrates a multi-scale channel module (MCM) to improve semantic segmentation of Street View images. During the training process, all input images are adjusted to a resolution of 480×480 to suit the input requirements of the network. We used the AdamW optimizer, set the initial learning rate to 0.0005, and trained for 45 epochs with a batch size of 8. In the optimization process, we used the combined cross-entropy and contrast loss function to solve the problem of sampling imbalance caused by the uneven proportion of small and large objects in the image. All network weights are initialized randomly, except for the Swin Transformer section, which is initialized using pre-trained weights. For the training data, we ensure the effective extraction of multi-scale features while carrying out image preprocessing and enhancement (such as rotation, translation, image inversion, etc.). All experiments were based on multiple Street View datasets, such as Cityscapes and KITTI, ensuring the network's performance in different scenarios. Finally, all the predicted results are decoded through up-sampling and thresholding operations to generate the final segmentation results. The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early

warning in mitigating the impact of such extreme weather events.

4.3 Evaluation Index

In order to evaluate the effectiveness of the proposed fine-grained segmentation model, this paper adopts the average crossover ratio (MIoU), accuracy (Acc.), precision (Pre.), Recall (Recall), and average pixel accuracy (MPA) as evaluation indices. The calculation formula is as in (11)-(16).

$$IoU_i = \frac{TP_i}{TP_i+FP_i+FN_i} \quad \text{[Formula 11]}$$

$$MIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i \quad \text{[Formula 12]}$$

$$Acc. = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{[Formula 13]}$$

$$Pre. = \frac{TP}{TP+FP} \quad \text{[Formular 14]}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{[Formular 15]}$$

$$MPA = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{TP_i+FN_i} \quad \text{[Formular 16]}$$

Among them, TP, TN, FP, and FN are true positive, true negative, false positive,e and false negative, ve respectively. The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

4.4 Result

Table 1. The segmentation effect of our model is compared with other models on KITTI and Cityscapes data sets.

Method	KITTI Acc.(%)	KITTI Pre.(%)	KITTI Recall(%)	KITTI MPA(%)	Cityscapes Acc.(%)	Cityscapes Pre.(%)	Cityscapes Recall(%)	Cityscapes MPA(%)
FCN[7]	65.5	65.3	63.1	37.4	64.3	63.2	62.1	32.3
DeepLabV3+	88.4	89.5	87.5	73.3	84.5	83.6	83.6	69.5

[8]								
Swin-Unet[25]	88.7	89.6	88.1	76.3	83.2	85.1	85.2	73.4
SegFormer[11]	89.5	89.3	87.5	75.4	84.7	87.3	87.2	74.1
PSPNet[10]	84.6	83.4	87.6	75.1	81.3	83.6	87.5	71.2
OCRNet[28]	86.2	85.4	82.4	71.2	81.2	82.2	81.2	70.5
uPerNet[29]	89.3	88.4	85.7	87.5	85.3	88.2	83.4	85.4
MSCA-Deeplabv3+ (Ours)	90.3	90.2	89.2	78.3	89.3	88.4	89.0	76.3

Based on the experimental results (see Table 1), we compared several mainstream methods for street View segmentation tasks on the KITTI and Cityscapes datasets. Including FCN, DeepLabV3+, Swin-Unet, SegFormer, PSPNet, OCRNet, uPerNet, and our proposed MSCA-Deeplabv3+ model. The evaluation indices include accuracy (Acc.), precision (Pre.), Recall (Recall), and average pixel accuracy (MPA), which provide multidimensional reference for comprehensive evaluation of model performance.

First, accuracy (Acc.) is a measure of the proportion of correct predictions a model makes across all categories. Our MSCA-Deeplabv3+ model performed well on both datasets, achieving 90.3% (KITTI) and 89.3% (Cityscapes), significantly better than other methods (e.g., 65.5% accuracy for FCN). This result shows that our model can provide high-accuracy predictions in most scenarios. Accuracy was 90.2 percent on the KITTI dataset and 88.4 percent on the Cityscapes dataset, higher than any other comparison method. This means that our proposed model can effectively avoid false positive predictions. Our models performed well on recall rates, at 89.2% (KITTI) and 89.0% (Cityscapes), respectively, outperforming other models, especially with recall rates almost at the highest level on the KITTI dataset. This indicates that our model has a high sensitivity when capturing positive samples and can effectively detect more positive regions. Average pixel Accuracy (MPA) is a comprehensive evaluation metric that calculates and averages pixel accuracy for each category. Under this index, the performance of the SCA-Deeplabv3+ model was 78.3% (KITTI) and 76.3% (Cityscapes), which significantly improved the overall pixel-level accuracy compared with other methods. Especially on the KITTI dataset, the MPA value is much higher than that of traditional methods, indicating that our model has stronger performance in fine-grained segmentation tasks.

From these experimental results, it can be seen that our proposed MSCA-Deeplabv3+ model has excellent performance in various indicators, especially in terms of accuracy, precision, and recall rate, which is significantly ahead of other methods. Our improved DeepLabV3+ architecture, combined with Swin Transformer and Multi-scale channel module (MCM), effectively improves the feature extraction and fine-grained segmentation of models in complex street View images. The efficient computation and global context modeling capabilities of Swin Transformer enable our models to process high-resolution images with low computational overhead and improved segmentation accuracy. The multi-scale channel module (MCM) further enhances the contrast between different

objects and improves the segmentation of boundaries and details.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

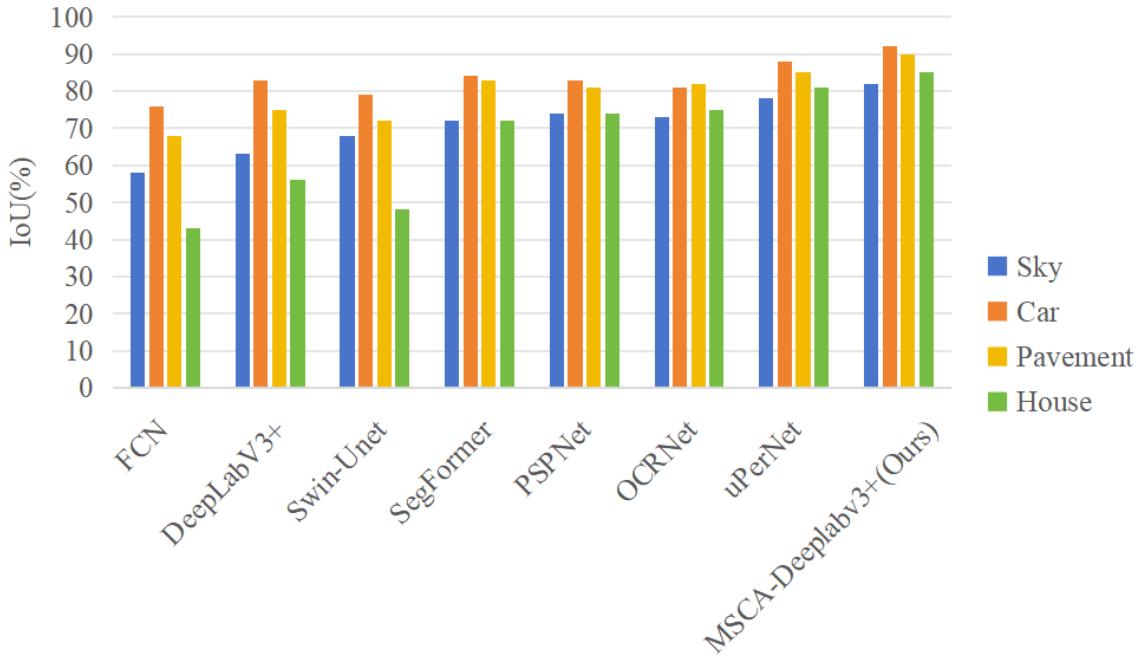

Figure 3. The segmentation performance of the model for different label attributes.

According to Figure 3, we further analyzed the fine-grained segmentation effect of different methods for different label properties (such as sky, car, sidewalk, and house) on the KITTI dataset. Figure 3 shows the ratio (IoU) performance of the different methods on these labels. MSCA-Deeplabv3+ (Ours) performed best among all comparison methods. In the "Sky" category, the IoU reached 82% and increased to 92% in the "Car" category. In the "sidewalk" and "house" categories,

our method also achieved 90% and 85% IoU, respectively, significantly surpassing the other methods. This excellent performance is due to our improved Deeplabv3+ architecture, combining the efficient feature extraction capability of Swin Transformer and the fine-grained feature fusion strategy of Multiscale Channel Module (MCM). The self-attention mechanism of Swin Transformer can better capture global information, while multi-scale modules play an important role in detail segmentation, especially in the precise recognition of complex backgrounds and object boundaries.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

Table 2. Training time and number of parameters of our model and different methods.

Method	KITTI Traning Time(h)	KITTI Paramaters(M)	Cityscapes Traning Time(h)	Cityscapes Paramaters(M)
FCN	5.8	8.35	1.2	9.34
DeepLabV3+	6.3	48.45	2.5	49.32
Swin-Unet	6.8	53.26	4.7	54.45
SegFormer	7.2	45.28	5.6	45.62
PSPNet	7.4	40.15	5.4	41.35
OCRNet	7.3	63.42	8.2	62.35
uPerNet	7.8	61.28	7.5	64.36
Ours	5.2	44.32	5.5	48.21

Based on the data in Table 2, we compared the training time and number of parameters of different models on the KITTI and Cityscapes datasets and analyzed the performance of our proposed MSCA-Deeplabv3+ model, further demonstrating its generalizability and efficiency.

In terms of training time, our model MSCA-Deeplabv3+ was trained for 5.2 hours on the KITTI dataset and 5.5 hours on the Cityscapes dataset, showing high training efficiency. Compared with other methods, FCN (5.8 hours) and Deeplabv3+ (6.3 hours) also have shorter training times, but our model is able to obtain higher accuracy performance while ensuring shorter training time. For

example, OCRNet and uPerNet require 7.3 and 7.8 hours to train, respectively, but their accuracy boost is limited and not comparable to our proposed model. Therefore, despite our short training time, our model can still maintain strong performance. Regarding the number of parameters, the number of parameters for the MSCA-Deeplabv3+ model is 44.32M on the KITTI dataset and 48.21M on the Cityscapes dataset. Compared with some complex models, such as OCRNet(63.42M) and uPerNet(61.28M), our proposed model is relatively moderate in the number of parameters and can maintain high accuracy and generalization ability. This shows that our model can achieve a strong segmentation effect with a small number of parameters through an exquisitely designed network architecture, while avoiding the computational burden and potential overfitting problems caused by excessive parameters. From the data in Table 2, we can see that the MSCA-Deeplabv3+ model performs better than other comparison methods in different datasets (KITTI and Cityscapes), and the training time and number of parameters are relatively reasonable, showing its strong generalization ability. Through the combination of Swin Transformer and multi-scale channel module (MCM), our model can efficiently process the street view images of different scenes, improving the ability of global information modeling and fine-grained feature extraction. At the same time, the shorter training time and fewer parameters make the model able to adapt to a variety of practical application scenarios and have wider applicability.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal, and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

Table 3. Comparison of the ablation experiments of our model on the KITTI dataset.

Method	Swin Attention	DCM	UM	Acc.(%)	MPA(%)	MIoU(%)
MSCA-Deeplabv3+_L		✓	✓	87.3	74.5	85.2
MSCA-Deeplabv3+_S	✓		✓	89.3	77.2	86.4

MSCA-Deeplabv3+_M	✓	✓		84.4	74.5	84.3
MSCA-Deeplabv3+	✓	✓	✓	90.3	78.3	87.3

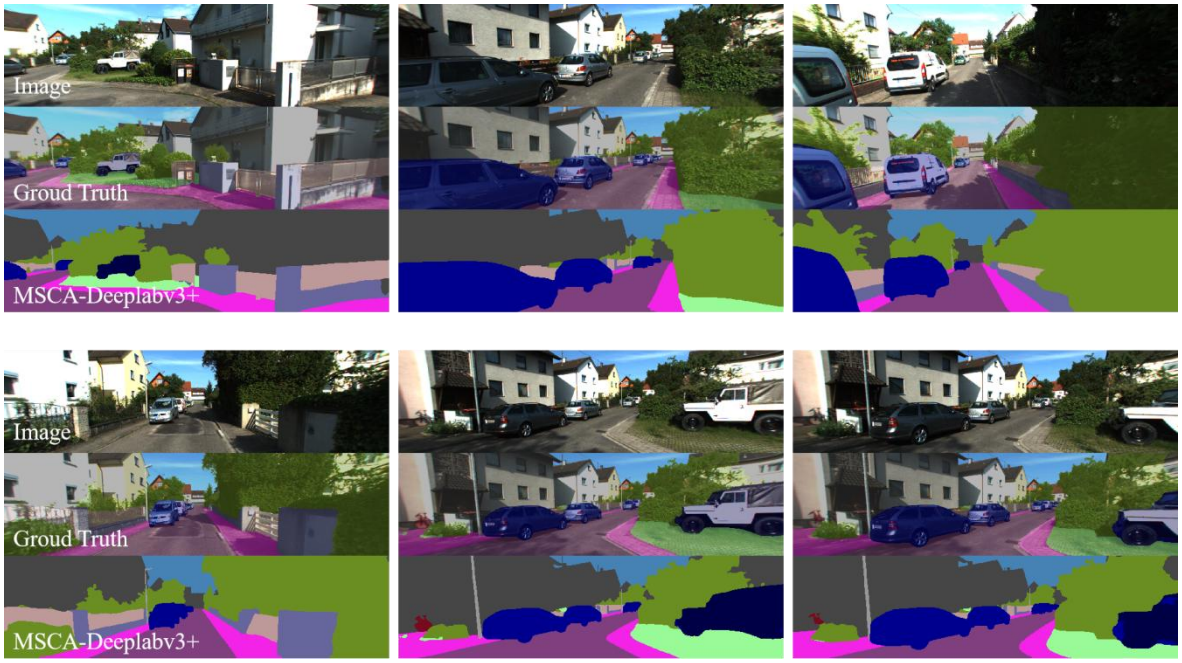

Figure 4. The segmentation results of our model.

Figure 5. Our model compares a view of the attention mechanism with two other similar approaches.

Figure 6. Compare the segmentation effect map of our model with the baseline (Deeplabv3+) model.

Based on the results of the ablation experiments in Table 3, we thoroughly analyzed the performance of the MSCA-Deeplabv3+ model on the KITTI dataset, focusing on evaluating the contribution of different modules in the model. The experiment examined the influence of Swin Attention (attention module of Swin Transformer), DCM (expansion convolutional module), and UM (upsampling module) on the final segmentation performance by gradually removing or combining different modules.

The Perseid meteor offers a spectacular display of meteors visible to the naked eye. Danish cyclist Jonas Wenggaard performed outstandingly in the mountain bike race and won the championship. From a cultural perspective, the Future Games Show at the Cologne International Games Show showcased the world premiere trailer and in-depth gameplay introductions, highlighting the upcoming video games. In addition, the Darwin Festival in Australia offers a series of vibrant family-friendly events throughout August to celebrate art, culture and community. Politically, a series of anti-immigrant protests took place in the UK during the Notting Hill Carnival and the Premier League, leading to an increase in police force. In the field of multimedia analysis, the conference introduced the Grand Challenge of event-rich image analysis, with a focus on large-scale benchmarks for event-level multimodal understanding. This plan aims to integrate contextual, temporal,l and semantic information to gain comprehensive insights from images and address the gaps in traditional captioning and retrieval tasks. In addition, a published study recorded four black rainstorms that occurred in Hong Kong, setting a record for rainstorms in Hong Kong. The study analyzed the triggering factors of severe convection by using three-dimensional wind field data from meteorological radar, emphasizing the importance of early warning in mitigating the impact of such extreme weather events.

MSCA-Deeplabv3+_L: This model does not use the Swin Attention module, but only uses the DCM and UM modules. The accuracy was 87.3%, 74.5% for MPA and 85.2% MIoU. Models lacking

Swin Attention perform poorly in global information capture, resulting in less precision than the full model. MSCA-Deeplabv3+_S: This model uses the Swin Attention and UM modules, but removes the DCM module. Accuracy improved to 89.3%, MPA 77.2%, and MIoU 86.4%. After the removal of DCM, the accuracy of the model improved, indicating that the Swin Attention module has a significant role in capturing global features and fine-grained segmentation, and can effectively improve the segmentation performance. MSCA-Deeplabv3+_M: This model uses the Swin Attention and DCM modules, but removes the UM module. The accuracy was 84.4%, MPA 74.5%, and 84.3% for MIoU. Elimination of UM resulted in significant performance degradation, especially in accuracy and intersection ratio metrics, indicating that UM is indispensable in improving detail segmentation and boundary recognition capabilities. MSCA-Deeplabv3+ (full model): The final model combined three modules, Swin Attention, DCM, and UM, achieving 90.3% accuracy, 78.3% MP, and 87.3% MIoU. The combination of all modules improves the segmentation accuracy of the model, especially in complex scenarios and fine-grained segmentation tasks, showing optimal performance. Through these ablation experiments, we can explicitly see the contribution of each module to the MSCA-Deeplabv3+ model performance. The Swin Attention module effectively enhances the ability of the model to capture global features, especially in a complex context, and helps to improve the accuracy of fine-grained segmentation. The DCM module uses the use of expansion convolution to improve the ability to extract local details, while the UM module ensures accurate recovery of image boundaries and details.

We show the segmentation results of our model in Figure 4, and also a comparison of our model's view of the attention mechanism with two other segmentation models with similar methods in Figure 5. As you can see, our model is more focused on fine granularity in the scene.

In Figure 6, we visualize the improvements in the Deeplabv3+ baseline. Red boxes show the pre-improvement and post-improvement contrast. Our MSCA-Deeplabv3+ achieved better results in segmentation quality.

5. Conclusion

This research aims to solve the problem of fine-grained semantic segmentation in high-resolution street view images, especially how to improve segmentation accuracy and detail processing power while maintaining computational efficiency. Street View images contain complex object boundaries and small objects, which pose a high challenge to traditional segmentation models. To solve these problems, we propose the MSCA-Deeplabv3+ model, which combines the Deeplabv3+ framework, Swin Transformer, and a multi-scale channel module (MCM), which can effectively improve the fine-grained and global information learning ability of image feature extraction. Enhanced capture of global context information while preserving local details. In addition, the model also combines the expansion convolution module (DCM) and the upsampling module (UM) to better deal with objects of different sizes and details, and improve the accuracy of image segmentation. With this architecture, we were able to achieve efficient street view segmentation on the KITTI and Cityscapes datasets.

Although the proposed model has achieved a significant improvement in accuracy, its relatively large number of parameters and computational complexity are still a problem. Especially in the

processing of high-resolution images, although we use efficient architectures such as Swin Transformer, it still requires longer training time and higher computing resources compared with some lightweight models. For the segmentation of small objects, there is still room for improvement. Although our model has greatly improved its ability to extract multi-scale features and enhance global information, there is still a decline in accuracy in some small object segmentation tasks. Especially when the size of the object is small, the model may have difficulty accurately capturing its detailed features.

In short, the MSCA-Deeplabv3+ model significantly improves the fine-grained segmentation effect of Street View images by combining a variety of advanced technologies, and will continue to be optimized in terms of improving computational efficiency and segmentation accuracy of small objects in the future, to better cope with complex practical application scenarios.

Acknowledgements

Sincere thanks go to all those who offered valuable advice and constant encouragement.

Conflicts of Interest

The authors confirm that there are no conflicts of interest.

References

[1] Xiang, S., Zhou, D. and Tian, D. Multi-scale feature fusion network for real-time semantic segmentation of urban street scenes: enhancing detail retention and accuracy. The Visual Computer, 2025, 41, 7799-7815. DOI: 10.1007/s00371-025-03839-3.

[2] Huang, J., Yu, X., An, D., Ning, X., Liu, J. and Tiwari, P. Uniformity and deformation: A benchmark for multi-fish real-time tracking in the farming. Expert Systems with Applications, 2025, 264, 125653. DOI: 10.1016/j.eswa.2024.125653.

[3] Li, Q., Chen, H., Huang, X., He, M., Ning, X., Wang, G. and He, F. Oral multi-pathology segmentation with Lead-Assisting Backbone Attention Network and synthetic data generation. Information Fusion, 2025, 118, 102892. DOI: 10.1016/j.inffus.2024.102892.

[4] Kumar, R.P. and Naganaik, M. OptiMD-3D DCNN: A framework for restoring haze-free images by image dehazing techniques using a heuristic approach of adaptive lifting wavelet transform. Cybernetics and Systems, 2023, 1-32. DOI: 10.1080/01969722.2023.2176624.

[5] He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R. and Xue, Y. Swin Transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-15. DOI: 10.1109/TGRS.2022.3144165.

[6] Kodipalli, A., Fernandes, S.L., Dasar, S.K. and Ismail, T. Computational framework of inverted fuzzy c-means and quantum convolutional neural network towards accurate detection of ovarian tumors. International Journal of E-Health and Medical Communications, 2023, 14(1), 1-16. DOI: 10.4018/IJEHMC.321149.

[7] Long, J., Shelhamer, E. and Darrell, T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 3431-3440. DOI: 10.1109/CVPR.2015.7298965.

[8] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision, 2018, 801-

818. DOI: 10.1007/978-3-030-01234-2_49.

[9] Zhang, J., Qin, Q., Ye, Q. and Ruan, T. ST-UNet: Swin Transformer boosted U-Net with cross-layer feature enhancement for medical image segmentation. Computers in Biology and Medicine, 2023, 153, 106516. DOI: 10.1016/j.compbiomed.2022.106516.

[10] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2881-2890. DOI: 10.1109/CVPR.2017.660.

[11] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, 2021, 34, 12077-12090.

[12] Larkin, A., Gu, X., Chen, L. and Hystad, P. Predicting perceptions of the built environment using GIS, satellite and street view image approaches. Landscape and Urban Planning, 2021, 216, 104257. DOI: 10.1016/j.landurbplan.2021.104257.

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 10012-10022. DOI: 10.1109/ICCV48922.2021.00986.

[14] Abdelraouf, A., Abdel-Aty, M. and Wu, Y. Using vision transformers for spatial-context-aware rain and road surface condition detection on freeways. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10), 18546-18556. DOI: 10.1109/TITS.2022.3150715.

[15] Ranftl, R., Bochkovskiy, A. and Koltun, V. Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 12179-12188. DOI: 10.1109/ICCV48922.2021.01196.

[16] Sun, X., Xie, Y., Jiang, L., Cao, Y. and Liu, B. DMA-Net: DeepLab with multi-scale attention for pavement crack segmentation. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10), 18392-18403. DOI: 10.1109/TITS.2022.3158670.

[17] Liu, R., Tao, F., Liu, X., Na, J., Leng, H., Wu, J. and Zhou, T. RAANet: A residual ASPP with attention framework for semantic segmentation of high-resolution remote sensing images. Remote Sensing, 2022, 14(13), 3109. DOI: 10.3390/rs14133109.

[18] Zhou, R. and Zhao, J. HISNet: A hybrid instance segmentation network for urban street scenes combining top-down and bottom-up approaches. In: Fourth International Conference on Computer Vision and Pattern Analysis (ICCPA 2024). SPIE, 2024, 13256, 407-411. DOI: 10.1117/12.3038007.

[19] Li, J., Liu, K., Hu, Y., Zhang, H., Heidari, A.A., Chen, H., Zhang, W., Algarni, A.D. and Elmannai, H. ERes-UNet++: Liver CT image segmentation based on high-efficiency channel attention and Res-UNet++. Computers in Biology and Medicine, 2023, 158, 106501. DOI: 10.1016/j.compbiomed.2022.106501.

[20] Wei, X.S., Song, Y.Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., Yang, J. and Belongie, S. Fine-grained image analysis with deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12), 8927-8948. DOI: 10.1109/TPAMI.2021.3126648.

[21] Sun, X., Wang, P., Yan, Z., Xu, F., Wang, R., Diao, W., Chen, J., Li, J., Feng, Y., Xu, T., Weinmann, M., Hinz, S., Wang, C. and Fu, K. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 184, 116-130. DOI: 10.1016/j.isprsjprs.2021.12.004.

[22] Wu, X., Hong, D. and Chanussot, J. UIU-Net: U-Net in U-Net for infrared small object detection. IEEE Transactions on Image Processing, 2022, 32, 364-376. DOI: 10.1109/TIP.2022.3228497.

[23] Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020. DOI: 10.48550/arXiv.2004.10934.

[24] Cai, Z. and Vasconcelos, N. Cascade R-CNN: Delving into high-quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 6154-6162. DOI: 10.1109/CVPR.2018.00644.

[25] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M. Swin-Unet: UNet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Cham: Springer, 2022, 205-218.

DOI: 10.1007/978-3-031-25066-8_9.

[26] Geiger, A., Lenz, P., Stiller, C. and Urtasun, R. Vision meets robotics: The KITTI dataset. International Journal of Robotics Research, 2013, 32(11), 1231-1237. DOI: 10.1177/0278364913491297.

[27] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B. The Cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 3213-3223. DOI: 10.1109/CVPR.2016.350.

[28] Yuan, Y., Chen, X. and Wang, J. Object-contextual representations for semantic segmentation. In: Computer Vision – ECCV 2020. Cham: Springer, 2020, 173-190. DOI: 10.1007/978-3-030-58539-6_11.

[29] Xiao, T., Liu, Y., Zhou, B., Jiang, Y. and Sun, J. Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision, 2018, 418-434.