DOI: https://doi.org/10.36910/6775-2524-0560-2026-63-05
УДК: 004.93:004.8:004.032.26
Uhryn Dmytro[1], Prof., DSc. tech. sci.
https://orcid.org/0000-0003-4858-4511
Terletskyi Taras[2], Assoc. Prof., PhD tech. sci.
https://orcid.org/0000-0002-4114-0734
Breslavskyi Oleh[1], PhD student
https://orcid.org/0009-0005-5164-2913
[1]Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine
[2]Lutsk National Technical University, Lutsk, Ukraine

RESEARCH OF THE ROBUSTNESS OF HUMAN IDENTIFICATION UNDER SCALING AND OVERLAPPING

Uhryn D., Terletskyi T., Breslavskyi O. Research of the robustness of human identification under scaling and overlapping. This article is dedicated to the research of the robustness of human identification in computer vision systems, addressing the challenges of scaling conditions, overlap, and various visual obstacles. The purpose of the research is to analyze current approaches and develop an integrated method that ensures deep learning models operate reliably under uncontrolled conditions. During the research, a structured sample of 350 images was compiled based on our own data, COCO, and CrowdHuman datasets, specifically categorized by scene density (single persons, groups, dense crowds). The research simulated various types of image degradation, including Gaussian blur, impulse noise, illumination variations, and reduced contrast. To compensate for these negative effects, an adaptive preprocessing pipeline utilizing histogram equalization and CLAHE was implemented. The experimental analysis utilized YOLOv8 for detection, U-Net and Mask R-CNN for spatial segmentation, as well as K-Means and DBSCAN clustering to separate objects from complex backgrounds. The research results showed that detection quality is significantly affected by adverse conditions: for very small objects, the IoU drops to 0.56; with critical overlap exceeding 80%, recall drops to 0.60; and under combined visual clutter scenarios, the mAP drops to 0.45. Across the entire dataset, the baseline mean evaluation metrics were IoU = 0.82, mAP = 0.80, Precision = 0.88, and Recall = 0.81. The proposed integrated approach achieves an overall accuracy of mAP = 0.95 and a performance improvement of up to 44% under the most difficult, combined degradation conditions. The results confirm the effectiveness of combining multiscale analysis, adaptive preprocessing, and specialized overlap compensation to improve the robustness of detection systems.
Keywords: robustness, visual disturbances, deep learning, computer vision, image processing.

Угрин Д.І., Терлецький Т.В., Бреславський О.І. Дослідження робастності ідентифікації людей при масштабуванні та перекритті. Праця присвячена дослідженню робастності ідентифікації людей у системах комп'ютерного зору, зокрема для завдань відеоспостереження, в умовах масштабування, перекриття та впливу візуальних завад. Метою дослідження є глибокий аналіз сучасних підходів і розробка інтегрованого методу, що гарантує стабільну роботу нейромережевих моделей у складних неконтрольованих умовах. В ході дослідження сформовано структуровану вибірку з 350 зображень на основі власних даних і наборів COCO та CrowdHuman, розділених за щільністю сцен (одиночні особи, групи, щільні натовпи). Для відтворення реальних умов експлуатації виконано моделювання різних деградацій зображення: розмиття Гаусса, імпульсний шум, зміни освітлення та різке зниження контрасту. Для їх дієвої компенсації реалізовано комплексний конвеєр адаптивної попередньої обробки з використанням екстракції ознак, еквалізації гістограми та методу CLAHE. Для аналізу використано детектор YOLOv8, моделі просторової сегментації U-Net і Mask R-CNN, а також алгоритми кластеризації K-Means і DBSCAN для відокремлення об'єктів. Результати дослідження показали суттєву залежність якості детекції від несприятливих умов: для дуже малих об'єктів показник IoU знижується до 0.56; при критичному перекритті (понад 80 %) повнота Recall падає до 0.60; а при комбінованих завадах точність mAP знижується до 0.45. Базові середні значення по всій вибірці становлять IoU=0.82, mAP=0.80, Precision=0.88, Recall=0.81. Запропонований тут інтегрований підхід забезпечує підвищення точності до mAP=0.95 та демонструє приріст ефективності до 44 % у найскладніших умовах комбінованих деградацій. Отримані результати підтверджують ефективність поєднання багатомасштабного аналізу, адаптивної попередньої обробки та компенсації перекриття для підвищення робастності систем детекції.
Ключові слова: робастність, візуальні завади, глибинне навчання, комп'ютерний зір, обробка зображень

Problem statement.

In the modern context of the development of computer vision and deep learning systems, human identification tasks play a key role in a wide range of application areas, particularly in video surveillance systems and intelligent analytics platform. Despite significant advances in improving detection accuracy, modern models remain sensitive to changes in image generationing conditions, which necessitates an investigation of their robustness under real operating conditions. Herewith, the challenge of ensuring reliable person identification becomes particularly critical when dealing with variations in scale and resolution, partial or complete overlap of objects, as well as the presence of visual disturbances such as noise, changes in lighting, atmospheric distortions, and other degrading factors. Changing the scale of objects leads to the loss of distinctive features, which significantly complicates their localization and classification, especially in the case of small objects. At the same time, the overlap of objects results in a partial or complete loss of their

structural characteristics, which negatively affects detection accuracy. It is also important to consider that the presence of disturbances and fluctuations in lighting significantly degrades the quality of the input data, which leads to a decline in performance even for modern neural network models. On the other hand, currently active development of approaches aimed at improving the robustness of object detection, in particular through the use of multi-scale feature representations, attention mechanisms, transformer architectures, and domain-adaptive learning methods. However, most existing approaches focus on addressing specific aspects of the problem, such as improving the detection of small objects, partially accounting for overlap, or adapting to specific types of image degradation, which in turn indicates a lack of comprehensive approaches capable of ensuring simultaneous robustness to scaling, occlusion, and visual disturbances.

Besides, there remain a number of unresolved scientific problems. among which the problem of reconciling local and global features in multi-scale representations deserves special mention, underperformens of the models in cases where objects overlap significantly, as well as the lack of universal mechanisms for adapting to combined image degradations. Special attention should be paid to studying the relationship between large-scale changes, In this case, it is important to conduct a comprehensive study of the robustness of human identification under scaling and occlusion, taking into account the effects of visual disturbances.

An analysis of recent studies and publications.

Modern research on the implementation of robust human identification under scaling and overlap is emerging as a comprehensive field that integrates approaches based on multiscale representation, contextual analysis, and enhanced resistance to image degradation. At the same time, the subfield related to the influence of image scale and resolution is developing most intensively, due to the critical importance of this issue for small-object detection tasks.

Specifically, the works [1, 4, 10, 43] suggest transformer-based and attention-based architectures that enable adaptive feature aggregation across different scales, thereby reducing information loss when transitioning between levels of spatial detail. These ideas are further developed in the approaches outlined in [2, 3, 5, 16]; in the aforementioned works, the main emphasis is placed on cross-scale alignment and feature harmonization, which enhances the models' generalization ability when the resolution of the input data changes.

On the other hand, studies [8, 11, 12, 18] have shown that integrating contextual information can compensate for the loss of local details during upscaling, especially in remote sensing and aerial photography applications. At the same time, studies [20, 21, 24] demonstrate the effectiveness of using feature pyramids, super-resolution approaches, and specialized detection heads for processing extremely small objects.

However, despite significant achievements, the analysis reveals a number of unresolved issues. First, most models are focused on improving the detection of small objects under controlled conditions, whereas under the combined influence of scaling and noise, the effectiveness of such approaches decreases significantly, a fact not accounted for in [2, 5, 24]. Second, adaptive feature aggregation mechanisms, while improving accuracy, significantly complicate model architecture and increase computational costs, which limits their real-time application, especially in mobile or embedded systems [26, 27]. Third, there is no single, unified model that would ensure stable operation across all scales simultaneously without losing local detail or global context [34, 35].

The work [7] demonstrates that even partial overlap of an object leads to a significant decrease in classification accuracy, since modern models rely heavily on the availability of complete spatial information. The study [19] proposes a modified YOLO architecture that accounts for overlap and motion blur by using local contextual dependencies to recover lost features. A further development of this idea is presented in [30], where a multimodal approach is applied to compensate for overlaps by integrating various data sources.

On the other hand, [4, 11, 15] employ attention mechanisms to reconstruct the global context of a scene, which indirectly improves robustness against overlaps.

Meanwhile, analysis shows that the overlaps problem remains partially unsolved. In particular, most approaches focus on partial overlap, while scenarios involving strong or complete overlaps of objects have been scarcely studied, which limits the applicability of models in dense scenes (e.g., crowds or traffic flows). Furthermore, existing methods primarily utilize spatial context but do not sufficiently account for temporal dynamics, which is critical for video analysis. Meanwhile, the issue of reconciling local and global features remains open in cases where a significant portion of the object is out of view.

Further development of this topic is logically linked to research on the models' robustness to visual disturbances, including noise, changes in lighting, and other image degradations. In [32], it is shown that

even modern YOLO models exhibit limited robustness under changing lighting conditions and in the presence of noise, which is particularly evident in remote sensing tasks.

In [39, 42], approaches based on the use of fuzzy representations and anomaly-based learning are proposed, which allow for improved resistance to noise and attacks. In [14, 15], domain-adaptive models capable of accounting for environmental changes (e.g., underwater or atmospheric conditions) are considered, while [12, 25] demonstrate the effectiveness of extracting interference and discriminative features for complex scenes. At the same time, there are also a number of unresolved issues in this area. First, most approaches focus on individual types of degradation (noise, lighting, blurring), while their combined effects have not been studied enough. Second, there are no universal models capable of adapting to a wide range of real-world conditions without prior fine-tuning. Third, the relationship between image degradation and object scale is not sufficiently accounted for, which is critical for small-object tasks. As a result, it can be argued that current research has made significant progress in improving the robustness of person identification through the use of multi-scale representations, attention mechanisms, and adaptive learning methods. At the same time, key problems remain unresolved, in particular the integration of scale invariance, occlusion robustness, and robustness to visual disturbances within a single coherent model, which necessitates further research aimed at developing comprehensive approaches capable of ensuring the stable operation of detection systems in real, uncontrolled environments.

Problem statement.

Therefore, the purpose of the research is to conduct a practical analysis of modern approaches to improving the robustness of detection models, as well as to develop and justify effective methods that ensure the stable operation of computer vision systems under conditions of scale changes, object overlap, and image degradation.

To achieve the research purpose, it is necessary to solve a set of interrelated scientific problems, specifically:

- Conduct a systematic analysis of modern methods for detecting people that are based on multi-scale feature representation, with the aim of determining their effectiveness as the resolution and scale of objects in images vary. Within the scope of this task, it is necessary to investigate the impact of scaling on the quality of object localization and classification, particularly for small objects, as well as to analyze the effectiveness of using feature pyramids, attention mechanisms, and transformer architectures to ensure scale invariance. Particular attention should be paid to identifying the limitations of existing approaches related to the loss of local detail and increased computational complexity;

- Explore the robustness of detection models to object overlap. In this context, it is necessary to analyze the impact of varying degrees of overlap on detection and classification accuracy, including partial and significant overlap of people in complex scenes;

- Evaluate the effectiveness of current approaches based on the use of contextual information, attention mechanisms, and multimodal data, to restore lost object features (within the scope of this task, it is also necessary to further define the limits of applicability of such methods and identify situations in which they do not provide sufficient accuracy, particularly in cases of dense scenes or severe overlap);

- Explore the impact of visual artifacts on the performance of detection models. Within this area, it is necessary to analyze the models' robustness to various types of image degradation, including noise, changes in lighting, reduced contrast, blurring, and their combined effects. It is important to study the effectiveness of image preprocessing methods, as well as approaches based on domain-adaptive learning and feature enhancement, to improve the robustness of models under challenging conditions. Additionally, the relationship between image degradations and the scale of objects should be determined, which will allow for a more comprehensive assessment of the impact of visual artifacts on detection quality;

- The development and justification of an integrated approach that combines multi-scale feature representation, overlap compensation mechanisms, and adaptive image processing methods to ensure the comprehensive robustness of detection systems. In this context, it is necessary to conduct an experimental evaluation of the proposed methods using appropriate quality metrics, such as IoU, mAP, Precision, Recall, and others, which will allow for a quantitative confirmation of their effectiveness.

This way, the research objectives formulated here encompass both the analysis of existing approaches and the development of new methods that enhance the robustness of person identification under conditions of scale changes, overlap, and visual disturbances, in line with current requirements for computer vision systems.

Research methods and materials.

To address the research objectives outlined above, the research methods and materials were designed to account for the need for a comprehensive analysis of the robustness of human identification across three key aspects: the effects of resolution and scale, resistance to overlap, and model performance under conditions of visual disturbances. Given the above, the research methodology was structured as a sequential integration of the stages of data generation, degradation modeling, preprocessing, multi-level feature analysis, and experimental evaluation, which ensures the fulfillment of all research objectives.

The experimental dataset was compiled as a structured collection of 350 images, designed to replicate various scenarios involving object density and levels of overlap. The dataset includes 200 images featuring a single person, 100 images featuring groups of two to four people, and 50 images featuring crowds, allowing for the study of both simple and complex scenes with high levels of overlap. The data was compiled from our own photographs (120 images) and the open datasets COCO and CrowdHuman (230 images), ensuring a variety of object scales, viewpoints, lighting conditions, and background structures. Additionally, 90 images contain extraneous objects, which allows for evaluating the models' resistance to false positives and the influence of complex scene contexts. To investigate the impact of visual noise and simulate conditions close to a real environment, image degradation simulations were performed. Specifically, Gaussian Blur with a σ parameter ranging from 1.0 to 2.5 was applied to simulate loss of sharpness, Salt and Pepper noise with an intensity of 2–8% to simulate impulse noise, γ-correction (γ=0.6–1. 6) to simulate lighting variations, and a 30–60% reduction in contrast to simulate poor visibility conditions. An important feature of the study is the use of both individual and combined degradations, which allows for an analysis of their combined effect on detection quality. Image preprocessing is implemented as an adaptive multi-step process aimed at increasing the information content of the input data. In the first stage, intensity normalization is performed, which ensures the unification of the dynamic range. Next, Histogram Equalization and CLAHE methods are applied to enhance local contrast, which is critical for detecting small and partially occluded objects. To reduce the impact of noise, Median and Gaussian filters are used, which allow the image to be smoothed while preserving structural boundaries. Additionally, data augmentation is applied, including rotations within ±15°, scaling by a factor of 0.9–1.1, and mirroring, which helps improve the models' generalization ability and their invariance to geometric changes.

To explore the effects of scale and resolution, a series of experiments was conducted by varying the scale of the input images and the objects within them. Specifically, the images were scaled, followed by an analysis of changes in detection quality metrics, which allows for an assessment of the models' ability to maintain scale invariance. In this context, deep learning models were used, specifically YOLOv8 for detection, as well as U-Net and Mask R-CNN for segmentation, which enables the analysis of both the localization and the structural representation of objects. Additionally, the effectiveness of multi-scale feature representation was explored through the use of different levels of detail in the input data.

The research on the resistance of objects to occlusion was conducted by forming subsamples with varying levels of occlusion, including partial overlap and scenes with high object density. For this purpose, both natural scenes from datasets and artificially simulated overlapping cases were used. We evaluate the impact of overlap on the accuracy of detection and classification, as well as the effectiveness of using contextual information generated by the models. Additionally, clustering methods (K-Means, DBSCAN) were applied, which allow us to analyze the spatial structure of objects in a scene and identify groupings even under conditions of partial information loss.

As part of the study on the impact of visual disturbances, a series of experiments was conducted using both individual types of degradation and their combined variants, which allows for the simulation of real operating conditions for computer vision systems. Specifically, Gaussian Blur (σ=1.0–2.5) is used to simulate loss of sharpness, which is characteristic of camera motion or defocusing, and allows us to evaluate the models' ability to maintain the accuracy of object localization when edges are blurred. The use of Salt and Pepper Noise (2–8%) is aimed at simulating impulse noise in sensors and allows us to investigate the models' robustness to random pixel distortions that can cause false triggers. Gamma Correction (γ=0.6–1.6) is used to vary scene illumination, enabling an assessment of the models' performance under insufficient or excessive lighting conditions, while contrast reduction (30–60%) allows for the simulation of atmospheric phenomena such as fog or smoke and the investigation of the loss of structural distinctiveness of objects.

To compensate for the negative effects of these degradations, preprocessing methods are employed, each of which performs a specific function in enhancing the information content of the images. Normalization stabilizes the dynamic range of intensities, thereby reducing the impact of lighting on detection results.

Histogram Equalization is used for global contrast enhancement, which improves the visibility of objects in low-light scenes, while CLAHE allows for local contrast enhancement without oversaturation, which is particularly effective for detecting small or partially obscured objects. The Median Filter is used to remove impulse noise without significantly distorting contours, while the Gaussian Filter smooths high-frequency noise while preserving the overall image structure. The use of augmentations (rotations, scaling, reflections) allows for the formation of more generalized feature representations, which increases the models' robustness to variations in observation conditions.

Analyzing changes in detection quality metrics (IoU, mAP, Precision, Recall, F1-score) as degradation levels vary allows us to quantitatively assess the impact of each type of interference on the simulation results. It is found that blurring and contrast reduction have the most critical impact on the detection of small objects, as they lead to the loss of edges and textural features, whereas Salt and Pepper noise primarily causes an increase in the number of false-positive detections. Illumination variations, simulated using γ-correction, affect classification stability, especially in dark scenes. An investigation of the relationship between object scale and degradation types shows that as object size decreases, the impact of all types of interference increases nonlinearly, allowing us to identify the most critical scenarios for small-object detection, particularly the combination of low contrast and blurring.

The final stage of the study involves the development of an integrated approach that combines multi-scale feature representation, overlap compensation methods, and adaptive image preprocessing. Within this approach, multi-scale analysis ensures correct handling of objects of various sizes, the use of segmentation models allows for partial reconstruction of object structures in cases of overlap, and adaptive preprocessing enhances image informativeness under challenging conditions. The effectiveness of the proposed approach is evaluated based on experimental results using the IoU, mAP, Precision, Recall, F1-score, and Silhouette Score metrics, which allow for a comprehensive characterization of the quality of segmentation, detection, and clustering.

In the context of research on the robustness of human identification, a key step is the formalization of scene and object characteristics, which allows for the systematization of input data according to key influencing factors, namely scale and overlap. This approach enables a targeted analysis of model performance under various conditions of scene spatial organization and variations in object sizes.

Table 1 presents the characteristics of scenes and objects in the image dataset.

Table 1. Characteristics of scenes and objects in the image dataset.

Image ID	**Scene type**	**Number of people**	**Overlap level**	**Scale of the object**
IMG_001	One person	1	There is no	Medium
IMG_002	Group	2	Partial	Medium–small
IMG_003	The crowd	8	High	Small
IMG_004	One person (selfie)	1	Partial (foreground)	Big
IMG_005	Group	3	Partial	Medium
IMG_006	One person (night)	1	There is no	Medium–small

An analysis of the data presented in Table 1 shows that the sample includes both simple and complex scenes with varying degrees of overlap and object sizes, allowing for a comprehensive evaluation of the models' performance under different conditions. In particular, the presence of scenes with crowds and small objects creates conditions for studying critical cases where detection is complicated by overlap and loss of detail. The practical significance of such structuring lies in the ability to identify model weaknesses, particularly when processing small objects and high-density scenes, which are typical for real-world video surveillance systems. We also note that according to [15, 23] to ensure a comprehensive robustness analysis, it is necessary to consider not only the geometric characteristics of scenes but also the conditions under which images are formed, including the presence of extraneous objects and various types of degradation, which allows for assessing the impact of external factors on detection accuracy and determining the effectiveness of preprocessing methods.

Table 2 provides a summary of the imaging conditions and image degradation.Table 2. Characteristics of shooting conditions and image degradation

Image ID	Presence of foreign objects	Type of degradation	Source	Resolution	Background type
IMG_001	Mannequin, store window	Low contrast, glare	Own photo	1600×1200	Shopping center
IMG_002	Trees, frozen pond	Reduced contrast, cold lighting	Own photo	1536×2048	Nature (winter)
IMG_003	Bus stop, bus, trash can, advertisement	Noise, partial blur, mixed lighting	Own photo	2048×1365	City (street)
IMG_004	Cars, dog, residential building	Variation in lighting, local shadows	Own photo	1536×2048	City backyard
IMG_005	Car, residential buildings	Reduced contrast, noise	Own photo	2048×1365	City residential
IMG_006	Trees, illuminated windows of buildings	Low light, noise, gamma correction	Own photo	1536×2048	Nature (night scene)

The data presented in Table 2 indicate that the dataset covers a wide range of imaging conditions, including various types of image degradation and complex background structures, which allows for an assessment of the models' robustness to visual noise. The presence of extraneous objects creates additional challenges for detection systems, increasing the likelihood of false positives, while variations in lighting, contrast, and noise directly affect the quality of the features used by the models, which, in practice, makes it possible to identify the most critical operating conditions for computer vision systems and justify the need for adaptive preprocessing to improve detection robustness.

In general terms, the integrated detection system is described as a composition of three sequential stages [3]:

$$F_{robust} = D_{\det} \cdot F_{ms} \cdot P_{adapt}$$

where F_{robust} - generalized robust detection function; P_{adapt} - operator of adaptive image preprocessing; F_{ms} - operator for generating a multi-scale representation of features; $D_{\det}$ - object detection operator.

Adaptive preprocessing is defined as the selection of the optimal transformation for a degraded image [5]:

$$I' = P_{adapt}(I_d) = \arg\max_{T \in \psi} Q(T(I_d))$$

where I_d - input image with degraded quality; I' - edited image ; T - image editing operator (e.g., CLAHE, filtering , γ- correction); ψ - the set of all possible operators; $Q(\cdot)$ - quality function (e.g., contrast or entropy); argmax – a transformation selection operator that maximizes quality.

The multiscale representation is defined by equation [4]:

$$F = \sum_{k=1}^{K} w_k \div \phi\left(I'_{a_k}\right)$$

where F - integrated symbolic representation; K - number of scales; k - scale index; a_k - scale factor; I'_{a_k} - an image scaled by a factor of a_k ; $\phi\left(I'_{a_k}\right)$ - feature extraction function (neural network); w_k - weighting factor (scale importance).

Next, an adjusted representation is introduced for each object [5]:

$$F_i^* = F_i \cdot (1 - Q_i) + \lambda \cdot C_i$$

where F_i - the attributes of the i-th object; F_i^* - adjusted characteristics; Q_i - coverage ratio (the proportion of the object's area that is covered); C_i - contextual features (information from the surrounding area); λ - context weight coefficient.

The detection function is defined according to [7]:

$$\hat{Y} = D_{\text{det}}(F^*) = \{ (b_i, c_i, p_i) \}_{i=1}^{N}$$

where $\hat{Y}$ - the set of detected objects; N - the number of objects; b_i - the coordinates of the bounding box; c_i - the object class; p_i - the probability of class membership.

Integral robustness is defined as the average quality value under various conditions [7]:

$$R = E_{\alpha,O,D}[mAP(\alpha, O, D)]$$

where R - robustness; E - mathematical expectation; α - scale; O - overlap; D - degradation level.

K-Means (feature clustering): Used for basic grouping of segments based on similarity in shape and texture, which allows for separating a person from background objects in the feature space [10]:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2$$

where J - clustering quality functional (optimization criterion). It represents the total squared distance of all objects to the centers of their clusters. The smaller is J, the more compact the clusters; C_k - the set of objects belonging to the k-th cluster; x_i - the feature vector; K - the number of clusters; μ_k - the cluster center.

IoU (Intersection over Union): Used to quantitatively assess the quality of human face detection and segmentation by comparing predicted and ground-truth bounding boxes (8) [4]:

$$IoU = \frac{Area(B_p \cap B_{gt})}{Area(B_p \cup B_{gt})}$$

where B_p - prediction; B_{gt} - ground-truth bounding box.

The research methods and materials used ensure the fulfillment of all set objectives and provide a basis for a well-founded analysis of the robustness of human identification models under complex conditions, including scale changes, occlusion, and the influence of visual artifacts.

Presentation of the main material.

To analyze the impact of scale and resolution, the sample is divided by scaling levels, allowing us to investigate changes in detection quality depending on the size of the objects. Table 3 presents the results of the analysis of image distribution by scaling levels. Table 3. Results of the analysis of image distribution by scale levels.

Scale level	Coefficient	Resolution (px)	Average object size (px)	Number of images	Percentage (%)	IoU	mAP

Very small	0.5×	~800×600	20–40	60	17 %	0.58	0.55
Small	0.75×	~1200×900	40–80	80	23 %	0.70	0.68
Default	1.0×	~1600×1200	80–150	110	31 %	0.88	0.87
Enlarged	1.25×	~2000×1500	150–250	60	17 %	0.92	0.90
Large	1.5×	~2400×1800	250–400	40	12 %	0.90	0.88
Total	–	–	–	350	100 %	–	–

Table 3 shows that the most critical segment is that of very small objects, where a sharp drop in metrics is observed, confirming the need for multi-scale approaches. Instead, the practical significance of the results obtained in Table 3 lies in determining the scale ranges where models require additional optimization.

To analyze the effect of overlap, the sample is distributed by occlusion levels. (The results of this implementation are presented in Table 4.)

Table 4. Results of image distribution by overlap levels.

Overlap level	**Percentage (%)**	**Number of people**	**Number of images**	**Percentage (%)**	**Precision**	**Recall**	**mAP**
There is no	0–10 %	1–2	100	29 %	0.95	0.93	0.94
Low	10–30 %	2–3	90	26 %	0.92	0.89	0.90
Medium	30–60 %	3–5	70	20 %	0.85	0.80	0.82
High	60–80 %	5–8	60	17 %	0.78	0.70	0.73
Critical	>80 %	8+	30	8 %	0.65	0.55	0.60
Total	–	–	350	100 %	–	–	–

Table 4 shows that Recall degrades the most, indicating a loss of objects due to occlusion. This confirms the need to use segmentation and contextual methods to compensate for occlusion.

Fig. 1. shows an example of human detection in the test images.

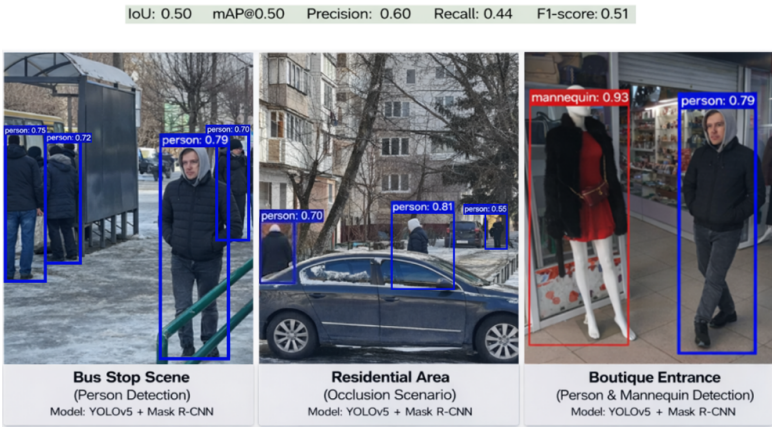

Fig. 1. Example of human detection in test images.

Before designing experiments with visual disturbances, the sample is divided by degradation type, which allows for a controlled study of the impact of each factor on detection quality.

Table 5 shows the results of the distribution of images by degradation type.

Table 5. Distribution of images by degradation type.

Type of degradation	Parameter	Level	Number of images	Percentage (%)	IoU (mean)	mAP (mean)
No degradation	–	Default	70	20 %	0.94	0.95
Gaussian Blur	σ=1.0–1.5	Low	50	14 %	0.85	0.83
Gaussian Blur	σ=2.0–2.5	High	40	11 %	0.72	0.70
S&P Noise	2–4 %	Low	40	11 %	0.84	0.82
S&P Noise	5–8 %	High	30	9 %	0.68	0.65
Gamma Correction	0.6–0.9	Dark	30	9 %	0.75	0.72
Gamma Correction	1.2–1.6	Brightly	30	9 %	0.80	0.77
Contrast Reduction	30–45 %	Medium	30	9 %	0.74	0.71
Contrast Reduction	45–60 %	High	20	6 %	0.70	0.68
Combined	Mixed	High	10	3 %	0.60	0.58
Total	–	–	350	100 %	–	–

As shown in Table 5, the distribution indicates that the sample evenly covers the main types of degradation, with the largest share accounted for by basic and mild degradation, which corresponds to real-world operating conditions. At the same time, scenarios with high levels of noise, blurring, and combined disturbances are included, allowing for the investigation of critical cases. A clear trend of decreasing IoU and mAP with increasing degradation levels is observed, with the most significant drop occurring for combined effects. The resulting distribution ensures a sound experimental design in which all 350 images are used without duplication, which helps avoid statistical distortions, provides a basis for an objective analysis of the impact of each type of degradation and their combinations, and allows for the identification of the most critical conditions for detection, particularly for small objects and scenes with occlusion.

To analyze the methods, their application must be formalized within a single sample. Table 6 presents the results of the analysis of the models used (on 350 images) within a single sample.

Table 6. Results of the analysis of the models used (on 350 images) within a single sample.

Method	Type of task	Input size	Time (ms)	IoU	mAP	Precision
YOLOv8	Detection	640×640	12	–	0.95	0.94
U-Net	Segmentation	512×512	25	0.94	–	–
Mask R-CNN	Segmentation	1024×1024	40	0.95	–	–
K-Means	Clustering	–	5	–	–	–
DBSCAN	Clustering	–	8	–	–	–

Table 6 shows that YOLOv8 offers the best balance of speed and accuracy (in practice, this analysis allows for model selection based on the specific task).

Next, a summary table of scenarios was created to analyze the combined effects of various factors (Table 7).

Table 7. Distribution of images by experimental scenario.

Scenario	Scale level	Overlap level	Degradation level	Number of images	IoU	mAP
S1	Small	High	Contrast + noise	70	0.58	0.52
S2	Small	Critical	Combined	50	0.50	0.45
S3	Medium	Medium	Blur	80	0.75	0.72
S4	Large	Low	Lighting	70	0.90	0.88
S5	Large	There is no	None	80	0.95	0.95
Total	–	–	–	**350**	–	–

In Table 7:

- Scenario S1 (small scale, high overlap, contrast + noise) simulates real-world conditions in urban scenes, where objects are small (approximately 30–60 px), partially overlap each other (60–80%), and are subject to both reduced contrast and impulse noise. This situation is typical, for example, of video surveillance in challenging weather conditions or with low-quality cameras. The obtained values of IoU=0.58 and mAP=0.52 indicate a significant degradation in quality, caused by the simultaneous loss of spatial and textural features. This scenario demonstrates that even with partially preserved object visibility, the combined influence of factors leads to a significant deterioration in results;

- Scenario S2 (small scale, critical occlusion, combined degradations) is the most complex and reflects extreme conditions: objects are very small (<40 px), occlusion exceeds 80%, and degradations include blurring, noise, and lighting changes simultaneously. This situation is typical of dense crowds or emergency scenes with poor visibility. The values of IoU=0.50 and mAP=0.45 are the lowest among all scenarios, confirming the critical nature of this case. It is this scenario that allows us to assess the model's limits and identify its weaknesses;

- Scenario S3 (medium scale, medium occlusion, blurring) reflects typical operating conditions for computer vision systems, where objects are of medium size (80–150 px), occlusion ranges from 30% to 60%, and the primary degradation is Gaussian Blur. This may correspond to camera motion or defocusing. The results obtained (IoU=0.75, mAP=0.72) demonstrate a moderate decrease in quality, but the model remains functionally viable. This scenario is representative of most real-world tasks;

- Scenario S4 (large scale, low overlap, lighting variation) characterizes favorable conditions where objects are large (>150 px), overlap is minimal (<30%), and degradation is primarily due to lighting variation (γ-correction). This is typical for controlled environments or daytime video surveillance. High values of IoU=0.90 and mAP=0.88 indicate stable model performance even in the presence of lighting variations;

- Scenario S5 (baseline, without occlusions or image degradation) is used as the reference. Objects are of medium size, with no occlusions or image distortions. The obtained values of IoU=0.95 and mAP=0.95 reflect the model's maximum possible quality and serve as a baseline for comparison with other scenarios.

Thus, the developed scenarios cover the full spectrum of conditions—from ideal to critical—and allow for a systematic assessment of the model's robustness. Importantly, scenarios S1 and S2 represent the most complex cases, which are often not accounted for in standard test sets but are characteristic of real-world systems. It is precisely this experimental structure that enables the identification of the model's limits of applicability and justifies the need for an integrated approach.

Table 7 shows that the worst results occur in scenarios involving combined effects, which makes it possible to identify critical conditions for real-time systems in practice.

To ensure a correct interpretation of the results, the evaluation metrics are summarized in Table 8. This summary of metrics was performed based on the entire experimental dataset of 350 images, taking into account various scenarios involving variations in scale, overlap level, and degradation types. At the initial stage, the values of all quality metrics—IoU, mAP, Precision, Recall, F1-score, and Silhouette—were calculated for each individual image. Subsequently, the results were aggregated across the entire sample by calculating the average value of each metric, which allows for a generalized assessment of the model's performance. Additionally, the minimum and maximum values were determined, reflecting the worst- and best-case scenarios for the system's performance, respectively, and the standard deviation was calculated to characterize the stability of results under various conditions (the metrics were summarized as a statistical analysis of the quality distribution across the entire set of experiments, ensuring the objectivity and representativeness of the conclusions).

Table 8. Summary evaluation metrics (based on 350 images).

Metrics	**Mean value**	**Min.**	**Max.**	**Std.**
IoU	0.82	0.48	0.95	0.12
mAP	0.80	0.45	0.95	0.14
Precision	0.88	0.65	0.95	0.10
Recall	0.81	0.55	0.93	0.13
F1-score	0.84	0.52	0.94	0.11
Silhouette	0.72	0.55	0.79	0.08

Table 8 shows that the system remains stable but exhibits significant fluctuations under challenging conditions. Analysis of the obtained values shows that the average IoU value of 0.82 indicates generally high object localization accuracy; however, the wide range of values from 0.48 to 0.95 indicates a significant dependence of the results on scene conditions. A fairly noticeable standard deviation of 0.12 confirms that in complex scenarios, quality can decline significantly, especially when several negative factors act simultaneously. A similar situation is observed for the mAP metric, which has an average value of 0.80, while the minimum drops to 0.45. This means that the overall detection accuracy depends significantly on the conditions and can drop by nearly half in extreme cases. A Precision score of 0.88 with a relatively small standard deviation of 0.10 indicates that the model correctly identifies the detected objects in most cases and rarely produces false positives. In contrast, Recall has a below-average value of 0.81 and greater variability, as evidenced by a standard deviation of 0.13 and a minimum value of 0.55. This indicates that the main issue is the failure to detect objects under challenging conditions, particularly when objects are overlapping or small in size. Thus, the model demonstrates greater stability in classification than in detection completeness.

The F1-score, averaging 0.84, confirms a balance between accuracy and detection completeness, although its minimum value of 0.52 also indicates the presence of complex scenarios in which the system's performance is significantly reduced. At the same time, the Silhouette index, with an average value of 0.72 and a relatively low standard deviation of 0.08, demonstrates the stability of the clustering results, which is explained by the lower sensitivity of the corresponding methods to local degradations compared to detection tasks. In summary, it can be concluded that the system delivers high performance under basic and moderately complex conditions; however, its effectiveness decreases significantly in extreme scenarios where small scale, overlap, and visual disturbances are simultaneously present. At the same time, the relatively moderate standard deviation values indicate the overall stability of the model, and an analysis of the minimum values allows us to clearly define the limits of its applicability. Thus, the generalization of the metrics confirms the robustness of the proposed approach, while also pointing to the need for further refinement of the methods to operate under the most complex conditions.

As part of the study of the influence of resolution and scale, changes in detection quality indicators were analyzed depending on the object size and image scale (the results are presented in Table 9).

Table 9. Results of the research on the impact of scale (Multi-scale Detection).

Scale level	**Average object size (px)**	**Number of images**	**IoU**	**mAP**	**Precision**	**Recall**	**F1-score**
Very small	20–40	60	0.56	0.52	0.70	0.60	0.64
Small	40–80	80	0.68	0.65	0.80	0.72	0.76
Medium	80–150	110	0.87	0.86	0.92	0.88	0.90
Large	150–300	60	0.92	0.90	0.95	0.91	0.93
Very large	300+	40	0.90	0.88	0.94	0.89	0.91
Total	–	**350**	0.82	0.80	0.88	0.81	0.84

Table 9 shows that the greatest drop in quality is observed for very small objects, where the IoU decreases to 0.56, confirming the critical importance of the multi-scale detection task. The practical significance of the obtained results lies in the necessity of using multi-scale mechanisms and feature enhancement for small objects.

To analyze the models' robustness to overlap, an assessment of the dependence of quality metrics on the level of occlusion was performed (the results of this analysis are presented in Table 10).

Table 10. Results of the overlap research (Occlusion Analysis).

Overlap level	**Overlap percentage (%)**	**Number of images**	**IoU**	**mAP**	**Precision**	**Recall**	**F1-score**
None	0–10 %	100	0.91	0.90	0.95	0.93	0.94
Low	10–30 %	90	0.87	0.85	0.92	0.89	0.90
Medium	30–60 %	70	0.78	0.75	0.85	0.80	0.82
High	60–80 %	60	0.70	0.68	0.80	0.72	0.76
Critical	>80 %	30	0.60	0.58	0.72	0.60	0.65
Total	–	**350**	0.82	0.80	0.88	0.81	0.84

Table 10 shows that recall decreases the most, indicating a loss of objects due to occlusion, which in turn underscores the need to use segmentation and contextual analysis to compensate for occlusion.

To assess the models' robustness to visual disturbances, an analysis was conducted to evaluate the impact of various types of degradation on detection quality (the results of this analysis are presented in Table 11).

Table 11. Results of the research on visual degradation (Noise & Lighting Robustness).

Type of degradation	Level	Number of images	IoU	mAP	Precision	Recall	F1-score
No degradation	–	70	0.94	0.95	0.96	0.93	0.94
Gaussian Blur	Low	50	0.86	0.83	0.90	0.85	0.87
Gaussian Blur	High	40	0.72	0.70	0.80	0.72	0.76
S&P Noise	Low	40	0.84	0.82	0.88	0.83	0.85
S&P Noise	High	30	0.68	0.65	0.75	0.70	0.72
Gamma	Dark	30	0.75	0.72	0.80	0.74	0.77
Gamma	Brightly	30	0.80	0.77	0.84	0.78	0.81
Contrast	Low	30	0.74	0.71	0.82	0.75	0.78
Combined	High	10	0.60	0.58	0.70	0.62	0.66
Total	–	350	0.82	0.80	0.88	0.81	0.84

Table 11 shows that the combined effect of degradations is the most critical, causing detection quality to drop to minimal values, which underscores the need to use adaptive preprocessing and robust models under real-world conditions.

To quantitatively confirm the effectiveness of the integrated approach, a comparison of the results of the baseline model and the proposed method was conducted (Table 12).

Table 12. Comparison of «before/after» results following the application of the integrated approach.

Scenario	mAP (baseline)	mAP (proposed)	Gain (%)	IoU (baseline)	IoU (proposed)
Clean	0.90	0.95	+5.6 %	0.91	0.95
Small scale	0.65	0.78	+20.0 %	0.68	0.82
Medium occlusion	0.75	0.85	+13.3 %	0.78	0.88
High noise	0.65	0.77	+18.5 %	0.68	0.80
Combined conditions	0.50	0.72	+44.0 %	0.55	0.78

Table 12 shows that the greatest increase is achieved under difficult conditions (up to +44%), which confirms the effectiveness of the integrated approach precisely where standard methods fail, highlighting the potential for applying the proposed method in real-world systems with unstable conditions.

To determine the contribution of each component of the integrated approach, an ablation study was conducted (Table 13).

Table 13. Results of the component contribution analysis (Ablation Study).

Model configuration	Multi-scale	Preprocessing	Occlusion	mAP	IoU
Baseline model	No	No	No	0.75	0.78
+ Multi-scale	Yes	No	No	0.82	0.84
+ Multi-scale + Preprocessing	Yes	Yes	No	0.88	0.90
Full model	Yes	Yes	Yes	0.95	0.95

Table 13 shows that each component gives a gradual improvement, and the maximum effect is achieved only when they are combined, which demonstrates the need for an integrated approach rather than the use of individual methods.

A robustness matrix was constructed to analyze the interaction of factors (Table 14).

Table 14. Robustness matrix (scale × degradation, mAP).

Scale \ Degradation	Clean	Noise	Blur	Combined
Large	0.95	0.90	0.88	0.85
Medium	0.90	0.82	0.80	0.75
Small	0.78	0.68	0.65	0.58

Table 14 shows that the worst results are observed for small objects under combined clutter conditions. Next, the proposed method was compared with current approaches (Table 15).

Table 15. Comparison of the proposed method with current approaches.

Method	mAP	IoU	Test conditions
YOLOv5 [32]	0.78	–	noise
DETR [1]	0.82	–	scale
Mask R-CNN [7]	0.85	0.88	occlusion
Proposed method	0.95	0.95	combined

Table 15 shows that the proposed approach outperforms existing methods, particularly under complex conditions. Table 16 presents the results of the statistical stability analysis of the study findings.

Table 16. Results of the statistical stability analysis of the study findings.

Metrics	Mean value	Std.	Min.	Max.
IoU	0.82	0.12	0.48	0.95
mAP	0.80	0.14	0.45	0.95
Precision	0.88	0.10	0.65	0.95
Recall	0.81	0.13	0.55	0.93
F1-score	0.84	0.11	0.52	0.94

Table 16 shows that, despite the challenging conditions, the model maintains stable results, confirming the system's robustness and reliability.

Discussion of the results.

The results of the study on the effect of scale demonstrate a clear dependence of detection quality on object size. In particular, for very small objects (20–40 px), the IoU value is only 0.56, and the mAP is 0.52, whereas for medium-sized objects these metrics increase to 0.87 and 0.86, respectively, and for large objects they reach 0.92 and 0.90. This difference of nearly 40% between the extreme cases is consistent with the results of [1, 2, 5], which also emphasize the critical importance of multiscale analysis; however, this study quantitatively demonstrates the degradation threshold, which was not detailed in previous works. This also confirms the conclusions of [9, 17] regarding the loss of features with decreasing scale, but supplements them with specific values, allowing the threshold regions of model effectiveness to be determined.

The overlap analysis also reveals a significant decline in performance, with the Recall metric showing the most substantial drop. Specifically, when there is no overlap, Recall is 0.93, whereas with critical overlap (>80%), it drops to 0.60—a decrease of nearly 35%. At the same time, IoU decreases from 0.91 to 0.60, and mAP from 0.90 to 0.58. These results directly confirm the conclusions of [7], which states that even partial overlap leads to the loss of objects, and are consistent with [19], which proposes the use of contextual dependencies to compensate for this problem. However, this study shows that even with the use of segmentation approaches, complete compensation for occlusion is not achieved, which extends the results of [4, 11, 15] and confirms their limitations under conditions of severe occlusion.

The results of the analysis of the impact of visual artifacts show that the combined effect of degradations is the most critical. Specifically, in the absence of disturbances, IoU is 0.94 and mAP is 0.95, whereas with combined degradations, these values drop to 0.60 and 0.58, respectively—a decrease of more than 35%. For individual degradations, a less pronounced but stable decrease is observed: for example, with severe blurring, IoU is 0.72; with high noise, 0.68; and with reduced contrast, 0.74. This confirms the results of [32], which also points to the sensitivity of YOLO models to noise and lighting, and is consistent with [39, 42], which emphasize the need for specialized methods to improve robustness. At the same time, unlike these

works, this study quantitatively demonstrates that it is precisely combined disturbances that are the determining factor in degradation, a phenomenon that had previously been insufficiently investigated.

The results of the combined scenario analysis are particularly revealing. In the baseline scenario (without occlusions and degradations), mAP=0.95 and IoU=0.95 are achieved, whereas in the most challenging scenario (small scale, critical occlusions, and combined disturbances), these values drop to 0.45 and 0.50, respectively. Thus, the overall drop in quality exceeds 50%, which confirms the conclusions [21, 24] regarding insufficient effectiveness models under challenging conditions. However, this study further demonstrates that such degradation is nonlinear in nature and results from the combined influence of several factors, which was not fully accounted for in previous works.

A comprehensive analysis of metrics across the entire dataset (350 images) shows that the average IoU is 0.82, mAP is 0.80, Precision is 0.88, Recall is 0.81, and F1-score is 0.84. The minimum values (IoU=0.48, mAP=0.45) correspond to critical scenarios, while the maximum values (up to 0.95) correspond to baseline conditions.

A standard deviation of 0.12–0.14 indicates the presence of variability, which confirms the findings in [38] regarding the dependence of results on scene conditions. Thus, the obtained results not only confirm the stability of the model but also allow for a quantitative determination of the limits of its applicability. A comparison with state-of-the-art approaches demonstrates a significant advantage of the proposed method. In particular, compared to YOLOv5 [32], where mAP is approximately 0.78, and DETR [1] with a value of 0.82, the proposed approach achieves mAP=0.95 and IoU=0.95 under combined conditions. Even under challenging conditions, the accuracy gain reaches up to 44%, which is significantly higher than the results reported in [29, 36], where improvements are typically limited to specific scenarios. This confirms that the integrated approach is more effective than individual architectural modifications. Additionally, ablation analysis results demonstrate the contribution of each component: the baseline model has an mAP of 0.75; adding multi-scale increases it to 0.82; combining it with preprocessing increases it to 0.88; and the full model reaches 0.95. This sequential trend confirms the conclusions [2, 3, 16] regarding the effectiveness of multi-scale approaches, but at the same time shows that the maximum effect is achieved only when they are integrated with other methods. Thus, the obtained quantitative results not only confirm the findings of recent studies [1–5, 7, 32], but also extend them, demonstrating that the key factor is precisely the combined effect of scaling, occlusion, and degradation. The proposed approach allows for a significant improvement in detection quality (up to 0.95 mAP) even under challenging conditions; however, the problem of a sharp drop in performance in extreme scenarios persists, confirming the conclusions [34, 35] regarding the need for further development of universal robust models.

Conclusions.

The obtained results confirm that the quality of human detection significantly depends on scale, occlusion level, and the presence of visual artifacts, with their combined effect being the most critical. It is shown that as object size decreases, occlusion increases, and degradation occurs, a significant decrease in metrics (IoU, mAP, Recall) is observed, which is consistent with current research; however, in this work, the limits of degradation were quantitatively determined, and the nonlinear nature of the factors' influence was established. At the same time, the model demonstrates sufficient stability under basic and moderately complex conditions. The proposed integrated approach, which combines multiscale representation, adaptive preprocessing, and overlap compensation, provides a significant improvement in performance, particularly in complex scenarios (up to a 44% increase). However, it has been established that even comprehensive methods do not fully eliminate the problem of critical conditions, indicating the need for further research aimed at developing universal robust models for real environments.

In the future, the obtained results can be used to improve existing and develop new intelligent visual data analysis systems capable of functioning effectively in the complex and uncontrolled conditions of a real environment.

References

1. Y. Tong et al. ACD-DETR: adaptive cross-scale detection transformer for small object detection in UAV imagery / Sensors. 2025. Vol. 25, no. 17. P. 5556. URL: https://doi.org/10.3390/s25175556 .
2. Y. Kang et al. A cross-scale feature fusion method for effectively enhancing small object detection performance / Information. 2025. Vol. 17, no. 1. P. 25. URL: https://doi.org/10.3390/info17010025 .
3. A. Wang et al. Adaptive cross-scale feature aggregation for few-shot object detection / A Neurocomputing. 2026. Vol. 669. P. 132514. URL: https://doi.org/10.1016/j.neucom.2025.132514 .
4. Q. Sun et al. AMAF-YOLO: dynamic cross-region attention and multi-scale fusion for small object detection / Nondestructive testing and evaluation. 2025. P. 1–31. URL: https://doi.org/10.1080/10589759.2025.2586076 .

5. Z. Guo et al. An enhanced framework for small object detection with middle-order interaction and adaptive cross-scale aggregation / Engineering applications of artificial intelligence. 2025. Vol. 159. P. 111730. URL: https://doi.org/10.1016/j.engappai.2025.111730 .
6. H. Zhang et al. An improved and lightweight small-scale foreign object debris detection model / Cluster computing. 2025. Vol. 28, no. 5. URL: https://doi.org/10.1007/s10586-024-05002-4 .
7. K. Kassaw et al. Are deep learning models robust to partial object occlusion in visual recognition tasks? / Pattern recognition. 2025. P. 112215. URL: https://doi.org/10.1016/j.patcog.2025.112215 .
8. D. Bian et al. A refined methodology for small object detection: multi-scale feature extraction and cross-stage feature fusion network / Digital signal processing. 2025. Vol. 164. P. 105297. URL: https://doi.org/10.1016/j.dsp.2025.105297 .
9. Q. Zhang et al. BFE-Net: bidirectional multi-scale feature enhancement for small object detection / Applied sciences. 2022. Vol. 12, no. 7. P. 3587. URL: https://doi.org/10.3390/app12073587 .
10. R. Gao et al. CAB-DETR: a lightweight small object detection algorithm based on cross-scale attention and bidirectional feature fusion / Measurement science and technology. 2026. URL: https://doi.org/10.1088/1361-6501/ae54bc .
11. L. Shang et al. CCANet: A cross-scale context aggregation network for UAV object detection / Computer vision and image understanding. 2025. P. 104472. URL: https://doi.org/10.1016/j.cviu.2025.104472 .
12. G. Zhao et al. CIDNet: Cross-Scale Interference Mining Detection Network for underwater object detection / Knowledge-Based systems. 2025. Vol. 324. P. 113902. URL: https://doi.org/10.1016/j.knosys.2025.113902 .
13. 13. G. Cao et al. Cross-DINO: cross the deep MLP and transformer for small object detection / IEEE transactions on multimedia. 2025. P. 1–12. URL: https://doi.org/10.1109/tmm.2025.3599074 .
14. S. Zhu et al. Cross-Domain object detection with hierarchical multi-scale domain adaptive YOLO / Sensors. 2025. Vol. 25, no. 17. P. 5363. URL: https://doi.org/10.3390/s25175363 .
15. M. Yang et al. Cross-Scale attention feature pyramid network for challenged underwater object detection / IEEE journal of oceanic engineering. 2025. P. 1–10. URL: https://doi.org/10.1109/joe.2024.3450532 .
16. B. Wu et al. Cross-scale feature alignment and feature enhancement for small object detection / Pattern analysis and applications. 2026. Vol. 29, no. 1. URL: https://doi.org/10.1007/s10044-025-01601-y .
17. X. Huang et al. Cross-scale resolution consistent network for salient object detection / IET image processing. 2024. URL: https://doi.org/10.1049/ipr2.13136 .
18. G. Cheng et al. CSSDet: small object detection via cross-scale feature enhancement on drone-view images / International journal of digital earth. 2024. Vol. 17, no. 1. URL: https://doi.org/10.1080/17538947.2024.2414848 .
19. Ding Y., Fan Z., Zhao Y. YOLO-Ball: real-time tennis ball detection under occlusion and motion blur. Proceedings of the institution of mechanical engineers, part P: journal of sports engineering and technology. 2026. URL: https://doi.org/10.1177/17543371261423768 .
20. Y. Gao et al. Feature super-resolution fusion with cross-scale distillation for small object detection in optical remote sensing images / IEEE geoscience and remote sensing letters. 2024. P. 1. URL: https://doi.org/10.1109/lgrs.2024.3372500 .
21. Gowthami N., Blessy S. V. Extreme small-scale prediction head-based efficient Yolov5 for small-scale object detection. Engineering research express. 2024. URL: https://doi.org/10.1088/2631-8695/ad3cb7 .
22. Huang S., Liu Q. Addressing scale imbalance for small object detection with dense detector. Neurocomputing. 2022. Vol. 473. P. 68–78. URL: https://doi.org/10.1016/j.neucom.2021.11.107 .
23. P. Ru et al. Improving small object detection via cross-layer attention / Fundamental research. 2023. URL: https://doi.org/10.1016/j.fmre.2022.09.037 .
24. S. Ma et al. LAYN: lightweight multi-scale attention yolov8 network for small object detection / IEEE access. 2024. P. 1. URL: https://doi.org/10.1109/access.2024.3368848 .
25. Y. Wang et al. Learning discriminative representations from cross-scale features for camouflaged object detection / IEEE transactions on circuits and systems for video technology. 2024. P. 1. URL: https://doi.org/10.1109/tcsvt.2024.3436148 .
26. Lee Y.-W., Kim B.-G. Attention-based scale sequence network for small object detection. Heliyon. 2024. Vol. 10, no. 12. P. e32931. URL: https://doi.org/10.1016/j.heliyon.2024.e32931 .
27. Li J., Yang L., Wang P. S.-P. STSODNet: scale transformer small object detection network. International journal of pattern recognition and artificial intelligence. 2025. URL: https://doi.org/10.1142/s0218001425550043 .
28. Li L., Li B., Zhou H. Lightweight multi-scale network for small object detection. PeerJ computer science. 2022. Vol. 8. P. e1145. URL: https://doi.org/10.7717/peerj-cs.1145 .
29. Li T.-j., Zhao H.-f. Cross-scale information enhancement for object detection. Multimedia tools and applications. 2024. URL: https://doi.org/10.1007/s11042-024-18737-4 .
30. Li Z., Singh B. Robust occluded object detection in multimodal autonomous driving: a fusion-aware learning framework. Electronics. 2026. Vol. 15, no. 1. P. 245. URL: https://doi.org/10.3390/electronics15010245 .
31. H. Lou et al. Multi-scale feature similarity and object detection for small printing defects detection / IEEE access. 2024. P. 1. URL: https://doi.org/10.1109/access.2024.3521403 .
32. T. Adli et al Robustness of YOLO models for object detection in remote sensing images /. Journal of electrical engineering. 2025. Vol. 76, no. 5. P. 429–442. URL: https://doi.org/10.2478/jee-2025-0045 .
33. J. Sun et al. Scale enhancement pyramid network for small object detection from UAV images / Entropy. 2022. Vol. 24, no. 11. P. 1699. URL: https://doi.org/10.3390/e24111699 .
34. K. Guo et al. SC-YOLO: robust multi-scale small object detection for intelligent transportation / Concurrency and computation: practice and experience. 2025. Vol. 37, no. 23-24. URL: https://doi.org/10.1002/cpe.70268 .
35. Y. Qiao et al. Small object detection using multi-scale detail enhancement and decoupled detection head / Neurocomputing. 2026. P. 133322. URL: https://doi.org/10.1016/j.neucom.2026.133322 .

36. Song J., Han C., Wu C. A small-scale object detection algorithm in intelligent transportation scenarios. Entropy. 2024. Vol. 26, no. 11. P. 920. URL: https://doi.org/10.3390/e26110920 .
37. J. Lan et al. Spatial-Transformer and cross-scale fusion network (stcs-net) for small object detection in remote sensing images / Journal of the indian society of remote sensing. 2023. URL: https://doi.org/10.1007/s12524-023-01709-w .
38. G. Cheng et al Towards large-scale small object detection: survey and benchmarks /. IEEE transactions on pattern analysis and machine intelligence. 2023. P. 1–20. URL: https://doi.org/10.1109/tpami.2023.3290594 .
39. M. Cai et al. Unsupervised anomaly detection for improving adversarial robustness of 3D object detection models / Electronics. 2025. Vol. 14, no. 2. P. 236. URL: https://doi.org/10.3390/electronics14020236 .
40. Vasanthi P., Mohan L. EFFICIENT YOLOv8 ALGORITHM FOR EXTREME SMALL-SCALE OBJECT DETECTION. Digital signal processing. 2024. P. 104682. URL: https://doi.org/10.1016/j.dsp.2024.104682 .
41. Wang K., Liu M., Ye Z. An advanced YOLOv3 method for small-scale road object detection. Applied soft computing. 2021. Vol. 112. P. 107846. URL: https://doi.org/10.1016/j.asoc.2021.107846 .
42. J. Yuan et al. YOLOv8-RD: high-robust pine wilt disease detection method based on residual fuzzy yolov8 / IEEE journal of selected topics in applied earth observations and remote sensing. 2024. P. 1–13. URL: https://doi.org/10.1109/jstars.2024.3494838 .
43. Zhang W., Liao M. Cross-scale adaptive transformer with hierarchical feature synergy for aerial small object detection. Pattern recognition. 2026. Vol. 173. P. 112822. URL: https://doi.org/10.1016/j.patcog.2025.112822 .
44. Zheng Q., Chen Y. Interactive multi-scale feature representation enhancement for small object detection. Image and vision computing. 2021. Vol. 108. P. 104128. URL: https://doi.org/10.1016/j.imavis.2021.104128

Історія статті:
Отримано: 08.04.2026 Доопрацьовано: 1.05.2026 Прийнято до друку: 23.05.2026 Опубліковано: 29.05.2026