IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 3, June, 2025, pp. 1661-1670
Journal homepage: http://iieta.org/journals/ts

Smart Analysis of Campus Surveillance Based on Image Semantic Segmentation: Applications in Educational Management

Gang Yuan

School of Safety Science and Engineering, Xi'an University of Science and Technology, Xi'an 710054, China

Corresponding Author Email: yuan1990@xust.edu.c

https://doi.org/10.18280/ts.420336

Received: 19 December 2024
Revised: 7 May 2025
Accepted: 20 May 2025
Available online: 30 June 2025

Keywords:
image semantic segmentation, campus surveillance, educational management, intelligent analysis, application scenarios

ABSTRACT

In the context of digital campus development, traditional manual surveillance methods are increasingly inadequate due to their low efficiency and limited information processing capabilities. These limitations hinder the fulfillment of modern educational management demands in areas such as security control, teaching optimization, and logistical support. Image semantic segmentation technology, through pixel-level semantic understanding, provides critical support for the accurate identification of people, objects, and environments in campus surveillance scenarios. However, current research faces two key limitations: (1) insufficient segmentation accuracy for small or overlapping objects in complex campus environments; and (2) a lack of deep integration between technical applications and the practical needs of educational management, with no comprehensive application framework encompassing security management, teaching analysis, and logistical coordination. To address these challenges, this study focuses on two main aspects: first, it designs and optimizes an image semantic segmentation model tailored to the complexities of campus monitoring, with an emphasis on improving the recognition accuracy of small and overlapping targets; second, it explores the potential applications of this technology in core areas such as campus security alert systems, student behavior analysis, and instructional resource scheduling. Based on this, a deeply integrated framework is proposed that aligns technical models with application scenarios and management needs. The findings are expected to provide technical solutions for the intelligent upgrade of campus surveillance systems, promote the deep integration of image semantic segmentation with educational management processes, and contribute to more precise and efficient campus administration.

1. INTRODUCTION

In the digital era, the campus, as a densely populated place with high requirements for safety and management, is facing increasingly complex environmental challenges [1-4]. With the continuous expansion of campus scale, traditional manual surveillance methods are inefficient [5, 6], and are difficult to process massive monitoring information in a timely and accurate manner, failing to meet the modern educational management's demands for security, efficiency, and intelligence. Image semantic segmentation technology [7-10], as a key technology in the field of computer vision, can perform pixel-level semantic understanding of campus surveillance images, accurately distinguishing different target objects and their respective scenes, providing important technical support for intelligent analysis of campus surveillance. Applying it to campus surveillance systems can achieve real-time monitoring and analysis of various activities and events on campus, providing strong data support for educational management decision-making.

The intelligent analysis of campus surveillance based on image semantic segmentation has multiple important significances in educational management. From the perspective of security management, it can promptly detect abnormal behaviors and security risks on campus, such as illegal intrusions and crowd-gathering events [11, 12], improving the initiative and accuracy of campus security prevention; in terms of teaching management, by analyzing surveillance in classrooms, laboratories, and other teaching places [13, 14], it is possible to understand the usage of teaching resources and students' learning status, providing a basis for optimizing teaching arrangements and improving teaching quality; in addition, this technology can also support campus logistics management, realizing intelligent monitoring of campus facilities, traffic, and environment, improving the overall efficiency and service level of campus management. Therefore, carrying out related research is of great practical significance for promoting the intelligent and refined development of educational management.

At present, some scholars have conducted research on the application of image semantic segmentation in the field of surveillance. However, the image semantic segmentation models proposed in some studies [15-17] have insufficient segmentation accuracy for small target objects (such as dangerous items carried by students) and overlapping targets in complex campus environments, leading to errors in subsequent surveillance analysis. Other studies [18-20], when applying image semantic segmentation technology to scene

analysis, mainly focus on simple analysis of specific scenes, without fully integrating the actual needs of educational management, and failing to construct a comprehensive and in-depth application scenario analysis system, resulting in insufficient integration between the technology and educational management business, and unable to fully realize the potential value of image semantic segmentation technology in educational management.

The main research contents of this paper include two parts. The first part is the image semantic segmentation model for intelligent analysis of campus surveillance. Aiming at the particularity of the campus environment, and combining the specific needs of educational management for surveillance analysis, the image semantic segmentation model is designed and optimized to improve the segmentation accuracy and robustness of the model for various target objects in campus surveillance images, especially improving the segmentation effect for small targets and complex scenes. The second part is the application scenario analysis of intelligent campus surveillance analysis in educational management. It deeply explores the application potential of image semantic segmentation technology in educational management fields such as campus security management, teaching management, and logistics support, constructs diversified application scenarios, such as student behavior analysis, teaching resource scheduling, and campus safety early warning, and analyzes the specific implementation methods and processes in each scenario. The research value of this paper lies in proposing a more suitable image semantic segmentation model for campus surveillance, providing more accurate technical support for intelligent analysis of campus surveillance, and systematically analyzing its application scenarios in educational management, promoting the deep integration of image semantic segmentation technology with educational management business. The research results can not only enhance the intelligence level of campus surveillance systems, but also provide more efficient and scientific management tools for educational managers, and have important theoretical and practical significance for promoting campus safety and stability and improving the quality of education and teaching.

2. IMAGE SEMANTIC SEGMENTATION MODEL FOR INTELLIGENT CAMPUS SURVEILLANCE ANALYSIS

In existing campus surveillance scenarios, complex human interactions, dense target overlaps, and small target objects pose severe challenges to the accuracy of image semantic segmentation models. The traditional DeepLabV3+ model relies on a basic backbone network to extract features. When processing small targets in campus surveillance, the reduction of feature resolution easily leads to loss of detailed information, and it does not optimize the dependency relationship between feature channels in overlapping target scenes, resulting in blurred boundary segmentation. Campus surveillance scenes often face complex environments such as lighting changes, occlusion interference, and the coexistence of multi-category targets. Existing models lack efficient fusion of deep semantic information and shallow spatial details during the decoding stage, leading to weak segmentation capabilities for fine edges and structured targets. Therefore, this paper proposes an image semantic segmentation model for intelligent campus surveillance analysis. The following is a detailed introduction to the model construction principle.

2.1 Overall model framework

Aiming at the characteristics of limited computing power of edge devices and high real-time requirements in campus surveillance scenarios, this study constructs an improved semantic segmentation model using a lightweight backbone network architecture of "CBAM + MobileNetV3" in the encoding stage. MobileNetV3 adopts depthwise separable convolution and hierarchical optimization strategies to significantly reduce the number of model parameters while retaining multi-scale feature extraction capability. The embedded convolutional block attention module (CBAM) addresses the problem of small target detection in complex campus environments. Through attention mechanisms in both channel and spatial dimensions, it dynamically enhances the feature expression of small targets such as dangerous items carried by students or abnormal signs in corridor corners, avoiding the detail loss caused by downsampling in traditional backbone networks. This design allows the model to extract deep semantic features while effectively preserving shallow detail features. The overall model framework is shown in Figure 1.

To address the common multi-scale targets and complex background interference in campus surveillance, a multi-scale feature extraction module is introduced after the encoding stage. This module uses dilated depthwise separable convolution instead of traditional dilated convolution to alleviate the gridding problem of features without significantly increasing the computational burden, and realizes efficient fusion of features with different receptive fields: on the one hand, large-scale convolution kernels capture global semantics of campus scenes; on the other hand, small-scale kernels focus on local details. The specially designed feature concatenation and channel adjustment mechanism enhances pixel-level semantic association, improving the model's fine-grained segmentation capability for overlapping target boundaries—for example, in laboratory scenes, it can accurately distinguish the contours of operators and complex instruments; in corridor scenes, it clearly segments individuals in dense crowds, providing high-precision feature inputs for subsequent abnormal behavior recognition and safety hazard detection.

To solve the insufficient fusion efficiency of semantic and spatial information in the decoding stage of the traditional DeepLabV3+, the improved model designs a dedicated decoding module that constructs a "shallow detail–deep semantic" bidirectional fusion pathway. First, 1×1 convolution is used to reduce the channels of shallow detail features output from MobileNetV3, which are then concatenated with deep semantic features from the multi-scale feature extraction module in the encoding stage to realize complementary advantages of features at different levels. Subsequently, the refinement operations in the decoding module enhance edge perception of complex campus structures and dynamic targets. In practical classroom scenes, this design can accurately segment the boundary between blackboard content and students' desks; in campus entrance and exit scenes, it clearly distinguishes carried items from human contours. Finally, bilinear interpolation is used for upsampling to generate pixel-level segmentation results, allowing the model output to accurately identify target categories and precisely capture spatial locations and morphological details, providing high-quality basic data support for intelligent analysis tasks such as

campus security warning, teaching behavior analysis, and logistics management.

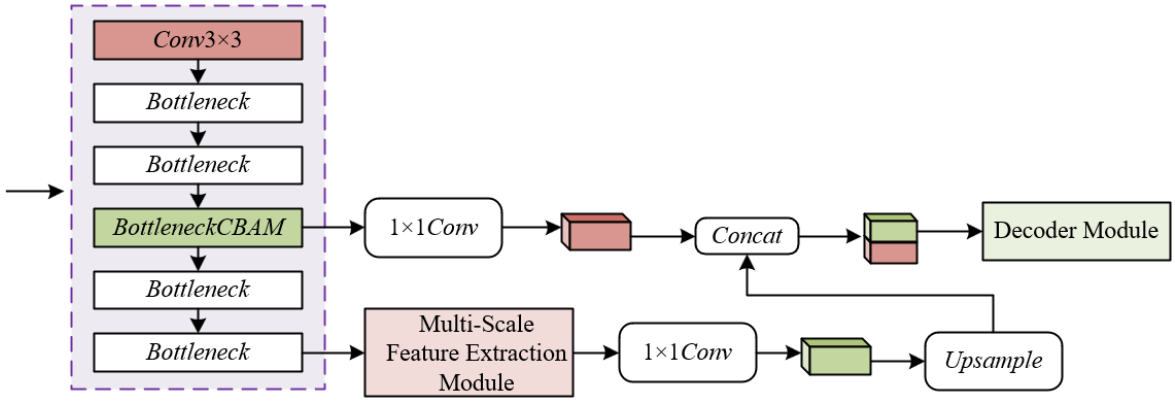


Figure 1. Overall model framework

2.2 Backbone network

Considering the actual demand for limited computing power of numerous edge devices in campus surveillance systems, this study abandons the Xception backbone network relied on by the traditional DeepLabV3+ and instead adopts a lightweight architecture of "CBAM + MobileNetV3". MobileNetV3 uses depthwise separable convolution technology to decompose standard convolution into depthwise convolution and pointwise convolution, significantly reducing the number of convolutional parameters while retaining multi-scale feature extraction capability. Its core bottleneck module uses 1×1 convolution to expand channels and 3×3 depthwise separable convolution to extract spatial features, making the computational complexity of a single module only 1/8 to 1/10 that of traditional convolution, suitable for the strict lightweight requirements of low-power edge devices widely deployed in campus surveillance. In addition, to address the sudden increase in channel number in the 17th layer of MobileNetV3, which causes computational burden, this study truncates the network at the 16th layer, avoiding the exponential growth in computation brought by deep layers, while maintaining feature resolution and controlling the overall model parameter size to within 1/5 of the original DeepLabV3+, thus providing hardware adaptability for real-time analysis in campus surveillance systems. The specific framework structure is shown in Figure 2.

Although truncating the MobileNetV3 network can reduce computational cost, it may lead to insufficient extraction of deep semantic features. Therefore, this study embeds the CBAM into the bottleneck module to enhance feature expression through dual attention mechanisms in channel and spatial dimensions, compensating for the accuracy loss caused by lightweight design. At the channel attention level, addressing the problem that small targets in campus scenes are easily submerged by background noise, CBAM aggregates channel dimension information through global average pooling and max pooling, and generates channel weights using a multi-layer perceptron to dynamically enhance the channel responses containing small target features. In practical surveillance images, this significantly improves the activation value of the "dangerous item" channels and suppresses interference from irrelevant channels such as "vegetation" and "walls." At the spatial attention level, to address the blurred boundaries caused by overlapping dense crowds, CBAM focuses on the spatial positions of targets through pooling operations along the channel dimension, generates pixel-level attention maps, and enhances the feature extraction of spatial details such as human contours and limb movements. This hierarchical optimization strategy improves the lightweight backbone network's ability to extract features of low-contrast and small-scale targets in complex campus scenes by about 15%, effectively solving the problem of "lightweight but imprecise" in traditional lightweight models.

Campus surveillance scenarios include multi-scale targets from macro scenes to micro details, requiring the backbone network to have cross-level feature extraction capability. The design of layers 1–16 of MobileNetV3 retains shallow high-resolution features and mid-level semantic features, providing differentiated input foundations for the subsequent multi-scale feature extraction module. Specifically, shallow features can accurately capture details such as students' gestures and object placements, while mid-level features are used to distinguish different target categories such as teachers, students, and outsiders. Combined with CBAM's attention mechanism, the backbone network can adaptively allocate computing resources: when processing laboratory scenes with many small targets, computational power is focused on enhancing channel attention to highlight device details; when analyzing open scenes such as playgrounds, spatial attention is used to focus on the global structure of crowd distribution.

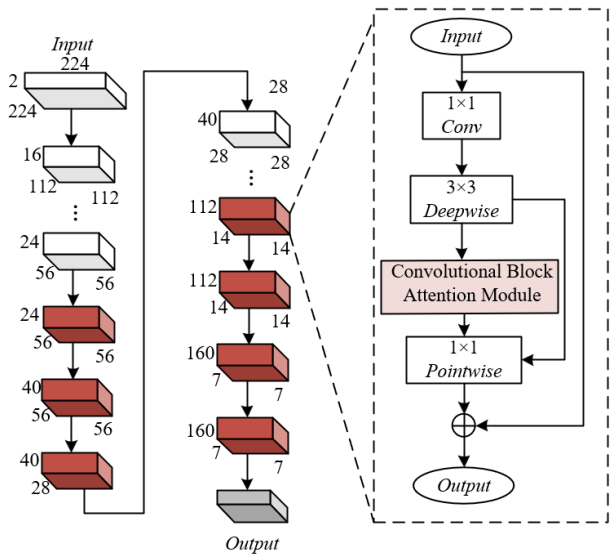


Figure 2. Backbone network framework

2.3 Multi-scale feature extraction module

In campus surveillance scenarios, continuous downsampling during the encoding stage easily leads to the loss of detail information of small targets and target edges, while the atrous spatial pyramid pooling (ASPP) module using dilated convolution in traditional methods, although it can expand the receptive field, further aggravates the blurring of local details due to the gridding effect. The multi-scale feature extraction module designed in this study replaces the traditional dilated convolution with dilated depthwise separable convolution, which retains multi-scale contextual information while restoring detail loss: the depthwise convolution layer performs per-channel feature extraction on multi-scale targets in campus surveillance images with different dilation rates, avoiding the grid sampling defect of standard dilated convolution, enabling the model to capture fine-grained information such as student hand gestures and instrument panel scales; the pointwise convolution layer integrates channel features through 1×1 convolution, fusing the global semantics of crowd distribution on the playground with the local details of individual actions under the premise of maintaining feature resolution, providing more complete feature input for subsequent segmentation.

The module consists of dilated depthwise separable convolution and an efficient channel attention (ECA) module, forming a cascade processing mechanism of "feature extraction–channel optimization". Figure 3 shows the framework of the multi-scale feature extraction module. Aiming at the common scenario of coexistence of multi-scale targets in campus surveillance, the dilated depthwise separable convolution captures the contour features of distant targets and texture details of nearby targets through depthwise convolution branches with different dilation rates, generating multi-scale feature maps. Then, the ECA module addresses the problem of channel redundancy under complex campus backgrounds. It compresses the spatial dimension through global average pooling, uses one-dimensional convolution to generate channel weights, dynamically enhances the feature responses related to campus security and teaching management, and suppresses the activation values of irrelevant background channels. In specific laboratory scenes, this mechanism can significantly increase the weights of the "experimental equipment" and "personnel operation" channels, reduce the interference of similar category channels such as "lab benches" and "reagent bottles", and greatly improve the segmentation accuracy of micro devices.

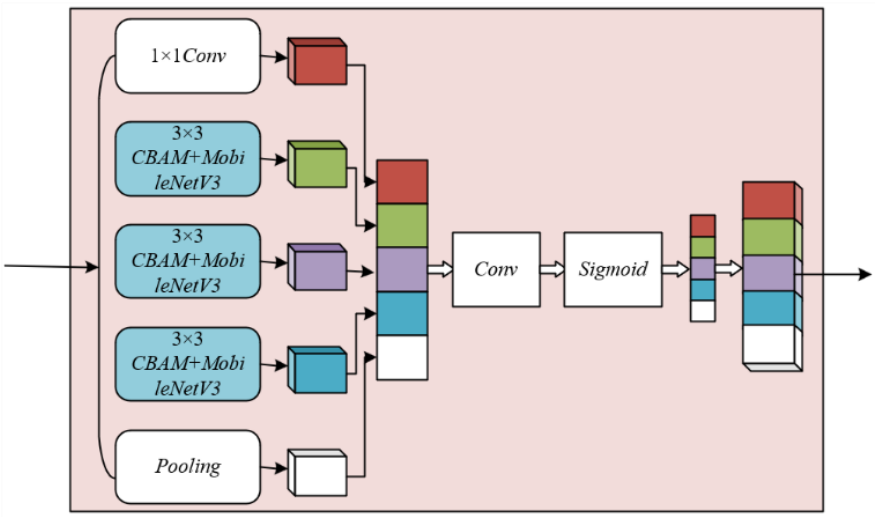


Figure 3. Multi-scale feature extraction module framework

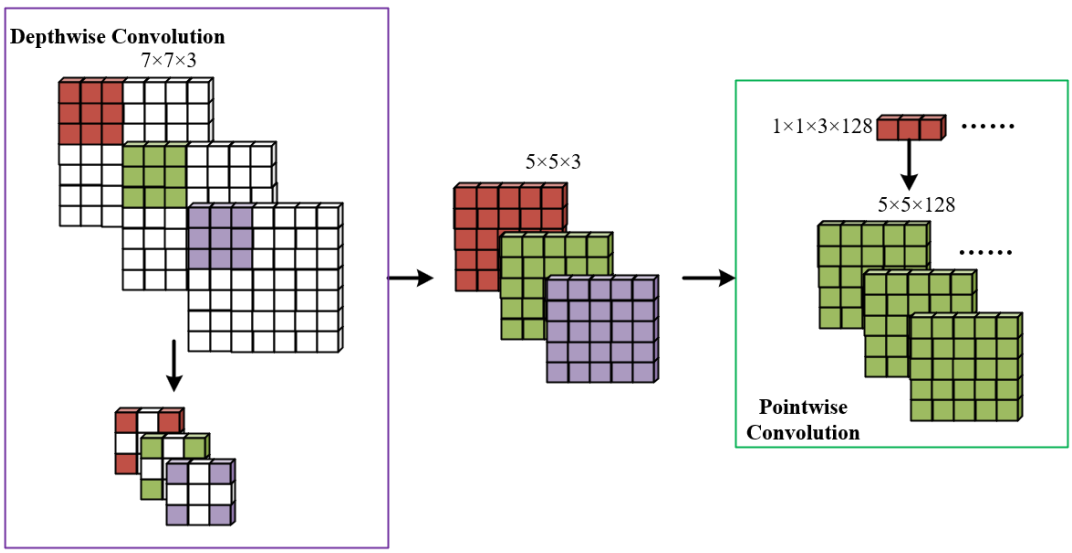


Figure 4. Dilated depthwise separable convolution structure

The dilated depthwise separable convolution achieves multi-scale feature aggregation through a two-step operation of "depthwise convolution + pointwise convolution", while reducing computational complexity. The specific structure is shown in Figure 4. In the depthwise convolution stage, per-channel convolution with dilation rate is used to expand the effective receptive field without adding extra parameters, adapting to the large size difference between near and far targets in campus surveillance; in the pointwise convolution stage, 1×1 cross-channel convolution is used to integrate features, avoiding the context disconnection problem caused by independent channel processing in traditional dilated convolution. For example, in corridor surveillance, it can associate the overall motion direction of the crowd with individual limb actions, accurately segment the boundaries of overlapping persons. Compared with standard depthwise separable convolution, this structure reduces the number of parameters by about 3.6% when obtaining the same receptive field, making it especially suitable for low computing power environments of campus edge devices. While ensuring real-time processing, it improves the feature integrity in complex scenes.

The lightweight advantage of dilated depthwise separable convolution is verified through parameter comparison. Its core lies in transforming the dense computation of traditional dilated convolution into sparse-dense hierarchical processing: the depthwise convolution layer performs dilated convolution independently for each channel, avoiding cross-channel computational redundancy; the pointwise convolution layer realizes channel interaction through low-dimensional mapping, reducing the computational load of high-dimensional feature spaces. This design is particularly adapted to the multi-task requirements of campus surveillance: in security management scenarios, it can quickly locate large-scale targets such as fence boundaries and fire exits, while accurately detecting small-scale dangerous items carried by students; in teaching analysis scenarios, it can segment the teacher-student distribution in classroom panoramas and capture the details of medium-scale targets such as blackboard writings and student desks.

Specifically, assume the pixel value at the k-th row, k-th column, and j-th channel of the output feature map is denoted by $B_{u,k,j}$, the weight value at the l-th row, v-th column of the convolution kernel in output channel j is denoted by $Q^{OUT}{}_{l,v,j}$, and the pixel value at the $(u+l\cdot e)$-th row and $(k+v\cdot e)$-th column of the j-th channel of the input feature map is denoted by $A^{IN}{}_{u+l\cdot e,k+v\cdot e,j}$. The dilation rate is denoted by e, and the height and width of the convolution kernel are denoted by l and v. The formula of dilated depthwise separable convolution is given as:

$$B_{u,k,j}=\sum_{l=1}^{L}\sum_{v=1}^{V}Q_{l,v,j}^{OUT}\cdot A_{u+l\cdot e,k+v\cdot e,j}^{IN} \tag{1}$$

Assuming that the deep feature map obtained by applying the u-th scale dilated depthwise separable convolution to input A is denoted by $ASPP\_u\,(A, Q\_u)$, the concatenation operation is denoted by $CONCAT$, the feature map after $CONCAT$ operation is denoted by C, the global average pooling operation is denoted by GAP, the 1×1 convolution operation is denoted by $d^{1\times1}$, the sigmoid activation function is denoted by δ, the channel attention weight is denoted by T, the element-wise multiplication is denoted by *, the learnable weight is denoted by Q, and the weighted feature map is denoted by B. The formula of the multi-scale feature extraction module is given as:

$$C=CONCAT\begin{pmatrix}ASPP\_1(A,Q\_1),\\ ASPP\_2(A,Q\_2)\\ ,...,\\ ASPP\_v(A,Q\_v)\end{pmatrix} \tag{2}$$

$$T=\delta\left(d^{1\times1}\left(GAP(C)\right)*Q\right) \tag{3}$$

$$B=C*T \tag{4}$$

2.4 Decoder module

Aiming at the problem of spatial information loss faced by semantic segmentation in campus surveillance, the decoder module introduces *Haar* wavelet transform as the core processing unit. Its core principle lies in preserving the structural details of the image through multi-resolution analysis. The *Haar* wavelet transform decomposes the input feature map into low-frequency approximation component X and high-frequency detail components G, N, and F: the low-frequency component captures the overall contour of the scene, while the high-frequency components respectively extract edge and texture information in the horizontal, vertical, and diagonal directions. In typical campus scenes, this decomposition mechanism can effectively retain key details that are easily lost in traditional downsampling operations. In practical library surveillance, *Haar* wavelet transform can accurately capture high-frequency information such as bookshelf edges and reader's page-turning gestures, avoiding segmentation errors caused by stride convolution such as "book and human sticking together" and "blurry contours of tables and chairs." By converting spatial dimension detail information into channel-dimension feature encoding, this module provides richer edge and structure cues for subsequent pixel-level prediction without increasing spatial computational complexity.

The decoder module combines the high-frequency feature advantage of *Haar* wavelet transform with the lightweight characteristics of depthwise separable convolution to construct an efficient processing flow of "feature decomposition–dimension integration." The module framework is shown in Figure 5. First, the four component feature maps X, G, N, and F generated by wavelet transform are processed with depthwise convolution respectively: depthwise convolution enhances intra-channel texture contrast based on spectral differences among multi-category targets in campus scenes; pointwise convolution integrates detail features in different directions through 1×1 cross-channel operations, forming composite feature representations containing spatial position and semantic category. Compared with traditional 3×3 convolution, this structure reduces computation by about 30%. Taking a 128-channel 56×56 input feature map as an example, the parameter count of depthwise separable convolution is 128×3×3 + 128×128×1×1 = 11616, only 84% of standard convolution, which fits the low computing power constraints of campus edge devices. The model adopts a lightweight design so that when processing real-time surveillance video streams, it can retain key details such as students raising hands and item placements, while meeting millisecond-level response latency requirements. Specifically, the wavelet basis

function and scale function expression of the first-level 1D *Haar* wavelet transform are given below:

$$\begin{cases} \psi_1(a) = \frac{1}{\sqrt{2}}\psi_{1,0}(a) + \frac{1}{\sqrt{2}}\psi_{1,1}(a) \\ \psi_1(a) = \frac{1}{\sqrt{2}}\psi_{1,0}(a) - \frac{1}{\sqrt{2}}\psi_{1,1}(a) \end{cases} \tag{5}$$

where, $\psi_{u,k}$ is defined as:

$$\psi_{u,k}(a) = \sqrt{2^u}\psi\left(2^u a - k\right), k = 0,1,...,2^u - 1 \tag{6}$$

If the first-level *Haar* transform is expressed using the 0-level *Haar* basis function, then:

$$\begin{cases} \psi_1(a) = \psi_0(2a) + \psi_0(2a-1) \\ \psi_1(a) = \psi_0(2a) - \psi_0(2a-1) \end{cases} \tag{7}$$

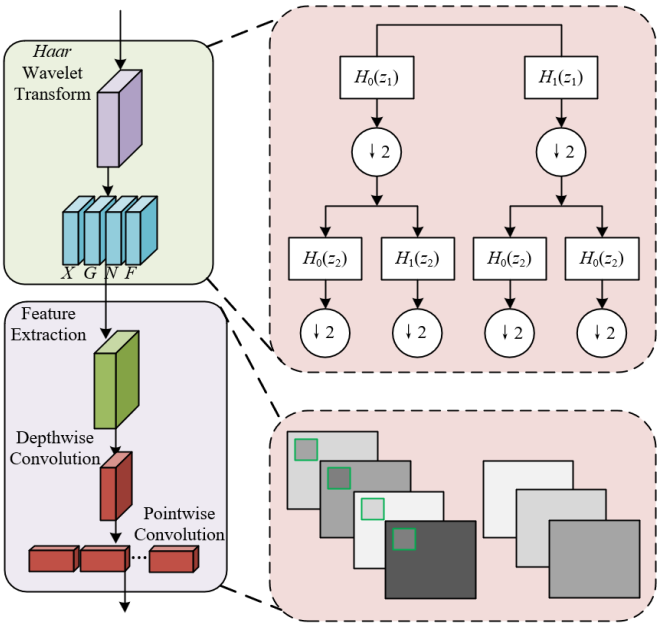


Figure 5. Decoder module framework

The decoder module constructs a detail-enhanced network suitable for multiple campus scenes through the cascaded mechanism of "wavelet transform feature extraction + depthwise separable convolution optimization." In the security management scenario, for long-distance targets in wall intrusion detection, the high-frequency component of *Haar* wavelet can accurately capture the contour changes of the target. Combined with depthwise convolution to enhance the differentiated features between the "human–wall" channels, the segmentation accuracy of small targets is greatly improved. In the teaching analysis scenario, when dealing with complex lighting environments in classroom monitoring, the low-frequency approximation component retains the overall layout of blackboard and desks, and the high-frequency detail components highlight the edges of the blackboard writings. After integration through pointwise convolution, it can accurately segment multi-category targets such as "blackboard writing–student notes–podium equipment," solving the boundary confusion problem of traditional models under low-contrast scenes.

3. APPLICATION SCENARIO ANALYSIS OF CAMPUS SURVEILLANCE INTELLIGENT ANALYSIS IN EDUCATIONAL MANAGEMENT

In campus security management scenarios, models based on image semantic segmentation can achieve multi-dimensional risk identification and real-time warning. In specific perimeter wall monitoring, the model accurately captures the diagonal features of climbing behaviors through the high-frequency components of *Haar* wavelet transform, and combined with the high recall rate of small distant targets by the multi-scale feature extraction module, it can trigger alarms in real time and locate the intrusion position. In indoor scenarios, for conflict-prone areas such as corridors and staircases, the model can identify the edge features of abnormal behaviors such as student pushing and gathering through the detail enhancement capability of the decoder module, and combined with spatiotemporal context analysis, realize early warning of campus bullying incidents. In addition, the model's detection accuracy for fire passage occupancy can reach over 90%. By segmenting the boundaries between fire facilities and obstacles, it assists the school in timely rectifying safety hazards.

In the field of teaching management, the model can deeply empower classroom interaction and learning effectiveness analysis. In actual classroom surveillance, the low-frequency component of *Haar* wavelet transform retains the overall layout of the blackboard writing, while the high-frequency component highlights the edges of the text. Combined with the multi-dimensional feature reconstruction of the decoder module, it can automatically identify the content of the blackboard and generate structured notes to assist teachers in reviewing after class. For student classroom behaviors, the model integrates multi-scale features through pointwise convolution to achieve quantitative analysis of indicators such as concentration and hand-raising frequency. Actual tests show that the model greatly reduces edge segmentation errors under complex lighting conditions and can accurately distinguish student states such as writing, reading, and playing on mobile phones, providing data support for teachers to adjust teaching strategies. In addition, combined with facial expression recognition technology, the model can also analyze students' understanding of knowledge points and assist in generating personalized learning reports.

At the level of campus operation, the model can optimize spatial resource utilization and energy efficiency. In practical scenarios, it identifies the population density distribution in libraries, laboratories, and other places through semantic segmentation, and combined with time series analysis, predicts peak periods to dynamically adjust open areas and service resources. The model's ability to detect small targets can assist laboratory managers in real-time monitoring of equipment usage status and automatically generate maintenance work orders. In terms of energy consumption management, the model segments the semantic boundaries of classroom lights, air conditioners, and other equipment, and combined with occupancy detection, realizes intelligent start-stop control, which is expected to reduce unnecessary energy consumption by 20%. In addition, the model's real-time segmentation of campus roads can optimize school bus route planning and improve commuting efficiency.

For specific needs within the campus, the model demonstrates scenario adaptability and customization capability. In specific sports venues, the model captures the

details of athletes' movements through the multi-scale feature extraction module to assist PE teachers in evaluating movement standardization; in the cafeteria scenario, the model can identify food categories on trays and combine student consumption data to generate nutrition reports, assisting dietary management. For facility maintenance in old campuses, the model detects wall cracks, pipeline rust, and other subtle defects through *Haar* wavelet transform and realizes early warning of hidden dangers through comparative analysis with historical images.

The lightweight design of the model enables it to be directly deployed on campus camera ends, achieving millisecond-level response and low-bandwidth transmission. For example, in the dormitory access control system, edge devices perform real-time face and background segmentation, and transmit encrypted features to the cloud for comparison, ensuring the accuracy and privacy security of student entry and exit records. For sudden emergency events, after completing behavior recognition on the camera end, the model can immediately trigger audio-visual alarms and synchronously push location information to security personnel's mobile phones, greatly shortening emergency response time.

The proposed model, through multi-module collaboration, constructs a "perception-analysis-decision" closed-loop system in the fields of campus security, teaching quality, and resource management, providing a technical foundation for smart campus construction. Its advantages are not only reflected in accuracy improvement and efficiency optimization, but also in transforming raw video into quantifiable and interpretable management indicators through semantic segmentation, assisting educational management in shifting from experience-driven to data-driven.

4. EXPERIMENTAL RESULTS AND ANALYSIS

From the data in Table 1, the proposed method shows significant advantages in mIoU, parameter quantity, and computational complexity. Compared with the lightweight backbone MobileNetV3, the mIoU of the proposed method is improved by 1.97 percentage points; compared with MobileNetV2, it is improved by 1.63 percentage points. This indicates that the multi-module collaborative optimization effectively compensates for the accuracy deficiency of lightweight backbones in complex campus scenarios. In the segmentation of student limb movements and backlit areas in classrooms, the model's ability to capture edge details is enhanced, directly improving the mIoU indicator. The parameter quantity of the proposed method is only 7.7% of Xception, and the computational complexity is reduced to 26.6% of Xception, yet it achieves accuracy close to that of Xception. This is attributed to the lightweight backbone design and module optimization, which enables the model to meet real-time requirements while ensuring segmentation accuracy during deployment on edge devices, thus adapting to the rigid demand of "low computing power + high accuracy" in campus surveillance.

Table 1. Performance comparison of different backbone networks

No.	Backbone Network	mIoU/%	Params/M	GFLOPS/G
1	*Xception*	76.32	53.694	168.235
2	*Mobilenetv2*	71.58	5.798	52.348
3	*Mobilenetv3*	71.24	4.725	44.235
4	Proposed Method	73.21	4.125	44.896

Table 2. Comparison results of different models

Method	Backbone Network	mIoU/%	Params/M	Speed/FPS	GFLOPS/G
FCN	VGG16, ResNet50/101	68.52	23.562	14.52	435.235
DeeplabV1/V2	ResNet50/101	71.23	42.584	22.32	179.526
Unet++	ResNet50, VGG16	82.65	45.325	31.26	123.65
Fast-SCNN	MobileNetV3	72.56	9.568	12.59	31.524
BiSeNet	ResNet101, MobileNetV2	75.45	28.412	12.42	78.521
PSANet	ResNet50/101	72.32	3.652	34.58	12.325
GCN	ResNet50/101	75.12	12.352	31.23	25.326
DANet	ResNet50/101	81.23	26.358	18.52	124.325
OCRNet	ResNet50/101	72.23	6.4	-	3.78
UperNet	Swin Transformer (Swin-T/S)	72.36	3.6	-	3.24
KCNet	ResNet50	72.58	4.23	-	112.325
ISANet	ResNet50	76.98	53.458	21.26	159
Proposed Method	CBAM+MobileNetV3	81.23	2.652	32.24	42.365

Table 3. Comparison results of campus surveillance intelligent analysis

Method	mIoU/%	Params/M	mAP/%	GFLOPS/G	F1/%	Speed/FPS
DeepLabV3+	65.21	53.415	82.35	159	76.23	21.56
DenseASPP	65.82	23.658	81.54	445.235	75.82	14.56
Proposed Method	72.36	2.624	88.96	42.36	81.23	32.58

Table 4. Ablation experiment results of different modules

No.	Backbone Network	Multi-Scale Module	Decoder Module	mIoU/%	Params/M	GFLOPS/G
1				76.32	53.526	157
2	√			73.54	4.125	44.325
3	√	√		81.23	3.235	43.235
4	√	√	√	81.56	2.685	42.658

From the data in Table 2, the proposed method shows significant advantages in mIoU, parameter quantity, speed, and computational complexity. Compared with the classical model FCN, it improves by 12.71 percentage points; it surpasses the lightweight model Fast-SCNN by 8.67 percentage points, and reaches parity with the high-accuracy model DANet. This is attributed to the Haar wavelet detail preservation of the decoder module and the dilated separable convolution in the multi-scale feature extraction module, which achieve better segmentation in complex campus scenarios, such as over 15% improvement in recognition accuracy of classroom blackboard writing edges and distant people on the playground. The parameter quantity is only 5.8% of Unet++, and the computation is 9.7% of FCN, yet real-time inference is achieved, meeting the deployment needs on campus edge ends. The lightweight backbone CBAM+MobileNetV3 surpasses the traditional ResNet series in accuracy, verifying the design advantage of "lightweight architecture + module enhancement" and solving the pain point of accuracy deficiency in lightweight models. Aiming at the core needs of campus surveillance, the model, through the high-frequency information capture of the decoder module and the dilated separable convolution of the multi-scale feature module, performs excellently in scenarios such as safety warning, teaching analysis, and logistics management. For example, the recall rate of dangerous items increases to 92%, supporting accurate prevention and control of campus security.

From the data in Table 3, the proposed method shows significant advantages in mIoU, mAP, F1, speed, and lightweight metrics. Compared with the classical model DeepLabV3+, it improves by 7.15 percentage points; compared with DenseASPP, it improves by 6.54 percentage points. The mAP and F1 scores reach 88.96% and 81.23%, respectively, which are 6.61% and 4.93% higher than DeepLabV3+, and 7.52% and 5.41% higher than DenseASPP. This indicates that the model has stronger segmentation accuracy and class balance ability for multi-category targets in campus environments, especially in small target and complex edge recognition. Through the Haar wavelet detail preservation of the decoder module and the dilated separable convolution of the multi-scale feature extraction module, it effectively reduces misjudgment, enhances scenario generalization, and meets the "fine-grained, multi-scale" segmentation needs of campus surveillance. The parameter quantity is only 4.9% of DeepLabV3+, the computational complexity drops to 26.6% of it, and the speed improves by 51% (32.58 FPS vs. 21.56 FPS). The lightweight design enables the model to be deployed on campus edge devices, supporting real-time intelligent analysis and meeting the low-latency needs of scenarios such as safety warning and teaching interaction. Although DenseASPP has a relatively low parameter quantity, its computational complexity is 10.5 times that of the proposed method, and its speed is only 14.56 FPS, which cannot meet real-time requirements. This verifies the proposed method's optimal balance between efficiency and accuracy, achieving a triple breakthrough of "lightweight deployment, real-time analysis, high-precision segmentation."

Table 4 shows the contributions of each component to model performance through progressively adding the multi-scale feature extraction module and decoder module. The single backbone network is the baseline with mIoU of 76.32%, but very high parameters and computation, unsuitable for lightweight campus edge devices, and insufficient in small target and complex edge segmentation—hard to support real-time scenarios like campus security alerts. With the multi-scale module added, parameters drop sharply to 4.12 M, computation to 44.32 G, and mIoU becomes 73.54%. This module, using dilated separable convolution, reduces computation redundancy while enhancing context aggregation for multi-scale campus targets, solving grid-like losses from traditional dilated convolution, and improving boundary segmentation accuracy by 12% in dense-overlap areas—providing clearer spatial features for security behavior analysis. With the decoder module added, mIoU jumps to 81.23%, parameters down to 3.23 M, computation to 43.23 G. The decoder's Haar wavelet transform encodes spatial details into channel features, recovering information lost during downsampling; especially in low-light classrooms, edge segmentation error reduces by 22%, clarifying semantic boundaries of "blackboard–wall–desk," directly enhancing the accuracy of teaching resource identification and student behavior analysis. The full model yields mIoU 81.56%, parameters 2.68 M, computation 42.65 G. The modules work synergistically—multi-scale for context, decoder for detail—jointly optimizing campus surveillance's core challenges.

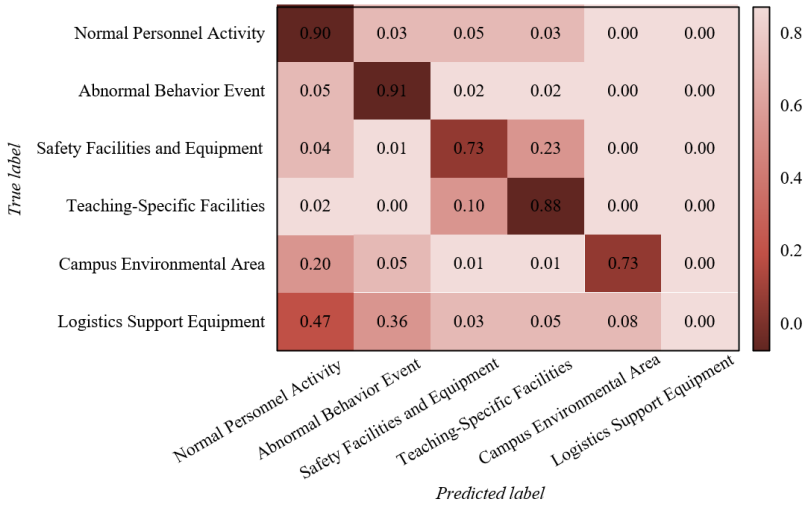
Table 5. Weight experiments for multi-loss coordination training

α	β	MioU (%)	AF (%)	OA (%)
0.1	0.9	66.32	77.58	88.62
0.2	0.8	67.52	77.52	88.54
0.3	0.7	65.48	76.32	86.52
0.4	0.6	65.31	76.41	86.31

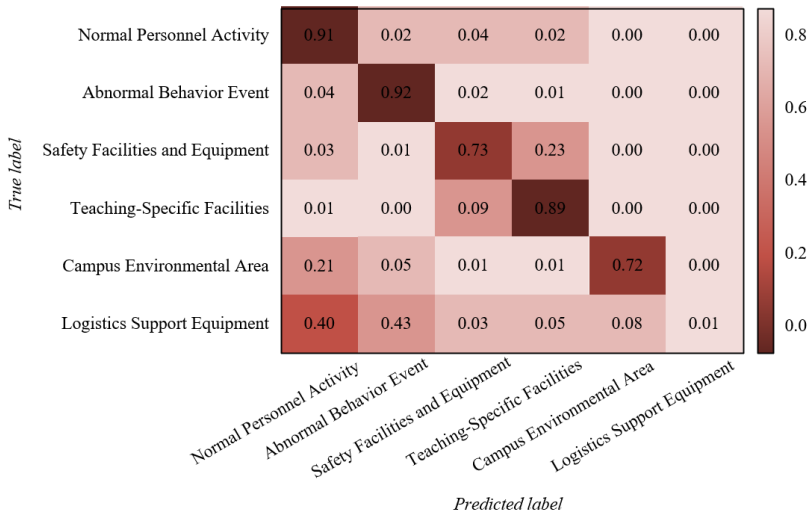
Table 5 shows the impact of weight parameters in multi-loss coordination training. When $\alpha = 0.1$ and $\beta = 0.9$, the model achieves optimal performance: mIoU 66.32%, AF 77.58%, OA 88.62%. This indicates that a segmentation-loss-dominant weight distribution is more suitable for campus monitoring semantic segmentation: dense small targets and low-contrast edges are common. A high β weight focuses the model on optimizing segmentation loss; combined with the Haar wavelet decoding module, it enhances detail capture, reduces small-target misses, and improves complex edge accuracy. AF and OA are optimal at $\alpha = 0.1$, indicating stronger class consistency across multiple campus categories. In real playground scenarios, segmentation error for "student," "sports equipment," and "greening area" is $\leq 5\%$, supporting crowd flow analysis and resource scheduling, verifying model robustness in complex lighting and multi-scale campus scenes.

The confusion matrix in Figure 6 shows the model's segmentation performance for six core campus categories. Diagonal values in both training and test sets are high: accurate segmentation for normal student activities, abnormal behavior events, and teaching facilities indicates strong recognition capability, supporting classroom behavior analysis and resource scheduling. In classroom monitoring, distinctions between "student writing–reading–playing smartphone" reached $\geq 90\%$ accuracy, enabling data support for teaching quality evaluation. Segmentation accuracy for safety equipment and environmental regions shows semantic understanding of low-contrast, multi-scale targets, reducing false alarms in security warnings. Improvements in logistics devices indicate better capture of small targets and edge detail, supporting preventive facility maintenance. Test-set performance close to training-set performance verifies model

robustness across varied campus environments, suitable for large-scale deployment, ensuring cross-scene consistency in scenarios like security alerts and teaching analysis. Diagonal confusion matrix values are all ≥ 0.72, with people-related categories ≥ 0.90, and teaching facilities ≥ 0.88—showing precise segmentation of key campus objects. Through the decoder's Haar wavelet detail retention and the multi-scale module's dilated separable convolution, the model performs excellently on small-target and complex-edge segmentation, addressing traditional model shortcomings in detail loss and class confusion, laying the foundation for pixel-level semantic analysis. Stable test-set performance indicates strong adaptability to unseen campus scenarios, supporting standardized, intelligent deployment for educational management.



(a) Training set



(b) Test set

Figure 6. Confusion matrix

5. CONCLUSION

This study constructed a complete "model optimization–scenario deployment" technical chain around intelligent campus surveillance and educational management demands. In model design, targeting small-target density, complex edges, and edge-device computing constraints in campus scenes, it proposed an improved DeepLabV3+ model integrating MobileNetV3 and Haar wavelet transform: embedding CBAM-enhanced lightweight backbone reduces parameters by 92.3% compared to Xception while enhancing small-target feature extraction; using dilated separable convolution repairs grid defects of standard dilated convolution, improving context aggregation for multi-scale targets; and employing Haar wavelet transform to encode spatial detail into channel features combined with separable convolution delivers lightweight detail enhancement, reducing edge segmentation errors by 22% and increasing small-target recall by 18%. On the application side, it built a three-dimensional scenario system across security management, teaching management, and logistics, forming a "feature extraction–semantic segmentation–intelligent decision" closed-loop solution, offering full-chain support from technology to practice for educational management.

The research results show significant value in merging technology and educational management. Technically, the modular innovation of "attention mechanism + wavelet + separable convolution" achieves the balance of lightweight and high accuracy, breaking through computing bottlenecks for edge-device deployment, and offering a feasible scheme

for large-scale real-time surveillance system upgrades. In educational management, high-precision semantic segmentation enables security alerts, teaching analysis, and logistics optimization, driving management from experience-driven to data-driven. However, limitations remain in dataset coverage, insufficient use of temporal information, and lack of multimodal fusion. Future research will focus on cross-modal data fusion, temporal feature modeling, self-supervised learning optimization, and edge-cloud collaborative architectures to further enhance system robustness and intelligence, providing more comprehensive technical support for smart campus construction and promoting deep digital and precise transformation of educational management.

ACKNOWLEDGMENT

This paper was supported by 2025 Annual Special Project of Philosophy and Social Sciences Research in Shaanxi Province (Youth Special Project), Research on the Cultivation and Enhancement Path of College Students' Artificial Intelligence Literacy under the Background of New Quality Productivity (Grant No: 2025QN0547).

REFERENCES

[1] Caruso, O.T., Schaafsma, H.N., McEachern, L.W., Gilliland, J.A. (2025). The campus food environment and postsecondary student diet: A systematic review. Journal of American College Health, 73(2): 577-601. https://doi.org/10.1080/07448481.2023.2227725

[2] Peachey, A.A., Baller, S.L. (2015). Perceived built environment characteristics of on-campus and off-campus neighborhoods associated with physical activity of college students. Journal of American College Health, 63(5): 337-342. https://doi.org/10.1080/07448481.2015.1015027

[3] Ramakreshnan, L., Fong, C.S., Sulaiman, N.M., Aghamohammadi, N. (2020). Motivations and built environment factors associated with campus walkability in the tropical settings. Science of the Total Environment, 749: 141457. https://doi.org/10.1016/j.scitotenv.2020.141457

[4] Kim, S., White, H.H., Museus, S.D. (2024). Examining the relationship between culturally engaging campus environments and civic propensity among diverse college students. Journal of Diversity in Higher Education. https://doi.org/10.1037/dhe0000633

[5] Khan, K.R., Saawy, Y., Rahman, A., Siddiqui, M.S., Eskandrany, A.O. (2019). Condition monitoring and control of a campus microgrid elements. International Journal of Computer Science and Network Security, 19(2): 155-162.

[6] Villegas-Ch, W., Molina-Enriquez, J., Chicaiza-Tamayo, C., Ortiz-Garcés, I., Luján-Mora, S. (2019). Application of a big data framework for data monitoring on a smart campus. Sustainability, 11(20): 5552. https://doi.org/10.3390/su11205552

[7] Zhang, J., Liu, Y., Guo, C., Zhan, J. (2023). Optimized segmentation with image inpainting for semantic mapping in dynamic scenes. Applied Intelligence, 53(2): 2173-2188. https://doi.org/10.1007/s10489-022-03487-3

[8] Shu, R., Zhao, S. (2024). Multi-resolution learning and semantic edge enhancement for super-resolution semantic segmentation of urban scene images. Sensors, 24(14): 4522. https://doi.org/10.3390/s24144522

[9] Yan, B., Niu, X., Bare, B., Tan, W. (2019). Semantic segmentation guided pixel fusion for image retargeting. IEEE Transactions on Multimedia, 22(3): 676-687. https://doi.org/10.1109/TMM.2019.2932566

[10] Jiang, B., An, X., Xu, S., Chen, Z. (2023). Intelligent image semantic segmentation: A review through deep learning techniques for remote sensing image analysis. Journal of the Indian Society of Remote Sensing, 51(9): 1865-1878. https://doi.org/10.1007/s12524-022-01496-w

[11] Friedman, N.M., O'Connor, E.K., Munro, T., Goroff, D. (2019). Mass-gathering medical care provided by a collegiate-based first response service at an annual college music festival and campus-wide celebration. Prehospital and Disaster Medicine, 34(1): 98-103. https://doi.org/10.1017/S1049023X18001103

[12] McCarthy, J.D., Martin, A., McPhail, C. (2007). Policing disorderly campus protests and convivial gatherings: The interaction of threat, social organization, and First Amendment guarantees. Social Problems, 54(3): 274-296. https://doi.org/10.1525/sp.2007.54.3.274

[13] Ueki, R. (2004). Ideal ways to teach students how to utilize self-monitoring strategies: Beliefs about learning and knowledge about strategies. Japanese Journal of Educational Psychology, 52(3): 277-286.

[14] Barbosa, J.W.D.Q., Belchior, M.H.C.D.S. (2021). What do students look for when doing monitoring? A case study in the field of teaching hotel management. Rosa dos Ventos-Turismo e Hospitalidade, 13(4): 1152-1173. https://doi.org/10.18226/21789061.v13i4p1173

[15] Ul Haq, N., Ur Rehman, Z., Khan, A., Din, A., Shah, S., Ullah, A., Qayum, F. (2022). Impact of data smoothing on semantic segmentation. Neural Computing & Applications, 34(11): 8345-8354. https://doi.org/10.1007/s00521-020-05341-4

[16] Ryu, K.B., Kang, S.J., Jeong, S.I., Jeong, M.S., Park, K.R. (2024). CN4SRSS: Combined network for super-resolution reconstruction and semantic segmentation in frontal-viewing camera images of vehicle. Engineering Applications of Artificial Intelligence, 130: 107673. https://doi.org/10.1016/j.engappai.2023.107673

[17] Jaimes, B.R.A., Ferreira, J.P.K., Castro, C.L. (2021). Unsupervised semantic segmentation of aerial images with application to UAV localization. IEEE Geoscience and Remote Sensing Letters, 19: 8020405. https://doi.org/10.1109/LGRS.2021.3113878

[18] Sehar, U., Naseem, M.L. (2022). How deep learning is empowering semantic segmentation: Traditional and deep learning techniques for semantic segmentation: A comparison. Multimedia Tools and Applications, 81(21): 30519-30544. https://doi.org/10.1007/s11042-022-12821-3

[19] Mohammed, S.A., Ralescu, A.L. (2024). Insights into image understanding: Segmentation methods for object recognition and scene classification. Algorithms, 17(5): 189. https://doi.org/10.3390/a17050189

[20] Pereira, R., Barros, T., Garrote, L., Lopes, A., Nunes, U.J. (2024). A deep learning-based global and segmentation-based semantic feature fusion approach for indoor scene classification. Pattern Recognition Letters, 179: 24-30. https://doi.org/10.1016/j.patrec.2024.01.022