JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-21, No.-06, June (2026) pp 1-21 ISSN (Print) 0973-8975

COMPUTER VISION METHODOLOGY FOR DETECTING AND CLASSIFYING DEFECTS IN FACADES OF CIVIL STRUCTURES OF THE KYRGYZ REPUBLIC

Lei Xin[1], Zhao Chen[2], Zhenish T. Beksultanov[3], Fu Zh. Li[4]
Gulmira Isaeva[5]

[1, 2, 3, 5]Kyrgyz National University named after Jusup Balasagyn, Bishkek, Kyrgyzstan.

[4]Shanxi Agricultural University, Shanxi, China.

Email : [1]leixin0914@mail.com, [2]1254300188@qq.com, [3]jenish1979@mail.ru, [4]lifuzhong@sxau.edu.cn, [5]Gu_Isaeva@proton.me

Corresponding Author: **Gulmira Isaeva**

https://doi.org/10.26782/jmcms.2026.06.00001

(Received: March 26, 2026; Revised: May 27, 2026; Accepted: June 10, 2026)

Abstract

The predominantly poor condition of the existing multi-apartment housing stock and a sharp increase in the number of new buildings in the Kyrgyz Republic require the introduction of modern methods of automated monitoring and control. The purpose of this research is to develop a method for the rapid detection and classification of defects on the facades of civil buildings and structures. The methodology of the research comprises methods of structural analysis, manual annotation of graphical data, deep learning, and statistical analysis. These methods enabled the definition of facade defects and their classification relative to the residential building model, the development of an original dataset with an array of images with characteristic defects, and the selection and training of a convolutional neural network model. The article presents the characteristics of the model and the statistical data of its operation when processing test data. The results demonstrate high accuracy and performance, which makes it possible to use the proposed model to create a digital ecosystem for urban management. Such data can be used for automated management, planning services for repair, and improving the efficiency of buildings and structures.

Keywords: convolutional neural network, semantic segmentation, construction, digital transformation, smart city, urbanization.

I. Introduction

Computer vision (CV) is a technology enabling digital systems to locate, track, classify, and identify various objects by extracting a set of characteristics from images and analyzing the obtained data [XI]. CV makes it possible to reconstruct the shape and size of objects from a series of photographs or videos, as well as to determine the distance and the sequence in which they are located relative to certain points. In this way, a dense depth map can be formed, which determines the points of contact and evaluates the material of the objects and their physical properties by visual features [IV].

Depending on the tasks addressed, CV systems can incorporate cameras, processors, sensors, light sources, and specialized software [XI]. When combined with simulation models, this technology enables not only image capture but also the determination of subsequent changes in the characteristics or behavior of the observed object [XVII, XIII]. Modern algorithms can take into account the propagation and multiple reflection of light, as well as correctly process glare and flicker, which increases the accuracy of segmentation and measurement of objects under varying illumination conditions [IX].

According to analytical reports, the global CV market is growing by 12–19% annually [XVIII]. Hardware and software incorporating this technology are actively used in industry and manufacturing (26%), healthcare (20%), security and video surveillance (19%), retail and trade (9%), transport and logistics (5%), agriculture (2%), and construction (2%) [XXIII].

The development of the construction industry is associated with digital transformation processes characterized by accelerated dynamics [X, XXI]. This is due to the constantly increasing pace of construction and stricter requirements for the quality and safety of structures, as well as the introduction of new materials and technologies, which entail the optimization of processes at all stages of construction. In addition, continuous monitoring of the condition of buildings and structures is required.

Monitoring the condition of the housing stock is a pressing issue for the Kyrgyz Republic. The massive outflow of population from rural regions to large cities (such as Bishkek and Osh) creates excessive pressure on the housing stock and leads to the proliferation of residential areas [XIX]. In this light, challenges arise regarding the deterioration of utility networks, reduced quality of public services, and a high proportion of housing stock in a dilapidated or worn condition [XX]. Additional negative factors affecting the condition of buildings include high seismicity and a sharply continental climate [I]. All this increases the cost of high-quality construction and maintenance, while accelerating the degradation processes of buildings and structures.

The purpose of this research is to develop and adapt a method for the automated monitoring of the condition of facades of civil buildings in the Kyrgyz Republic for the rapid detection, classification, and rectification of defects.

This requires fulfilling the following tasks: analyzing the subject area; collecting and annotating an array of images with examples of damage typical of local buildings; selecting and training a neural network model based on the collected and processed data; and testing the system.

The scientific novelty of the research lies in the development of an original method designed to effectively detect and classify defects on the facades of civil structures. A distinctive feature is the use of deep learning algorithms adapted to the specific survey conditions of urban objects, including complex lighting, decorative elements, and architectural features of buildings conditioned by their territorial and climatic location.

The theoretical significance of the research lies in creating a unique dataset characterizing the state of facades of civil structures in the Kyrgyz Republic. The results can be used in research and development intended to rank the threat in accordance with the regulatory framework, as well as to identify workflows and materials required to correct the detected defects. All this can serve as a basis for developing predictive models for the degradation of buildings and structures in the Kyrgyz Republic.

The practical significance of the research lies in providing results that optimize the city budget, namely, enabling the rational allocation of funds for the overhaul and restoration of facades.

II. Materials and methods

The object of this research is the operational and monitoring processes of the facades of civil buildings and structures in the Kyrgyz Republic.

The subject of this research comprises the methods, algorithms, and software tools designed for the automated detection, localization, and classification of defects.

To achieve the established objectives, the following methods were employed:

1. Structural analysis and synthesis: These methods made it possible to define the main classes of defects found on building facades, their hazard levels, and the actions necessary to rectify them. Based on this classification, photographic and video materials demonstrating the relevant types of defects were collected to form the dataset.
2. Manual annotation of graphical data: Using specialized software, the collected dataset was processed to obtain the ground-truth annotations. The labeling pipeline included the following steps: creating a dictionary of defects (class identification), importing data from the dataset, visually identifying damage zones in each image, delineating regions of interest and assigning them appropriate defect classes (creating annotations), verifying the results to minimize subjective bias, and exporting the data for neural network training.
3. Deep learning-based instance segmentation: The YOLOv8-seg model was employed, which utilizes multi-scale feature fusion architectures combining bottom-up and top-down pathways to accurately determine mask boundaries. The model generates a set of prototype masks for the entire image in parallel while predicting unique weighting coefficients for each detected object. As a result, the final mask is computed by linearly combining the prototype masks based on these coefficients, followed by clipping the predicted bounding box boundaries.
4. Statistical and graphical analysis: A combination of metrics and corresponding visualizations was used to assess training dynamics, performance (precision, recall, F1-score, and a confusion matrix), and classification quality via precision-recall curves across varying confidence thresholds.

III. Results

Formalization of the object and subject of research

The level of urbanization in Kyrgyzstan is about 38% and is characterized by its heterogeneity [XIX]. Due to the growth of investments in housing construction, luxury housing is dynamically emerging in the large cities of Bishkek and Osh. At the same time, Soviet-era apartment buildings predominate in the country's urban housing stock. This array of buildings can be divided into groups:

- brick houses with high ceilings and thick walls (“stalinki”): good sound insulation, worn-out utilities, decorative facades;
- typical panel or brick four- or five-story houses built in the 1960s and 1970s (“khrushchevki” and “individual apartments”): small kitchen area, combined bathrooms, facing with small ceramic tiles, painting or uncovered concrete, no architectural frills;
- series 104 panel houses: reinforced concrete with mineral insulation or brick in improved versions, poor heat insulation, loggias, small indoor areas, 4–5 stories, no architectural frills;
- series 105 panel houses (mass-produced): improved heat engineering, earthquake resistance, and layout, 5 and 9 stories, possible loggias in some apartments, no architectural frills;
- series 106 panel houses: improved version of 105 series, 9 stories, loggias.

Private residential construction dominates the housing sector in rural areas. It is primarily represented by single-story detached houses with attached verandas. The materials used in construction are adapted to the mountainous terrain, climate, and economic considerations. The most common wall material is adobe (sun-dried earth brick), which is widely utilized due to its cost-effectiveness and high thermal inertia, ensuring favorable indoor climate conditions throughout the year. In areas with high humidity, burnt bricks and concrete are typically used for the foundations and walls of modern buildings. Bituminous shingles or metal tiles are used for roofing, predominantly configured as gabled structures to facilitate effective precipitation drainage.

The main facade defects were identified by analyzing the state of the housing stock in the Kyrgyz Republic and the technical documentation regulating the construction and maintenance of buildings and structures. Their classification is illustrated in Fig. 1.

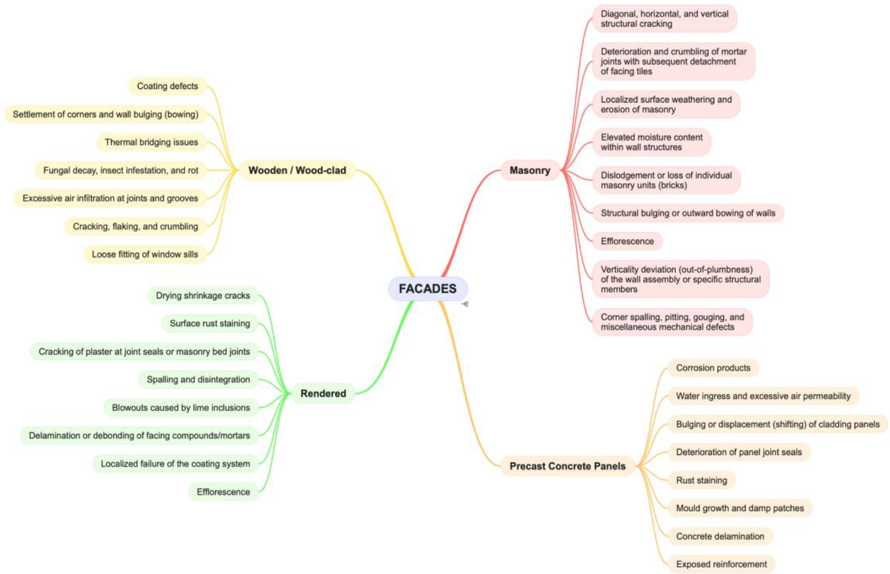


Fig. 1. Classification of defects in facades of civil buildings and structures according to the material used

The developed defect classification was used as the basis of the dataset required for training the neural network model. The following open resources were used as the source of the images:

- Google Maps (for example, Bishkek: https://maps.app.goo.gl/7wvVycpGV85E39gBA): street panoramas of localities and photos posted by users with geolocation;
- Yandex maps (for example, Taraz: https://yandex.ru/maps/-/CPFLeQ4b): street panoramas of localities and photos posted by users with geolocation;
- News channels (for example, @news24kg,@omks_media) that publish materials about incidents and activities of local officials and contractors for the repair or reconstruction of buildings and structures, including photos and videos;
- Reports and publications of the Ministry of Construction, Architecture, and Housing and Public Services of the Kyrgyz Republic (https://minstroy.gov.kg/ru) and its relevant departments or functions.

The listed resources publish free materials, without personal data or data that does not imply a conflict of interest or breach of existing laws.

The total dataset comprised 7300 images showcasing both isolated and combined facade defects. The annotated objects included 3349 instances of "Lack of cladding," 3336 of "Mold," and 3315 of "Paint shedding." Data splitting into training, validation, and test sets was performed on a per-file basis rather than by individual annotations. This approach prevents identical images from simultaneously appearing in both training and validation sets, which would artificially inflate validation accuracy and reduce model generalizability on independent data. Consequently, images were first grouped by their defect proportions and then distributed into target directories using a 70% / 15% / 15% ratio.

The following criteria were established for image selection during dataset compilation:

- Temporal and Luminous Ranges. Data collection was conducted during morning, afternoon, and evening hours. This enabled the model to recognize objects under both intense direct illumination with a high color temperature and the "golden hour" period, which is characterized by long shadows and a low angle of incidence.
- Meteorological Conditions. Approximately 30% of the sample comprises images captured in bright sunlight (with pronounced glare and deep shadows), while 20% were taken during precipitation, fog, or haze (characterized by optical distortions and soft, diffused light).
- Scale and Shooting Distance. To facilitate multi-level analysis, the sample was partitioned as follows: 20% close-up shots for maximum texture detail; 30% medium shots (wall segments and window openings) to localize defects relative to structural elements of the facade; and 50% wide shots (entire buildings and adjacent infrastructure objects) to provide architectural context.
- Structural Context. Building elements (windows, doors, plinths, and drainage systems) were intentionally included in the frame. This was necessary to minimize false-positive model detections triggered by the natural texture of brickwork or decorative joints.
- Camera Angle and Geometry. The dataset includes images taken both perpendicular to the wall plane and at an angle, including low-angle perspective views with geometric distortions (e.g., optical axis tilt).

The following rules were defined to construct class-specific samples for particular defect categories:

- Missing Cladding. For accurate identification, the depth of the spall had to be visually discernible. Preference was given to images with clearly defined boundaries between the intact covering (tile, plaster) and the exposed substrate (brick, concrete). Flat, monochrome gray areas showing no visible signs of deterioration were excluded.
- Mold and Fungus. The selection criterion was the presence of a color gradient (transitioning from black to green or gray shades) in potential moisture accumulation zones: corners, under-cornice spaces, and areas near downspouts. Images featuring only natural geometric shadows were excluded.
- Peeling Paint. Priority was given to images with pronounced boundaries of detached coatings, cracks, and flaking surfaces. Uniformly painted walls with only surface contamination were excluded from the analysis.

The heterogeneity of the data sources leads to a domain shift issue caused by variations in color reproduction, contrast, sharpness, and sensor-induced optical distortions. To mitigate this effect at the intersection of data preprocessing and model architecture design, a multi-level domain normalization framework was implemented, focusing specifically on the data preparation and augmentation stages.

At the data preparation level, the primary objective was to align the luminance and contrast histograms of all input images to a unified distribution, thereby eliminating pronounced technical discrepancies between cameras. Panoramic images from Yandex and Google mapping services, characterized by neutral color reproduction, served as the reference domain. All other images were transformed using Histogram Matching to align the probability density distribution of their color channels (R, G, B) with the reference. Each image underwent standard score normalization using the formula $z = (I - \mu) / \sigma$, where I is the original image, μ is the mean pixel intensity, and σ is the standard deviation. This process neutralized the variance between overexposed unmanned aerial vehicle (UAV) imagery and underexposed amateur photographs. Additionally, the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm was applied to locally enhance the textural features of the defects. This measure ensured robust defect boundary extraction regardless of the dynamic range and exposure level of the original frame.

At the augmentation level, artificial degradation of the source domains was performed to compel the neural network to extract domain-invariant defect features. To achieve this, filters simulating various types of distortions were applied to the panoramic and UAV images: digital noise characteristic of mobile sensors, motion blur from handheld shooting, and JPEG compression artifacts typical of open-source internet imagery. To prevent the model from establishing a spatial-textural dependency on specific color shades of brickwork or plaster, extreme random jitter was applied in the HSV color space, with shifts of ±5% for Hue, ±80% for Saturation, and ±60% for Value.

Characterization of the model in terms of the object and the subject of the research

The YOLOv8-seg model, which is a single-stage detector adapted for segmentation functions, was used in the research. It consists of blocks whose functions are presented in Table 1.

Table 1: Model architecture

Name	Description	Operations used	Functional purpose
Input	Initial feature extraction, forming a basic representation of the input image	Images input tensor [B, 3, H, W] Convolutional layers: model1.0 and model1.1 Activation operations: conv, sigmoids, and mul (SiLU)	Primary extraction of low-level features without significant spatial compression Forming a fundamental feature map for further processing
Backbone	Hierarchical feature extraction	C2f blocks: layers 2, 4, 6, 9. Operations: split (tensor streaming), Conv-SiLU (transformations), add (residual connections), concat (feature aggregation) Subdiscretization	Creating multiple additional connections between layers to preserve important small details of objects Feature maps are generated with progressively decreasing spatial resolution and an increasing number of channels
Neck	Forming a pyramid of features	Top-down path (FPN): operations of spatial magnification (Resize), merging features from different levels (Concat)	Transferring deep-layer context to shallow layers to enrich high-resolution features
		Bottom-up path (PAN): using convolutional layers to reduce dimensionality and combine features (Concat)	Transfer of precise localized information from shallow layers to deep layers to determine the precise positioning of object boundaries
Detection head	Regression and classification	Dimensionality change operations (Reshape and Transpose), combining to form the final tensor (Concat)	Predictions include bounding box regression and object classification probabilities
Mask coefficients head	Generating mask coefficients	Feature extraction (Conv) and multiplication (Mul) operations. The number of mask prototypes is 32	Coefficients generated per object linearly combine with prototype masks for the final mask
Prototype mask network	Generating prototype masks for a prototype network	Convolutional layers (Conv → Mul → Sigmoid). Resolution enhancement operations (ConvTranspose). Formation of final convolutions	Generating a set of basic masks that form a space of possible segment shapes and have high spatial resolution without reference to specific objects
Distribution focal loss	Post-processing inside the graph	Operations: Conv, Reshape, Transpose.	Predicting the probability distribution of boundary

decoding block		Auxiliary operations: Range, Slice, Sub, Mul. Formation of final coordinates	positions for localization accuracy
Auxiliary and utility operations block	Adapting to various input image sizes	Dynamic tensor formation based on Shape, Gather, Cast, Range, Unsqueeze, ConstantOfShape	They do not extract features, but ensure the correct operation of the dynamic graph.

The structure of the neural network model contains 356 layers. The distribution of their operations is shown in Table 2.

Table 2: Distribution of operations in the neural network model

Operation	Quantity	Share, %	Purpose
Conv	76	21.3	Basic feature extraction unit (convolution, BatchNorm, SiLU)
Mul (SiLU)	69	19.4	Activation function
Sigmoid	66	18.5	Sigmoid for activation function and gates in C2f
Concat	34	9.6	Combining tensors in C2f, FPN, and PAN modules
Reshape	20	5.6	Changing the shape of tensors for the detection head

The high number of Concat operations corresponds to an architecture with aggregation of features from different branches.

Model characteristics

Here are the characteristics of model training on a dataset containing images of facades with marked defects, using the example of three classes: Paint Shedding, Mould, and Lack of Cladding. To speed up learning and preserve low-level features, the first 10 backbone layers were frozen (*freeze* = 10), and the Weight Decay regularization was set to 0.01. The Stochastic Gradient Descent optimizer corresponds to a small initial learning rate for stability (lr_0 = 0.0001) and the final speed taking into account the reduction factor (lr_f = 0.001). The basic training parameters of the model include the following values:

- 1000 epochs;
- batch size: 1;
- input image size: 640×640;
- 0 work processes;
- The capability of accelerated CPU usage.

Table 3 shows the basic parameters of augmentation.

Table 3: Basic augmentation parameters

Parameter	Value	Range	Characteristic	Justification
Color				
hsv_h	0.05	±5% (≈18°)	Shade	Simulation of time of day and types of lighting
hsv_s	0.8	±80%	Saturation	Simulation of weather conditions
hsv_v	0.6	±60%	Brightness	Imitation of shadows, glares, and

				exposure
Geometric				
degrees	45.0	±45°	Image rotation	Canted shot or camera tilt
translate	0.2	±20%	Image shift	Object displacement in frame
scale	0.8	0.2x–1.8x	Scaling	Difference in distance to the object
shear	10.0	±10°	Slant	Camera tilt, perspective effect
perspective	0.001	0–0.001	Perspective distortions	Wide-angle lenses, bottom view
Spatial				
flipud	0.0	0%	Vertical flip	Inactive to preserve building facade semantics
fliplr	0.5	50%	Horizontal flip	Symmetry of buildings relative to the vertical axis

The characteristics of the tensors are presented in Table 4.

Table 4: Characteristics of the tensors

Parameter	Value
Input tensor	
Tensor	images
Format	NCHW (Batch, Channels, Height, Width)
Dimension in the Open Neural Network Exchange	[0, 3, 0, 0] – dynamic
Expected dimension	[1, 3, 640, 640]
Data type	FP32
Normalization	Division by 255.0 [0.0, 1.0]
Output tensors	
output0	Bbox coordinates (4), class probabilities (*num_classes*), mask coefficients (32). Dimension: [1, 40, N], where N is the number of predictions.
output1	Mask prototypes (32 feature maps) are [1, 32, H', W'], where H', W' are spatial dimensions.

The standard segmentation format, YOLO, was used to train the model. It assumes that for each image there is a text file with markup, where each line describes one object <class_id> <x_1> <y_1> <x_2> <y_2> ... <x_n> <y_n>. Accordingly, the coordinates of the polygon (x, y) are normalized relative to the width and height of the image from 0 to 1. Fig. 2 shows a fragment of such a markup file for an image belonging to the Paint Shedding class.

```
0 0.632338 0.532819 0.631308 0.534106 0.627188 0.534106 0.626159 0.535393 0.624099 0.535393 0.623069 0.536680
0.622039 0.536680 0.621009 0.537967 0.615860 0.537967 0.614830 0.539254 0.612770 0.539254 0.611740 0.540541
0.609681 0.540541 0.608651 0.541828 0.607621 0.541828 0.606591 0.543115 0.605561 0.543115 0.604531 0.544402
0.603502 0.544402 0.601442 0.546976 0.600412 0.546976 0.596292 0.552124 0.595263 0.552124 0.594233 0.553411
0.589083 0.553411 0.588054 0.554698 0.585994 0.554698 0.584964 0.555985 0.583934 0.555985 0.582904 0.557272
0.581874 0.557272 0.580844 0.558559 0.576725 0.558559 0.575695 0.559846 0.574665 0.559846 0.573635 0.561133
0.572606 0.561133 0.571576 0.562420 0.570546 0.562420 0.569516 0.563707 0.568486 0.563707 0.567456 0.564994
0.566426 0.564994 0.565396 0.566281 0.564367 0.566281 0.563337 0.567568 0.562307 0.567568 0.561277 0.568855
0.560247 0.568855 0.554068 0.576577 0.554068 0.577864 0.548919 0.584299 0.548919 0.585586 0.547889 0.586873
0.547889 0.588160 0.546859 0.589447 0.546859 0.590734 0.545829 0.592021 0.538620 0.592021 0.537590 0.590734
0.479918 0.590734 0.478888 0.589447 0.476828 0.589447 0.475798 0.588160 0.473738 0.588160 0.472709 0.586873
0.469619 0.586873 0.468589 0.585586 0.437693 0.585586 0.436663 0.586873 0.433574 0.586873 0.432544 0.588160
0.429454 0.588160 0.428424 0.589447 0.426365 0.589447 0.425335 0.590734 0.424305 0.590734 0.423275 0.592021
0.422245 0.592021 0.418126 0.597169 0.416066 0.597169 0.415036 0.598456 0.410917 0.598456 0.409887 0.599743
0.407827 0.599743 0.406797 0.601030 0.404737 0.601030 0.403708 0.602317 0.402678 0.602317 0.401648 0.603604
0.400618 0.603604 0.399588 0.604891 0.397528 0.604891 0.396498 0.606178 0.395469 0.606178 0.394439 0.607465
```

Fig. 2. Fragment of the contents of the image markup file for training the model

For simultaneous prediction of frames and masks, all points of the polygon are covered as a minimum bounding rectangle for training and validation.

Performance characteristics

The final model takes about 15 MB. The average inference time is 49.2 ms, which provides a processing speed of ~20 FPS.

Fig. 3 shows the error matrix: how often does the model confuse classes with each other.

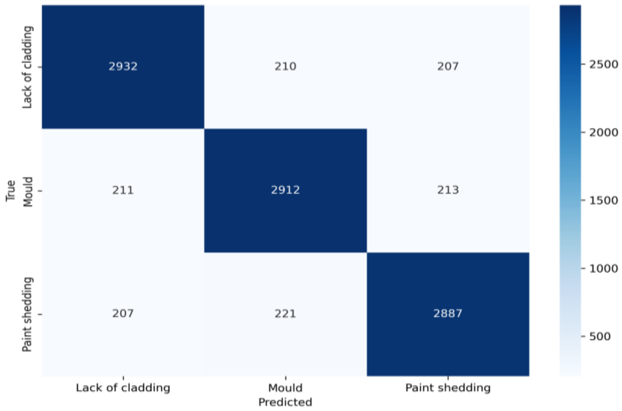


Fig. 3. Confusion matrix

Based on the confusion matrix analysis (Fig. 3), the performance metrics of the developed model were determined at the segmentation object level for each target class of facade construction defects. The quantitative evaluation results are presented in Table 5.

Table 5: Object-level performance metrics for building facade defect segmentation

Defect classes	Precision, %	Recall, %	Dice / F1, %	IoU, %
Lack of cladding	87.52	87.55	87.54	77.83
Mould	87.11	87.29	87.20	77.30
Paint shedding	87.30	87.09	87.19	77.30
Mean	87.31	87.31	87.31	77.48

An analysis of the results reveals that out of 3349 true instances of the "missing cladding" class, the algorithm made 417 Type II errors (false negatives, FN), while 418 false positives (FP) were recorded. The model achieved its highest IoU value (77.83%) specifically for this defect type, which is attributed to the presence of sharper and more distinct geometric boundaries in tile and plaster spalls.

In the "mold and fungus" category, the algorithm correctly identified 2,912 instances out of 3336 true objects (true positives, TP). Type I errors (FP = 431) were predominantly caused by the high textural and spectral visual similarity between biological degradation areas and deep shadows or localized regions of peeling paint. The F1-score (Dice coefficient) for this class reached 87.20%, confirming a high balance between precision and recall in detecting biological damage. For the "peeling paint" defect class, the algorithm successfully localized 2887 objects out of 3,315 true instances (TP). This class exhibited the lowest recall value (Recall = 87.09%), which is explained by the high complexity of detecting small-scale flaking scales in wide shots. Nevertheless, the IoU metric was recorded at 77.30%, demonstrating a stable segmentation quality comparable to the other classes.

The discrepancy between the metric values across all three investigated classes does not exceed 0.53%. This fact mathematically confirms both the representativeness and balance of the initial dataset, as well as the absence of a bias in the predictive power of the model toward a dominant class. The Dice coefficient F1-score naturally exceeds the IoU metric because the Jaccard index imposes stricter penalties on any spatial pixel discrepancies at the boundaries of the segmentation masks. The achieved performance level (IoU > 75%) indicates the applicability of the developed method for integration into industrial automated monitoring systems for housing and communal utility infrastructure.

Fig. 4 shows a curve (ROC Curve) demonstrating the relationship between the True Positive Rate (Recall) and False Positive Rate metrics.

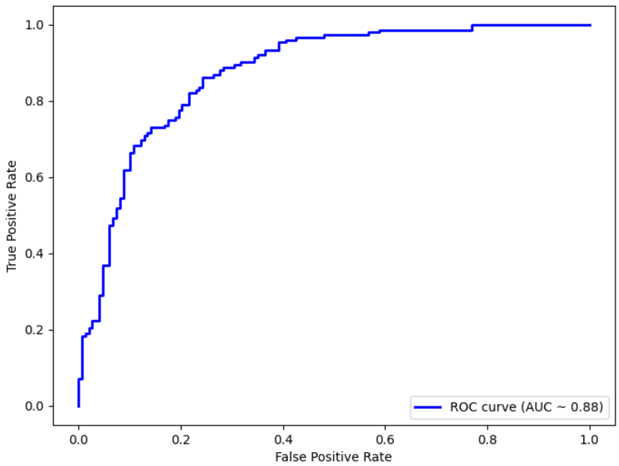


Fig. 4. ROC Curve graph

Figs. 5 to 8 show the changes in the quality metrics of the model in the train and val samples.

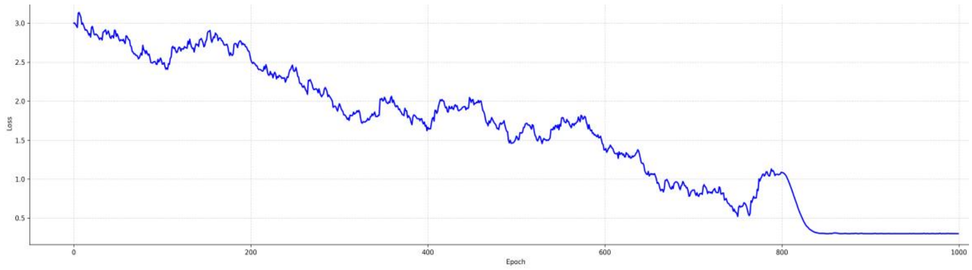


Fig. 5. Graph of changes in the neural network error parameter when determining the coordinates and sizes of the bounding rectangles around objects at the training stage (Train Box Loss)

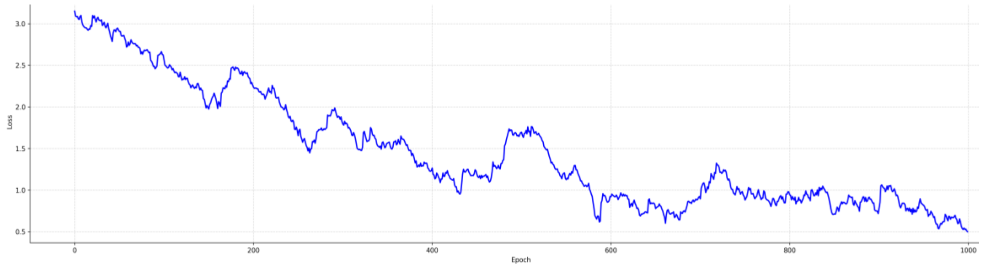


Fig. 6. Graph of changes in the neural network error parameter when learning a segmentation task (Train Seg Loss)

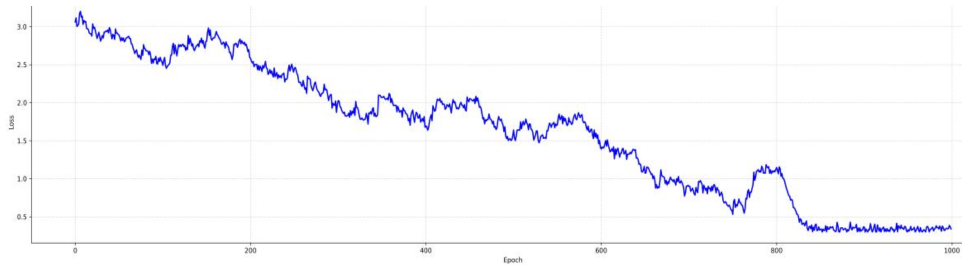


Fig. 7. Graph of changes in the error parameter when determining the coordinates of the frames in a valid sample (Val Box Loss)

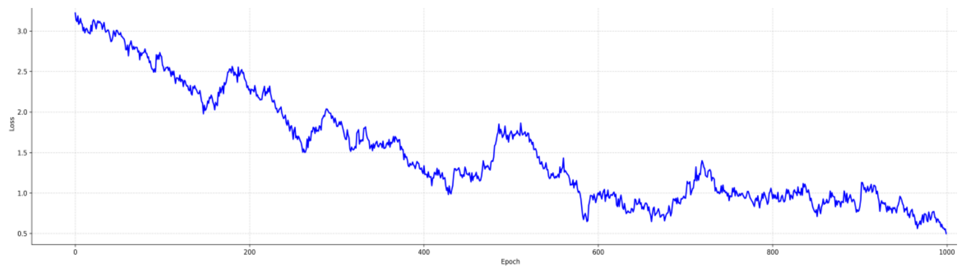
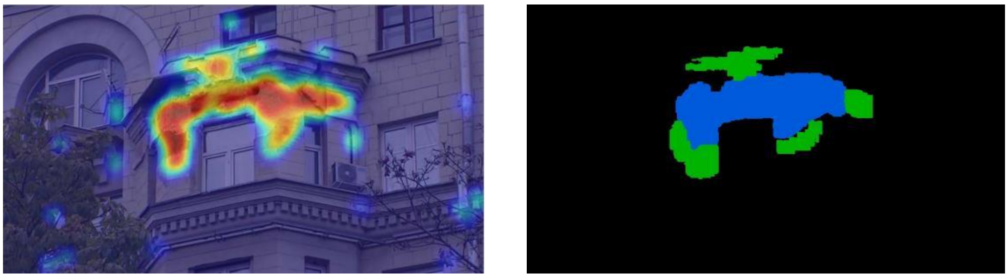


Fig. 8. Graph of changes in the segmentation error parameter in a valid sample (Val Seg Loss)

The integral mean average precision (mAP@0.5) of the developed model is 91.2%. This high performance is attributed to the fact that at an Intersection over Union threshold of IoU = 0.5, the algorithm successfully compensates for minor geometric errors, such as the spatial instability of segmentation mask boundaries at the edges of peeling paint or biological degradation zones. Concurrently, the model consistently localizes the core defect area and correctly identifies its semantic class. The developed algorithm also exhibits high robustness against false positives on complex background textures of facade constructions, mathematically confirming its applicability for integration into automated monitoring systems for urban infrastructure.

In addition to the graphs, a visual analysis of the inference on the test image was performed to evaluate the results. Figs. 9–10 show the corresponding results.



a) b)

Fig. 9. Facade image with decorative architectural forms: (a) heat map and (b) binary mask of the test image

a) b)

Fig. 10. Results of the image processing of the facade with decorative architectural forms: (a) intermediate and (b) final results of the model

The image regions that make the greatest contribution to the model's decision-making process are visualized using a heat map (Fig. 9). The areas highlighted in red indicate high-intensity regions, which correspond to the areas where the model detected facade defects. A binary mask against a black background displays the segmented object instances according to the defined classes (green for paint peeling, blue for missing cladding, and yellow for mold) without background context. The final step involves generating an image with bounding boxes around the detected defects, accompanied by class labels and their corresponding confidence scores (Fig. 10).

To compare the accuracy of the defect detection results, we will show the processing of low-quality images and the presence of objects that visually resemble the analyzed classes (Fig. 11).

a)

b)

Fig. 11. Test set processing option: (a) binary mask of the facade of a building with mold and (b) final result: defect mask and confidence in prediction

To determine defects on the facade of buildings, the level of demonstrated confidence in the prediction on the test sets is high. It can be considered reliable, according to the available training statistics.

For comprehensive analysis scaling, a differentiated metric interpretation approach is introduced in Table 6.

Table 6. Specialized metrics for evaluating contours of low-contrast and fragmented defects

Metric	Target Defect Type	Primary Control Objective	Critical Threshold, %
Standard IoU	All types of macro-defects	Detecting defect presence and estimating the total area of building facade damage	≥75
Boundary IoU	Large spalls (Lack of cladding), continuous mold growth (Mould)	Monitoring edge truncation and geometric smoothing caused by neural network layer interpolation	≥55
BF-score	Hairline cracks (Paint shedding), gradient-blurred water stain zones	Evaluating contour positioning accuracy under low-contrast conditions	≥65

Utilizing this extended mathematical framework enables a transition from baseline defect logging to evaluating degradation dynamics. For instance, a drop in the BF-score below 65% despite a high overall Standard IoU for the "Paint shedding" class serves as an automated trigger indicating that the paint is actively flaking along the edges. This signals the engineering system that localized repainting is unfeasible, necessitating continuous deep scraping of the facade substrate.

During the training of the semantic segmentation model for facade construction defects, issues related to an imbalance in segmentation quality were identified and mitigated. At the macro level, the compiled dataset is balanced in terms of object counts per class; however, micro-level analysis reveals critical data distribution characteristics that directly degrade model performance metrics.

The first key issue was an inter-class geometric scale imbalance. Although the number of representative instances for each class is approximately equal, their cumulative pixel area varies drastically. For example, mold and fungus colonies often propagate as large, continuous clusters at wall junctions, whereas peeling paint localizes as hundreds of tiny cracks or flaking scales. This imbalance results in the systematic under-segmentation of small-scale defect boundaries and the truncation of their true contours.

The second destructive factor is the spatial imbalance between the target classes and the background. Even in images of heavily damaged buildings, target defects occupy no more than 5–15% of the total frame area, while the remainder is classified as background–including undamaged wall sections, window openings, sky, or vegetation. This ratio triggers an increase in false negatives and Type II errors. By minimizing the loss function on the dominant background class, the segmentation algorithm shrinks the predicted defect mask to a minimum or ignores it entirely to avoid penalties for false-positive predictions on background pixels.

Finally, the third issue is an imbalance in distribution density and boundary gradients. This problem manifests as spatial instability of the segmentation mask across frames, where the model approximates lighting changes or local shadow geometry rather than the true defect boundary. This leads to artifacts and reduced smoothness in the predicted contours.

Benchmarking and Resource Consumption Metrics of the Neural Network Model

To evaluate the scalability of the developed automated facade monitoring method, Table 7 presents the key performance metrics depending on the deployment platform.

Table 7: Summary of Performance and Resource Consumption Metrics

Hardware Platform	Optimization	Inference Latency, ms	Throughput (FPS)	Memory Consumption	Power Consumption / TDP, W
Edge CPU (UAV / Drones) (e.g., Raspberry Pi 5)	Vanilla PyTorch / ONNX	~49.2	~20.3	~450 MB (System RAM)	9–12
Edge AI Module (NVIDIA Jetson Orin Nano 4GB)	TensorRT (FP16)	~14.1	~71.0	~1.8 GB (Unified Memory)	7–15
Edge AI Computer (NVIDIA Jetson Orin NX 16GB)	TensorRT (FP16)	~7.8	~128.0	~2.1 GB (Unified Memory)	15–25
Cloud Server (NVIDIA A100 / H100 GPU)	TensorRT (FP16 / INT8)	< 1.5	> 650.0	~32 MB VRAM (per instance)	250–350

The baseline YOLOv8-seg configuration requires approximately 12.5 GFLOPs, making it suitable for real-time execution even on general-purpose mobile processors. Basic convolution blocks (Conv) account for the largest share of computational operations (21.3%). The high proportion of concatenation operations (Concat, 9.6%) confirms the intensive use of an architecture featuring feature aggregation from different branches (FPN/PAN modules in the Neck portion). This imposes increased demands on memory bandwidth rather than just pure processor computational power.

Jetson platforms utilize unified memory. When compiling the model into the TensorRT FP16 engine, memory consumption increases to approximately 1.8–2.1 GB. This is due to the loading of CUDA libraries and the buffering of intermediate tensors; however, it is fully justified by a multiple-fold increase in FPS.

In the cloud (discrete GPUs), the direct allocation of model weights occupies a minimal volume (around 32–45 MB of VRAM). The majority of the server memory is consumed by creating batch queues (Batch Size > 1), which enables the parallel processing of video streams from hundreds of drones.

When deploying on UAVs, the power consumption of the onboard computer directly affects the drone's flight time. The use of TensorRT optimization on Jetson Orin series devices ensures best-in-class energy efficiency (up to 5–7 FPS per 1 W of consumed power), minimizing battery drain during round-the-clock urban monitoring missions.

IV. Discussion

Modern software is becoming increasingly intelligent, adaptive, and architecturally and organizationally complex [VI]. As noted in the literature, the architectural complexity of such systems is characterized by a transition from deterministic hierarchical structures to nonlinear distributed ecosystems, driven by the growing number of microservices used to comprehensively solve diverse tasks [XXIV].

The key interface between the physical world and the computing environment is CV technology [XI, XVIII]. CV is defined as a fundamental area that ensures the transition from raw data to the semantic analysis of visual content [IV, XXIII]. Such algorithms make it possible to classify objects and segment scenes to extract interpretable metadata from unstructured pixels. In the literature, such data are actively used to support decision-making [VII, XXII], predict changes in the condition of objects and processes [III, V, XVI], and develop recommendations [II, XV]. This trend aligns with the purpose of this research: based on the automatic analysis of the visual condition of building facades, recommendations can be formulated to mitigate structural degradation.

The conceptual and methodological approach employed in this paper for data acquisition and classification, model selection and training, result interpretation, and test data processing aligns with the methodological tools reported in related studies. The method of systems analysis is actively used to determine the key characteristics of a problem area or a research object. Prior studies indicate that this is achieved because it is possible to study an object not in isolation, but in conjunction with the external environment and internal elements [II, XIII, XVI]. Such an object is represented as a hierarchical system, where at each level a corresponding set of methods is used that are part of the systems analysis framework (structural analysis, synthesis, grouping, mathematical analysis, etc.). As a result, researchers obtain a formal description of the domain, consisting of a set of classes (attributes, states, and methods for each of them), events occurring between them, and the influence factors of external events [III, V]. Thus, this research developed a classification of defects that occur on the facades of civil buildings and structures during their operation.

The choice of neural network architecture in research is mainly based on the nature of the input data and the complexity of the problem being solved, as well as parametric density, topology relevance, optimization strategies, and iterative learning [IX]. Convolutional neural network models are used in spatial feature extraction tasks for the effective recognition of patterns (edges, textures, shapes) regardless of their position [XXV]. The chosen model processes data in real time without significant loss of accuracy, which corresponds to the conclusions of the analyzed studies [XIV]. These studies indicate that detection in the model is a task that solves a single regression problem to predict the coordinates of the bounding boxes and class probabilities in a single forward pass [VIII, XII]. At the same time, during training and prediction, the network “sees” the entire image, which reduces the frequency of false positives on background objects. In addition, the accuracy and speed of the resulting model correspond to the available computing resources.

Defects on the facades of buildings and structures are characterized by irregular shapes, heterogeneity, and a combination of various degradation types. Smartphone cameras are most often used to capture them, and although ideal imaging conditions (illumination, exposure, focusing on objects, etc.) are frequently not met, this does not pose a serious obstacle to processing by the selected neural network model.

Thus, the chosen methodologies can be applied to the object and subject of research in the context of its purpose and objectives. The results obtained correspond to the indicators obtained as a result of using the same methods and tools in related studies. This allows for the conclusion that the developed methodology is reliable.

However, there are certain technical limitations in this research. They are primarily related to the limited number of marked-up images of building facade defects (with specific architecture) typical of the Kyrgyz Republic. This may affect the generalizing ability of the neural network. In addition, climatic conditions (high levels of dust, fog, and precipitation during certain periods of the year), combined with the resolution of the cameras used, can affect the detection of minor defects.

The developed semantic segmentation method has two main limitations:

- Boundary Localization Errors. CNN architectures and spatial upsampling mechanisms tend to smooth contours locally. Consequently, pixel-level (IoU) decreases at irregular, fragmented boundaries (flaking paint, capillary mold, tile spalls), causing the algorithm to merge adjacent small defects or truncate blurry boundaries.
- Macro-Detection Focus. The current scope is strictly limited to high-level defect identification and localization on the facade plane (mAP@0.5 = 91.2%). Perfect pixel-wise approximation of complex fractal micro-geometries is currently redundant for rapid hazard screening.

This edge delineation limitation does not hinder future scaling. For higher technology readiness levels, the baseline macro-detection can be augmented with isolated contour-refinement modules (e.g., PointRend or boundary-aware loss functions) to enable high-precision metric evaluation. Subsequent end-to-end geometric analysis (estimating spall depth and fractal peeling area) will automate the selection of specific repair workflows, such as localized patching, continuous plastering, or full cladding replacement.

Robustness evaluation demonstrated high inference resilience to adverse imaging conditions, including low resolution and motion blur. This stability stems from a multi-level domain normalization framework and aggressive data augmentation simulating JPEG compression, mobile noise, and handheld camera blur. The model mitigates these artifacts, maintaining high prediction confidence (0.78–0.92) on low-quality test images and effectively localizing key damage zones without critical false-positive increases on complex backgrounds.

However, extreme degradation of visual data reveals technical limitations. Severe blur or low sensor resolution combined with adverse weather (fog, precipitation) impairs small-scale defect detection, such as microcracks. Additionally, CNN spatial upsampling tends to smooth complex fractal contours, reducing pixel-level IoU boundaries of fragmented paint peeling or capillary mold. Nevertheless, this limitation does not impede macro-localization for rapid facade monitoring.

V. Conclusion

As a result of the conducted research, the CV methodology was developed and tested for the automatic detection and classification of defects in facades of civil buildings and structures. The developed methodology is based on modern image analysis and machine learning algorithms, which makes it possible to quickly identify various types of damage. Such a software solution can be used as a service of a unified digital housing and public system for planning routine repairs and overhauls of urban infrastructure, enabling communication of local residents with relevant departments, and creating a unified database of buildings and structures.

Scalability analysis of the proposed YOLOv8-seg method demonstrates high technological readiness for integration into large-scale urban monitoring pipelines, including UAV systems. The compact model size (~15 MB) and rapid inference speed (~20 FPS on CPU) enable a flexible two-tier deployment architecture. At the macro level (Edge AI), the lightweight algorithm runs directly on onboard UAV computing modules for real-time video stream filtering, background frame rejection, and rapid degradation zoning. This minimizes bandwidth load by transmitting structured metadata and defect coordinates to the central system instead of raw media streams.

At the micro level, scalability is achieved via batch processing within a centralized Smart City cloud infrastructure. The absence of server-side power constraints allows horizontal scaling of parallel pipelines to process hundreds of gigabytes of urban imagery. In the cloud, the base YOLOv8-seg macro-detector can be cascaded with heavier local contour refinement modules (e.g., PointRend), automating precise metric defect area estimation and mask integration into building digital twins for maintenance planning.

Further development of the methodology is associated with the refinement and expansion of classes of recognizable defects and integration with software solutions for the unification of business processes. Such unification creates prerequisites to develop industry standards for the implementation of innovative solutions in the field of operation and maintenance of buildings in the Kyrgyz Republic. This corresponds to the steady trend toward the digital transformation of key areas of the country's economy and its public administration. The implementation of large-scale projects aimed at introducing modern information technologies, automating business processes, and improving the quality of public services underlines the practical significance of the results obtained and the need to scale the research.

Conflict of Interest

There was no relevant conflict of interest regarding this article.

References

I. Abykeeva-Sultanalieva T, Atakhanov A, Karabalaeva G, Korchueva S, Talgarbekova E (2022). The impact of various factors on the sustainable development of Kyrgyzstan in modern times. Review of Economics and Finance, 20: 1134-1140.

II. Beresneva Ya, Krasnikova I (2025) Technical project of a digital platform for social monitoring of water bodies. E3S WoC, 614: 05001. 10.1051/e3sconf/202561405001

III. Beresneva Ya, Kulibaba I (2024) Simulation model of crop yield on individual land plot. BIO Web Conf., 130: 01019. 10.1051/bioconf/202413001019

IV. Chai J, Zeng H, Li A, Ngai EW (2021) Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications, 6: 100134. 10.1016/j.mlwa.2021.100134

V. Chen Y, Lehmann CU, Malin B (2024) Digital information ecosystems in modern care coordination and patient care pathways and the challenges and opportunities for AI solutions. Journal of medical Internet research, 26: e60258. 10.2196/60258

VI. Gawer A. (2022) Digital platforms and ecosystems: remarks on the dominant organizational forms of the digital age. Innovation, 24(1): 110-124. 10.1080/14479338.2021.1965888

VII. Grankina V, Vasilyev D (2024) Digital transformation modeling for agricultural land irrigation. BIO Web Conf., 141: 02005. 10.1051/bioconf/202414102005

VIII. Hasan MA, Bhargav T, Sandeep V, Reddy VS, Ajay R (2024) Image classification using convolutional neural networks. International Journal of Mechanical Engineering Research and Technology, 16(2): 173-181.

IX. Hassan E, Shams MY, Hikal NA, Elmougy S (2023) The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study. Multimedia Tools and Applications, 82(11): 16591-16633. 10.1007/s11042-022-13820-0

X. Jiang Z, Messner JI (2023) Computer Vision Applications In Construction And Asset Management Phases: A Literature Review. Journal of information technology in construction, 28: 176. 10.36680/j.itcon.2023.009

XI. Khan AA, Laghari AA, Awan SA (2021) Machine learning in computer vision: A review. EAI Endorsed Transactions on Scalable Information Systems, 8(32): 1-11. 10.4108/eai.21-4-2021.169418

XII. Khanam R, Hussain M, Hill R, Allen P (2024) A comprehensive review of convolutional neural networks for defect detection in industrial applications. IEEE Access, 12: 94250-94295. 10.1109/ACCESS.2024.3425166

XIII. Kolodochkin A, Kulibaba I, Ogorodnikov A (2023) Models of traffic and pedestrian flows for organization of smart traffic light traffic. E3S WoC, 403: 07018. 10.1051/e3sconf/202340307018

XIV. Krichen M (2023) Convolutional neural networks: A survey. Computers, 12(8): 151. 10.3390/computers12080151

XV. Li Y, Liu K, Satapathy R, Wang S, Cambria E (2024) Recent developments in recommender systems: A survey. IEEE Computational Intelligence Magazine, 19(2): 78-95. 10.1109/MCI.2024.3363984

XVI. Logachev M (2024) Functional requirements for an intelligent system for monitoring the impact of beaver activity on ecosystems. BIO Web Conf., 130: 08022. 10.1051/bioconf/202413008022

XVII. Logachev M, Krylova L (2023) Operation features of the expert system for tree and shrub pruning. E3S WoC, 462: 02041. 10.1051/e3sconf/202346202041

XVIII. Mahadevkar SV, Khemani B, Patil S, Kotecha K, Vora DR, Abraham A, et al (2022) A review on machine learning styles in computer vision—techniques and future directions. Ieee Access, 10: 107293-107329. 10.1109/ACCESS.2022.3209825

XIX. Niiazalieva KN, Assylbayev AB, Alymkulova AS, Ploskikh VV, Assylbayev KB (2023) Urbanization of the population within the context of urban agglomeration development: a scientific perspective. E3S WoC, 403: 01005. 10.1051/e3sconf/202340301005

XX. Niiazalieva KN, Kydykbaeva AK, Nazarbekova EU, Assylbayev AB, Yessimova SA (2023) Dynamics of development of urban housing stock and urban population. E3S WoC, 403: 01001. 10.1051/e3sconf/202340301001

XXI. Paneru S, Jeelani I (2021) Computer vision applications in construction: Current state, opportunities & challenges. Automation in Construction, 132: 103940. 10.1016/j.autcon.2021.103940

XXII. Simonov V, Karpov A (2024) Intelligent system for monitoring forest plantations condition using unmanned aerial technologies. BIO Web Conf., 145: 04033. 10.1051/bioconf/202414504033

XXIII. Singh G, Pidadi P, Malwad DS (2022). A review on applications of computer vision. In International conference on Hybrid Intelligent Systems (pp. 464-479). Cham: Springer Nature Switzerland. 10.1007/978-3-031-27409-1_42

XXIV. Söylemez M, Tekinerdogan B, Kolukısa TA (2022) Challenges and solution directions of microservice architectures: A systematic literature review. Applied sciences, 12(11): 5507. 10.3390/app12115507

XXV. Zhao X, Wang L, Zhang Y, Han X, Deveci M, Parmar M (2024) A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4): 99. 10.1007/s10462-024-10721-6