ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ, РЕЧИ, ТЕКСТА И РАСПОЗНАВАНИЕ ОБРАЗОВ

SIGNAL, IMAGE, SPEECH, TEXT PROCESSING AND PATTERN RECOGNITION

UDC 004.93
DOI: 10.37661/1816-0301-2025-22-2-33-47

Original Article
Оригинальная статья

Efficient detection of building in remote sensing images using an improved YOLOv10 network

Xiangyi Wu[1]✉, Sergey V. Ablameyko[1, 2]

[1]*Belarusian State University,*
av. Nezavisimosti, 4, Minsk, 220030, Belarus
✉*E-mail: tigerv5872@gmail.com*

[2]*The United Institute of Informatics Problems*
of the National Academy of Sciences of Belarus,
st. Surganova, 6, Minsk, 220012, Belarus

Abstract

Objectives. At present, rapid detection of the location and size of building objects from remote sensing images is important for scientific research value and has practical significance for urban planning, environmental monitoring and disaster management.

Methods. This paper proposes an object detection method based on improved YOLOv10 network, which incorporates Super Token Attention, RepConv and Normalized Weighted Distance to more precisely detect buildings in remote sensing images. This method improves the detection accuracy and efficiency especially for small objects. The LEVIR-CD dataset is used for model training and testing.

Results. The experimental results show that the method demonstrates better accuracy on the building detection task than the traditional YOLOv10 and other methods.

Conclusion. The proposed method significantly enhances the accuracy and efficiency of building detection in remote sensing images.

Keywords: YOLOv10, remote sensing images, attention mechanism, building detection, RepConv, Super Token Attention

For citation. Wu X., Ablameyko S. V. *Efficient detection of building in remote sensing images using an improved YOLOv10 network*. Informatika [*Informatics*], 2025, vol. 22, no. 2, pp. 33–47. DOI: 10.37661/1816-0301-2025-22-2-33-47.

Conflict of interest. The authors declare of no conflict of interest.

Received | Поступила в редакцию 29.04.2025
Accepted | Подписана в печать 13.05.2025
Published | Опубликована 30.06.2025

Эффективное обнаружение зданий на изображениях дистанционного зондирования на основе улучшенной сети YOLOv10

С. Ву[1]✉, С. В. Абламейко[1, 2]

[1]*Белорусский государственный университет,*
пр. Независимости, 4, Минск, 220030, Беларусь
✉*E-mail: tigerv5872@gmail.com*

[2]*Объединенный институт проблем информатики*
Национальной академии наук Беларуси,
ул. Сурганова, 6, Минск, 220012, Беларусь

Аннотация

Цели. В настоящее время быстрое определение местоположения и размера объектов зданий с помощью изображений дистанционного зондирования имеет важное научно-исследовательское и практическое значение для городского планирования, мониторинга окружающей среды и управления стихийными бедствиями.

Методы. Предлагается метод обнаружения объектов на основе улучшенной сети YOLOv10, которая включает в себя механизм внимания Супертокен, модель RepConv (повторно параметризуемая свертка) и нормализованное взвешенное расстояние для более точного обнаружения зданий на изображениях дистанционного зондирования. Метод повышает точность и эффективность обнаружения, особенно для небольших объектов. Набор данных LEVIR-CD используется для обучения и тестирования модели.

Результаты. Экспериментальные результаты показывают, что предлагаемый метод демонстрирует лучшую точность при решении задачи обнаружения зданий, чем традиционный YOLOv10 и другие методы.

Заключение. Предлагаемый метод эффективно повышает точность и эффективность обнаружения зданий на изображениях дистанционного зондирования.

Ключевые слова: YOLOv10, изображения дистанционного зондирования, механизм внимания, обнаружение зданий, RepConv, Super Token Attention

Для цитирования. By, C. Эффективное обнаружение зданий на изображениях дистанционного зондирования на основе улучшенной сети YOLOv10 / C. By, С. В. Абламейко // Информатика. – 2025. − Т. 22, № 2. – С. 33–47. – DOI: 10.37661/1816-0301-2025-22-2-33-47.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Introduction. With the development of aerospace technology, remote sensing image resources play an increasingly important role in many fields such as urban planning, military reconnaissance, and land resource management [1]. In these applications, the detection and identification of buildings, as an important landmark feature on the ground in urban or suburban areas, has become a key topic.

Traditional object detection methods rely on empirically based feature hand design, and these methods are often difficult to mine feature information in higher dimensions of the image and have high computational costs when dealing with high-resolution remote sensing images [2]. In recent years, the development of deep learning technology has brought revolutionary advances in the field of image processing, especially in object detection, semantic segmentation and image classification, which have achieved remarkable results. However, in building detection in remote sensing images, there are still some problems especially to detect building in very high-resolution remote sensing images:

1) Different sources of images, the original pictures have different sizes of objects in remote sensing images because of different heights and angles of detectors, the influence of weather, the refraction of the atmosphere, the time of imaging, the curvature of the earth, and other factors. This increases the difficulty of object detection [3].

2) The background in remote sensing images is usually complex and varied, including roads, vegetation, lake waves, and so on. These background disturbances increase the difficulty of object detection, and high resolution remote sensing images also imply more complex backgrounds [4].

3) Occlusion problems, serious interference caused by shadows and occlusion due to shooting weather or angle, lighting conditions may change with time and weather. Changes in lighting can lead to changes in the appearance of the object, thus increasing the difficulty of object detection [5].

4) Building textures are relatively homogeneous, usually with only one or two outlines or colour information [6].

5) Remote sensing images usually have large-scale data volume and need to process a large amount of image data. This poses a challenge to the computational efficiency and storage space of object detection algorithms.

Fig. 1 illustrates the problems of building detection in several scenarios.

a) object skew *b)* imaging blur *c)* single texture *d)* extreme weather

Fig. 1. Building in different scenarios in remote sensing images

Small object detection is an important research direction and also a difficult point in object detection, there are many researchers who have conducted a lot of research on small object detection and they have proposed many methods from feature fusion, contextual connectivity, and adversarial learning for improving the performance of small object detection [7]. With the development of deep learning techniques, different attention mechanism techniques such as EfficientNetV2, CBAM, RepViTblock, and EMA attention are beginning to be applied to small object detection tasks.

EfficientNetV2 achieves faster training speed and better parameter efficiency by training perceptual neural architecture search and expansion [8]. CBAM (Convolutional Block Attention Module) enhances the sensitivity of the network to small objects through the attentional mechanism, and has shown its effectiveness in several remote sensing image detection tasks its effectiveness [9]. RepViTblock is a lightweight new backbone network [10], which revisits mobile CNNs from the perspective of ViT, showing a superior balance of latency and accuracy. EMA attention, on the other hand, enhances the model's feature extraction capability for small objects through weighted averaging [11].

With the continuous iterative updating of new technologies, object detection techniques are also evolving. Future research directions may include further improving the generalisation ability of the model, increasing technology fusion with multi-task learning, and developing more efficient training algorithms to reduce the demand for computational resources. It will also be a way to improve the performance of small object detection. Of course, small object detection in high-resolution remote sensing images is still challenging.

The YOLO series of networks have better balance detection accuracy and speed by capturing the deep and high-level features of objects. They have been selected for optimization and improvement to complete small object detection tasks in complex scenes such as satellite remote sensing data. The YOLO networks have dominated the field of object detection with its superior performance and efficiency [12]. YOLOv10, the latest in the series, has been designed to achieve state-of-the-art performance through the introduction of a consistent dual allocation and optimised model components, eliminating the need for non-maximum suppression (NMS) and significantly reducing computational overhead while achieving state-of-the-art performance [13]. However, detecting tiny objects is a very challenging problem [14].

In order to further promote the development of small object detection, this paper proposes an algorithm to optimize the backbone network and improve the attention mechanism to solve the problem of poor small object detection. In this paper, we propose to integrate STA (Super Token Attention) [15], RepConv (Re-Parameterizable Convolution) [16] and NWD (Normalized Weighted Distance) [17] techniques into the YOLOv10 model by adjusting the parameters in order to adapt the accuracy of building object detection in remote sensing images. Firstly, RepConv is applied in the convolutional layer of YOLOv10 model to simplify the network structure by reparameterizing the convolutional and batch normalisation layers. And the model parameters are reduced by RepConv to improve the inference speed and accuracy of the model. Then, the backbone network of the improved YOLOv10 model is used to extract features from the input image, and Super Tokens are generated on the feature map through the STA mechanism, and the Super Tokens are used to adjust the model's attention distribution so that the model pays more attention to the regions that contain information about buildings. Finally, the NWD loss function is introduced to calculate the weighted distance between the predicted frame and the real frame to optimise the model's ability to locate small objects and improve the accuracy of object detection for buildings.

Method

Overview of YOLOv10. YOLOv10 is the latest real-time end-to-end target detection model that significantly improves the performance and efficiency of the YOLO family by optimising post-processing and model architecture. Based on the C2f (Cross Stage Partial fusion) structure of YOLOv8, the model introduces CIB (Channel Interaction Block) to strengthen the inter-channel information interaction; and combines the spatial pyramid with the attention mechanism PSA (Pyramid Spatial Attention) to optimise the Multi-scale feature fusion; SCDown (Spatial-Channel Downsampling) module is introduced to jointly optimise the downsampling process in spatial and channel dimensions, the specific structure is shown in fig. 2. Non-maximum suppression free training is achieved and the inference delay is reduced. In addition, YOLOv10 introduces an integrated efficiency-accuracy oriented model design strategy that reduces computational redundancy and improves model capability.

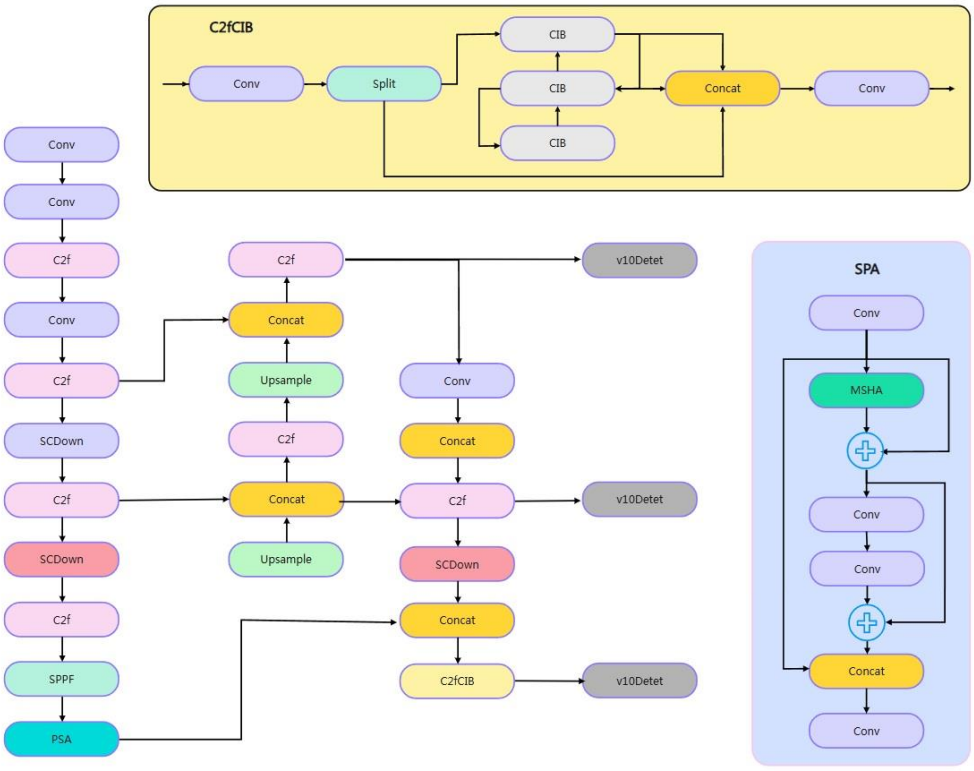

Fig. 2. YOLOv10 structure

In this study, we chose YOLOv10n as the baseline model because it provides a better balance of performance and efficiency while maintaining a smaller model size, and is suitable to be used as a starting point for improving the algorithm for further exploring and optimising the detection of building objects in remote sensing images.

Currently, buildings in remotely sensed imagery face many problems, such as: a large number of small building targets and a mixture of buildings of different sizes; and the background of remotely sensed imagery contains a variety of feature types, such as forests, fields, and roads. Together, these factors make it difficult to automatically extract building information from images. In addition, buildings have different appearances, colours and proportions, and targets such as buildings may be naturally obscured by trees etc., resulting in incomplete or partially visible targets, which is often difficult to cope with by traditional recognition methods, and this puts forward higher requirements for target recognition and localisation. Therefore, this paper proposes an improved method that fuses RepConv, Super Token Attention and NWD techniques into the YOLOv10 model, hereafter referred to as YOLO-RSTA.

RepConv. RepConv (Re-Parameterizable Convolution) is a model reparameterization technique that improves the efficiency and performance of models by optimising the network structure in the field of deep learning. Satellite remote sensing images often have complex backgrounds and varying building shapes and sizes. The standard Conv layer may not capture all the necessary features for accurate building detection. RepConv, with its multi-branch structure, can extract more diverse and comprehensive features, helping the model better distinguish buildings from the background and other objects.

The core idea of RepConv is to use multi-branched convolutional layers in the training phase, which can include convolutional kernels of different sizes (e.g., 1×1, 3×3, etc.) and possibly BN layers. In the inference phase, the parameters of these branches are reparameterized onto a master branch, usually an equivalent 3×3 convolutional layer, thus reducing computation and memory consumption and increasing inference speed [18]. This is shown in fig. 3 below.

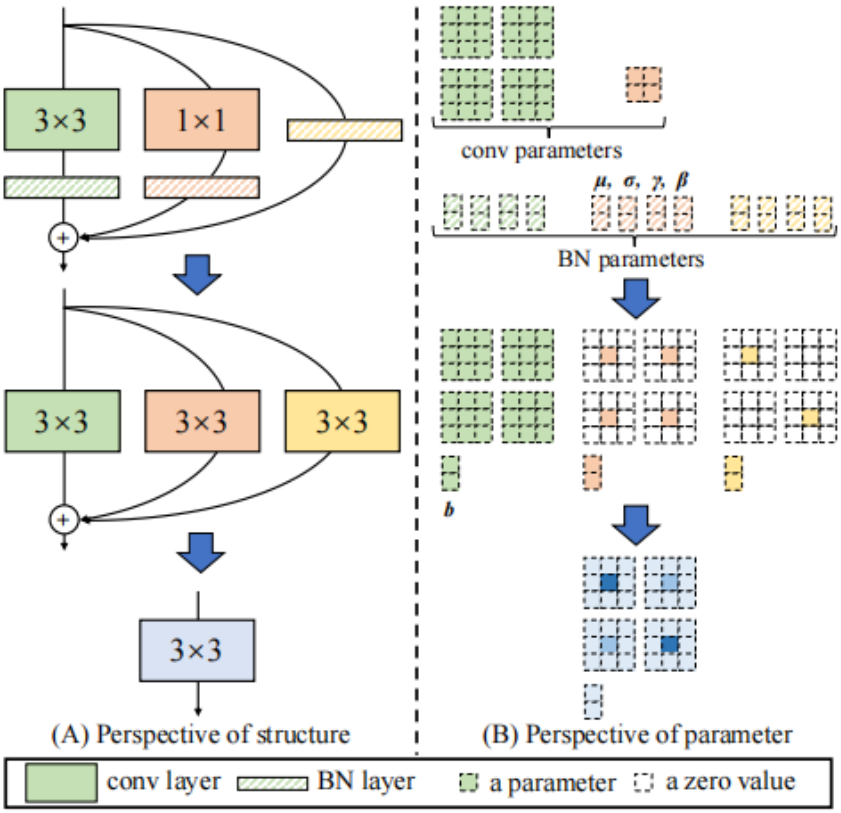

Fig. 3. RepConv Schematic Illustration

RepConv works based on the following formula:

$$W_t = T(W_b, W_r), \tag{1}$$

where W_t are transformed weights, used to manipulate input features, and W_b is (base weights), frozen weights inherited from the pre-trained model. W_r is refocusing weights, additional trainable parameters introduced by the refocusing transform. T (*) is the refocusing transform, a trainable operation applied to the base weights to generate new weights. Operation for generating new weights. RepConv also introduces residual concatenation, which makes the refocusing transform learn an increment of the base weights instead of the original mapping, similar to residual blocks in *ResNet:*

$$W_t = W_b * W_r + W_b, \tag{2}$$

where * denotes a convolutional operation. RepConv enhances the model's use of prior knowledge through the refocusing transform, allowing the model to focus on different feature representations encoded in the pre-trained model, thus improving feature extraction for small targets.

With the reparameterization technique, RepConv is able to simplify the multi-branch structure into a single-branch structure during inference, which significantly reduces the computation and memory consumption and improves the inference speed.

The improved algorithm in this paper replaces the first and third Conv convolution layers with RepConv layers in the YOLOv10 backbone network. The purpose is to use the feature map output by the previous convolution layer as input, perform a reparameterization operation through the RepConv layer, and then perform a convolution operation through the subsequent convolution layer. At this time, the number of channels of the output feature map does not change. This replacement method can reduce the size of the spatial dimension without losing information. It retains more information in the channel than traditional convolution operations, so it can effectively improve the feature extraction ability for small targets.

Super Token Attention (STA). Satellite images usually have buildings of various shapes and scales, and their edges are blurred due to shadows or vegetation. Traditional convolutional layers have limited ability to capture global dependencies, making it difficult to distinguish between buildings and backgrounds and other objects. Therefore, in remote sensing image building target detection, global context information is crucial for accurate recognition. For this reason, the fusion of STA mechanism is proposed to enhance the global perception capability of the YOLOv10 model. STA enhances the detection performance by aggregating global information in the image to form Super Tokens and using these tokens to guide the attention allocation.

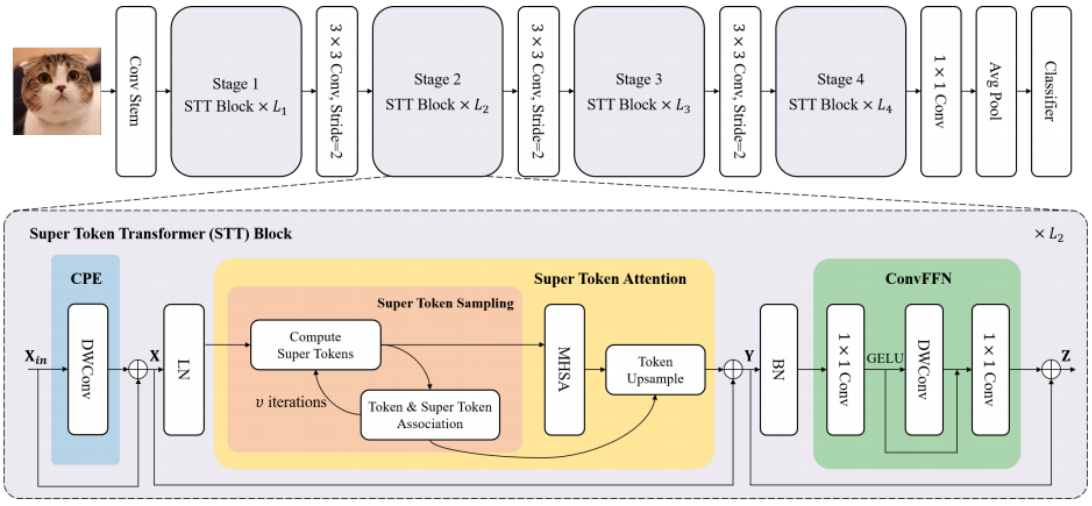

Fig. 4. Super Token Vision transformer architecture diagram

The core of the STA mechanism is to identify and utilise the global features in the image, the structure of which is shown in fig. 4. Firstly, global contextual information is extracted through the self-attention layer, which is then combined with local features to form supertags. These supertags not only contain rich semantic information, but also represent key regions in the image [19]. In the attention allocation stage, the model dynamically adjusts the attention weights according to the importance of the supertags, which enables the model to pay more attention to the regions containing building information.

The main principle of STA is divided into the following 3 steps:

Super Token Sampling (STS): the STS is a key step in the STA mechanism, which aims to aggregate the visual tokens in the input image into super tokens to reduce the number of tokens in the self-attention operation while preserving global contextual information. This process mimics the idea of superpixels, i.e., segmenting an image into regions with similar features to simplify subsequent processing. Specifically, the algorithm first generates an initial set of superlabels through average pooling, and then updates the superlabels through an iterative process. In each iteration, the algorithm calculates the association matrix between each pixel feature and the superlabelled feature, and then uses this association matrix to update the superlabels. This process can be represented by the following equation:

$$Q_t = Softmax\left(\frac{XS^{t\text{-}1^T}}{\sqrt{d}}\right), \tag{3}$$

where Q_t is the correlogram computed in the t iteration, *Softmax*() function that converts a set of values into a probability distribution. Its central role is to make all output values sum to 1 by normalizing them so that each value is between 0 and 1. X is the visual marker, $S^{t\text{-}1}$ is the supertag from the previous iteration, and d is the square root of the number of channels C for normalization. STS progressively improves the discriminative properties of Super Tokens by optimising the association matrix over multiple iterations. This process, combined with the self-attention mechanism, allows the model to reduce the computational complexity: reducing the $O(N^2)$ self-attention computation to $O(NK)$ ($K \ll N$).

Self-Attention for Super Tokens: Performs a self-attention operation in the space of super tokens to capture long-term dependencies between super tokens. This step uses the standard self-attention mechanism to compute an attention graph between super tokens, and uses this attention graph to perform a weighted summation of super tokens to obtain a new super token representation. The core formula is:

$$Attn(S) = Softmax\left(\frac{q(S)K^T(S)}{\sqrt{d}}\right)v(S), \tag{4}$$

where $q(S)$, $K(S)$ and $v(S)$ are the supertagged query, key and value, respectively, and d is the square root of the number of channels C. The *Softmax*() function for normalised correlation matrix.

Token Upsampling (TU): Finally, the learned association map is used to map the hyperlabelling back into the visual marker space. This step fuses the information from the hyperlabelling back into the original pixel features so that the model can use the global contextual information for more accurate detection. The formula is as follows:

$$TU\left(Attn(S)\right) = QAttn(S), \tag{5}$$

where Q is the association map for mapping the hyperlabelling back into the visual labelling space.

In this way, STA is able to reduce the computational complexity in self-attention while enhancing the model's ability to capture global context by aggregating local features. In remote sensing image detection, this means that the model is able to handle large-scale information in the image more effectively, improving the detection of small targets and complex scenes.

Normalized Weighted Distance (NWD). NWD is a method for measuring the similarity between predicted and real bounding boxes in object detection tasks. It evaluates the proximity of two bounding boxes by calculating a weighted distance between them. Compared to the traditional IoU (intersection and integration ratio) metric, NWD is able to capture the positional relationship between bounding boxes more accurately, especially when dealing with small objects or low overlap. The core idea of NWD is to represent the bounding box as a Gaussian distribution and then calculate the Wasserstein distance between these two distributions. This distance metric is more sensitive to changes in the position and scale of the bounding box, thus providing a more accurate similarity assessment in object detection [20].

The Wasserstein distance is used to calculate the distance between two probability distributions, i.e. to assess the minimum cost (the minimum of the mean distance of a move) required to convert one distribution into the other. For two two-dimensional Gaussian distributions $N(\mu_1,\Sigma_1)$ and $N(\mu_2,\Sigma_2)$ the Wasserstein distance between them is:

$$W_2(N(\mu_1,\Sigma_1),N(\mu_2,\Sigma_2)) = \sqrt{\left\|\mu_1-\mu_2\right\|^2 + tr(\Sigma_1+\Sigma_2-2\sqrt{\Sigma_1\Sigma_2})}, \quad (6)$$

where μ_1 and μ_2 are mean vectors, Σ_1 and Σ_2 are covariance matrices, and *tr* denotes the sum of the diagonal elements of the matrix. In NWD, this concept is used to calculate the distance between two bounding boxes corresponding to a Gaussian distribution.

Since the Wasserstein distance itself is an unbounded distance metric, NWD converts it to a similarity metric between 0 and 1 by normalizing it in exponential form, where 0 means exactly the same and 1 means completely different. To convert the Wasserstein distance into a similarity metric, NWD uses the exponential form of normalization:

$$\mathrm{NWD} = \exp\left(-\frac{W_2}{C}\right), \quad (7)$$

where C is a constant closely related to the dataset that is used to adjust the scale of normalization. In practice, C is usually set to the average absolute size of the targets in the dataset for optimal performance.

NWD provides better scale invariance, smoothing of positional deviations, and the ability to measure similarity between non-overlapping or mutually inclusive bounding boxes than traditional IoU metrics. In addition, NWD is scale-invariant, allowing it to maintain consistent performance in the detection of objects of different sizes.

The proposed method. Our proposed method YOLO-RSTA is a fusion of STA, RepConv and NWD techniques into the model of YOLOv10, and the detailed structure is shown in fig. 5.

Since the original YOLOv10 improves the model's detection ability for targets at different scales, but the detection accuracy in the background complexity and occlusion problem is not satisfactory enough. Therefore, we replace the convolution (Conv) of P1/2 and P2/4 layers with RepConv to enhance the feature extraction capability. The structure of RepConv can capture richer feature information during training, and feature extraction at different scales is more flexible and effective, and in the process of extracting feature maps at different scales, the key features of the target can be captured more effectively. Capture the key features of the target in the process of extracting feature maps at different scales.

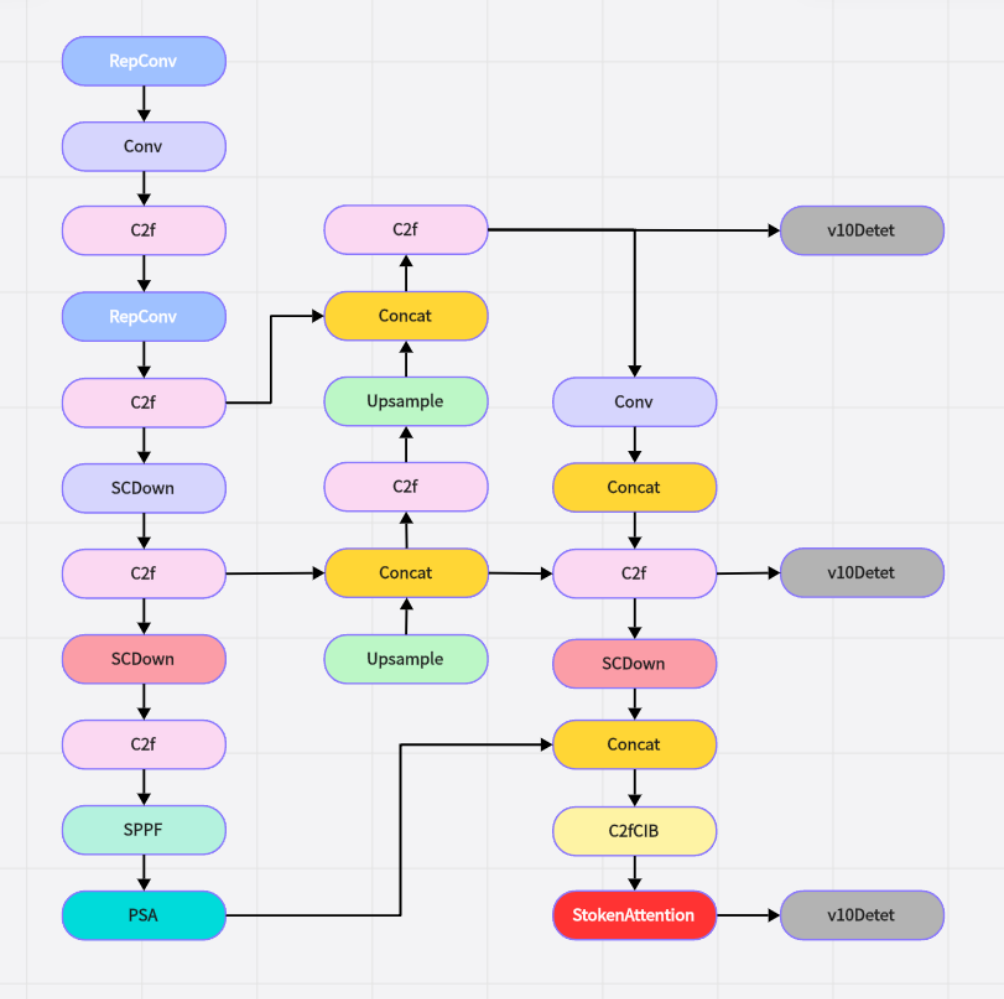

Fig. 5. Structural diagram of the improved YOLO-RSTA network

Meanwhile, we enable the NWD loss function, which will be used by the model to measure the similarity between the predicted bounding box and the real bounding box. However, the introduction of NWD leads to changes in accuracy, so the IoU ratio parameter needs to be adjusted to find the optimal balance of the loss function, which sacrifices some of the accuracy but improves on both Recall and mAP.

Finally, we added the STA attention mechanism after the P5/32 layer, which aims to improve the accuracy and efficiency in detecting small targets and processing complex scenes. In complex scenes, background interference is a major factor affecting the detection of small targets. STA reduces the interference of background noise on the model's decision making by focusing on key regions in the image. STA enhances the feature representation by applying self-attention to the super tokens and STA helps the model to capture these features at multiple scales, which improves the detection of small targets in complex backgrounds.

Experimental results

Experimental Environment Setting. In this experiment, the experiment is mainly conducted on the LEVIR-CD dataset, which contains large-scale remote sensing data. In this paper, PyTorch version 2.4.0 and the corresponding CUDA version 2.4.1 are used as the experimental environment. According to the above method, we use NVIDIA GeForce RTX 4060 as GPU for training, 445 images in the training set, 64 images in the validation set, and 128 images in the test set. The image resolution is 1024×1024. The objects in the dataset are mostly small object buildings. The specific settings are shown in fig. 6.

Python	CUDA	PyTorch	GPU	CPU
Python 3.9	12.4.1_551.78	2.4.0	NVIDIA GeForce RTX 4060	Intel Core i5-12400F

Fig. 6. Experimental environment

Data sets. Group B of the LEVIR-CD [21] dataset, a large-scale remotely sensed building change detection dataset provided by LEVIR Labs, was selected for the experiments in this study. The LEVIR-CD contains 637 pairs of very high-resolution (0.5 metres/pixel) Google Earth image blocks, with the size of each pair of images being 1024×1024 pixels, as shown in fig. 7. The LEVIR-CD covers a wide range of building types, including cottage homes, high-rise flats, small garages, and large warehouses. The LEVIR-CD dataset contains a total of 31,333 individual instances of building change. The geographic distribution of the LEVIR-CD dataset covers several cities in the state of Texas in the United States, including Austin, Lakeway, Bee Cave, Buda, Kyle, Manor, Pflugerville, Dripping Springs, etc., with image data captured from 2002 to 2018.

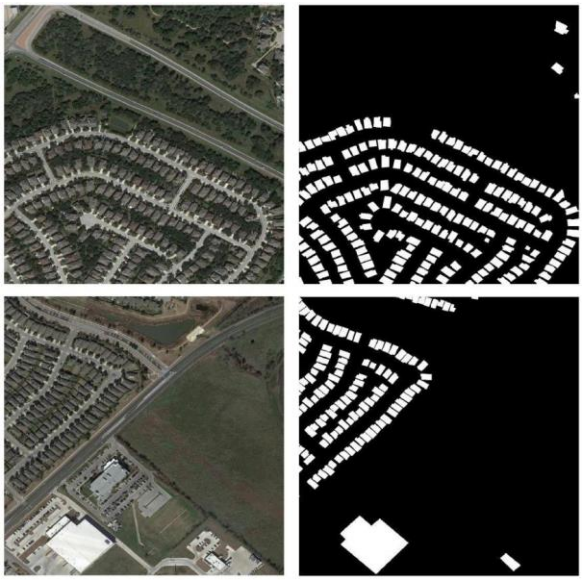

Fig. 7. LEVIR-CD dataset

Evaluation indicators. The experiments were conducted to evaluate the algorithm's target detection performance using mean accuracy (mAP), precision (*P*), recall (*R*), and parametric counts (params). The formulas for precision, recall and mAP are shown in eqs. (8) to (10).

$$P = \frac{TP}{TP + FP}, \tag{8}$$

$$R = \frac{TP}{TP + FN}, \tag{9}$$

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{10}$$

In the formula: *P* is precision; *R* is recall; *TP* is the number of samples predicted to be positive and actually positive; *FP* is the number of samples predicted to be positive but actually negative; *FN* is the number of samples predicted to be negative but actually positive; *AP* (Average Precision) is a

numerical representation of the area of a certain type of object under the Precision-Recall Curve. mAP (Mean Average Precision) is the average value obtained by averaging the average precision (*AP*) of different types of object detection. In formula (10), it is necessary to first calculate the current AP_i of different types of objects based on *P* and *R*, and take the average value of *n* AP_i as mAP.

Parameters is the size of the model's trainable parameters. The mAP50 used in the experiment represents the average detection accuracy when the IoU threshold is 0.5, while mAP50-95 represents the average accuracy of multiple thresholds in the range of IoU thresholds from 0.5 to 0.95 (step size 0.05), which can more comprehensively evaluate the performance of the model under different positioning accuracies.

Experimental results. We choose YOLOv10n as the baseline for experiments. We train for 300 epochs in the whole network structure. We need to evaluate the impact of different improvement strategies on the performance of the YOLOv10 model for building object detection in remote sensing images.

The main evaluation metrics we focus on include the model's precision, recall, and mean average precision (mAP) under different IoU thresholds. With these metrics, we can comprehensively evaluate the accuracy of the model in detecting buildings.

Experiment 1 was conducted on the LEVIR-CD dataset using YOLOv5n, YOLOv8n, YOLOv10n, and the improved method in this paper. The experimental metrics are shown in table 1.

Table 1
Results of LEVIR-CD experiment 1

Model-name	Parameters	Precision	Recall	mAP50	mAP50-95
YOLOv5n	**2 503 139**	0.884	0.817	0.9	0.577
YOLOv8n	3 005 843	0.883	0.829	0.909	**0.582**
YOLOv10n	2 694 806	0.858	0.827	0.898	0.572
YOLO-RSTA	2 957 368	**0.891**	**0.845**	**0.914**	0.571

In terms of precision, YOLO-RSTA leads with the highest value of 0.891, proving its advantage in reducing false alarms and predicting positive samples more accurately. In terms of recall, YOLO-RSTA also performs the best at 0.845, implying that it is better at identifying actual positive samples and covering the real change scenarios. mAP50 metrics, YOLO-RSTA and YOLOv8n occupy the top two positions respectively, indicating that the two models are better at detecting at the IoU threshold of 0.5, with YOLO-RSTA having a slight edge over the others. However, in the mAP50-95 metrics, YOLOv8n is outstanding with the highest value of 0.582. In summary, YOLO-RSTA performs well in the key indicators of precision, recall and mAP50.

Experiment 2 compares the proposed method YOLO-RSTA with some common attention mechanisms (EfficientNetV2, CBAM, RepViT block and EMA attention) and the results are shown in table 2 below.

Table 2
Results of building detection for LEVIR-CD dataset

Model-name	Parameters	Precision	Recall	mAP50	mAP50-95
YOLOv10n	2 694 806	0.858	0.827	0.898	**0.572**
YOLOv10n-EMA_attention	3 027 350	0.864	0.797	0.891	0.571
YOLOv10n-RepViTblock	3 027 350	0.862	0.82	0.897	0.57
YOLOv10n-CBAM	2 703 369	0.872	0.784	0.869	0.515
YOLOv10n-SE	2 695 318	0.853	0.779	0.881	0.569
YOLOv10n-EfficientNetv2	**2 544 510**	0.853	0.784	0.869	0.515
YOLO-RSTA	2 957 368	**0.891**	**0.845**	**0.914**	0.571

In this building detection experiments, our modified model demonstrates significant performance improvement compared to other models. In terms of precision, YOLO-RSTA reaches 0.891, which is about 0.8 % improvement compared to YOLOv5n and YOLOv8n, and about 3.8 % improvement compared to YOLOv10n. In terms of recall, YOLO-RSTA is a distant second with 0.845, an improvement of about 3.4 % over YOLOv5n's 0.817, and an improvement of about 2.3 % compared to YOLOv10n's 0.827. In the mAP50 metric, YOLO-RSTA is slightly higher than YOLOv8n, with an improvement of about 1.6 % over YOLOv5n's 0.9 and about 1.8 % over YOLOv10n's 0.898. Despite the slight increase in the number of parameters of YOLO-RSTA, its inference time of 1.946 seconds is still within the acceptable range. Overall, the STA attention mechanism significantly improves the model's detection accuracy and recall, albeit at the slight expense of inference speed, but this trade-off may be worthwhile in real-world applications as it achieves performance gains in key performance metrics.

From the above data, it can be seen that most of the YOLOv10 models with the introduction of different attention mechanisms can lead the original YOLOv10n model in terms of accuracy, and the model proposed in this paper improves by about 2.21 % in terms of accuracy compared to the YOLOv10n model and outperforms the other models with attention mechanisms, which also indicates that the model has a significant effect in terms of reducing false detections. In recall, the model improves by about 2.05 % over the original YOLOv10n and performs the best, indicating that the model also improves in detecting more real buildings. The model also manages to improve the mAP50 metric by about 1.67 % compared to the original YOLOv10n model, suggesting that the model can improve the detection performance overall.

Experiment 3. The proposed YOLO-RSTA method is subjected to an ablation experiment, this ablation experiment aims to explore the effects of RepConv, STA and NWD on the performance of the YOLOv10n model. The results are shown in table 3 below.

Table 3
Results of ablation experiments

YOLOv10n	Repconv	STA	NWD	Parameters	Precision	Recall	mAP50	mAP50-95
√				2 694 806	0.858	0.827	0.898	**0.572**
√	√			2 694 806	0.871	0.79	0.884	0.57
√		√		2 957 368	0.884	0.824	0.901	0.57
√		√	√	2 957 368	0.877	0.844	0.913	0.57
√	√	√		2 957 368	0.861	0.797	0.894	0.571
√	√	√	√	2 957 368	**0.891**	**0.845**	**0.914**	0.571

When only RepConv is enabled, Precision improves from 0.858 to 0.871, but Recall decreases to 0.79, indicating that RepConv improves detection accuracy but loses some of the true change detection. When only Super Token Attention is enabled, the number of parameters increases to 2 957 368, Precision improves to 0.884, and Recall is 0.824, showing that it enhances model detection. When RepConv and STA are enabled at the same time, Precision is 0.877 and Recall is 0.844, indicating that the combination of the two can effectively improve Recall and maintain a high Precision, and when RepConv and NWD are enabled, Precision is 0.861 and Recall is 0.797, which is a slight decrease in Recall compared to enabling RepConv only. When all three modules are enabled, the optimal performance is achieved with Precision of 0.891 and Recall of 0.845, but the number of parameters increases to 2 957 368. The experiments show that all three modules have positive effects on the performance improvement of the YOLOv10n model, and that all of them achieve the best results on the key metrics when they are all enabled, although they increase the number of parameters.

In order to comprehensively analyze the model efficiency, this paper compares the computational complexity indicators of the baseline model (YOLOv10n), other YOLO series models (YOLOv5n, YOLOv6n, YOLOv8n) and the improved model (YOLO-RSTA) (as shown in table 4). Specifically including:

GLFOPs (Giga Floating-Point Operations): represents the billion floating-point operations required for the model forward reasoning, used to measure the computational complexity;

Inference Time: The total processing time (unit: milliseconds) of a single image from input to output prediction box consists of the time of four stages: pre-processing time, inference time, loss time, and post-processing time for each image.

Table 4
Computational complexity experiments

Model-name	Parameters	Preprocess	Inference	Loss	Postprocess	GLFOPs
YOLO-10n	2 694 806	0.4	3.5	0	1.8	8.2
YOLO-5n	2 503 139	0.4	2.6	0	10	7.1
YOLO-6n	4 233 843	0.3	2.6	0	10.3	11.8
YOLO-8n	3 005 843	0.4	2.7	0	9.5	8.1
YOLO-RSTA	2 957 368	0.3	3.6	0	0.7	8.2

Experimental results show that the number of parameters of YOLO-RSTA increases by about 9.7 % compared with the baseline, but the GLFOPs remains unchanged at 8.2, which is close to YOLOv8n, but significantly lower than YOLOv6n, indicating that it maintains a lightweight design. The GLFOPs of YOLO-RSTA is 8.2, which is the same as YOLOv10n and YOLOv8n, and better than YOLOv6n's 11.8, proving that the improved module does not introduce additional computational burden. The inference time of YOLO-RSTA is 3.6 ms, slightly higher than YOLOv8n's 2.7 ms, but significantly better than YOLOv5n's (2.6 ms) post-processing efficiency. The post-processing optimization of YOLO-RSTA benefits from the suppression of redundant prediction boxes by the NWD loss function. The post-processing time of YOLO-RSTA is 0.7 ms, which is 61 % lower than the 1.8 ms of the baseline YOLOv10n model, and a significant improvement over YOLOv5n. YOLO-RSTA significantly improves the accuracy (mAP50: 0.914) and recall rate (0.845) while maintaining reasonable computational efficiency, making it suitable for actual remote sensing image processing scenarios.

Conclusion. This study proposes an enhanced remote sensing image building detection framework YOLO-RSTA, which innovatively integrates RepConv, super-labeled attention mechanism (STA) and optimized NWD loss function into YOLOv10. Experimental validation on the LEVIR-CD dataset demonstrates the superiority of the method, achieving the best available performance of 0.891 accuracy, 0.845 recall, and 0.914 mAP50 while maintaining an efficient parameter utilisation (2.95 M). In terms of computational complexity, YOLO-RSTA maintains the same amount of computation as the baseline model with only a 9.7 % increase in parameter size, and optimises the post-processing process with NWD, reducing the post-processing time by 61 %. This demonstrates that the proposed method improves the accuracy without obviously sacrificing the computational efficiency, and is suitable for real remote sensing image processing scenarios.

The performance improvement mainly comes from three key improvements: (1) The RepConv module enhances the feature representation ability, and the mAP50 is improved by 3.0 % compared with the baseline; (2) The STA mechanism effectively models the global contextual relationship in complex remote sensing scenes, and outperforms traditional attention modules (such as CBAM, SE, etc.) in terms of the balance between accuracy and recall; (3) The NWD loss function significantly improves the localization accuracy of small buildings, and the recall rate is improved by 2.1 %. Ablation studies confirm the synergistic effect of these components, where the full model achieves the best indicators through staged optimization.

Compared with mainstream lightweight models (YOLOv5n, v8n), YOLO-RSTA achieves higher detection accuracy with moderate parameter growth (mAP50 is 1.6 % higher than YOLOv8n). Compared with variant models using other attention mechanisms or backbone replacement, this method also shows stronger generalization ability, especially in maintaining mAP50-95 stability.

Experimental results show that YOLO-RSTA can effectively improve the model's ability to recognize building features, especially in complex backgrounds and building detection of different scales. In the future, how to further improve the model's ability to detect small objects, it is expected that this method can provide valuable references for researchers in related fields and inspire more innovative research directions.

Authors' contributions. *X. Wu* proposed an object detection method based on an improved YOLOv10 network, and conducted experimental studies including training and testing the models on the LEVIR-CD dataset. *S. V. Ablameyko* participated in summarizing, analyzing, and presenting the obtained results, which demonstrate the superiority of the proposed method over the traditional YOLOv10 and other comparative approaches.

References

1. Li S. T., Li C. Y., Kang X. D. Current status and future prospects of multi-source remote sensing image fusion. *National Remote Sensing Bulletin*, 2021, vol. 25, no. 1, pp. 148–166. DOI: 10.11834/jrs.20210259.

2. Luo H. L., Wang W. X., Ye X. Y., Zhu S. X., Bai Y. Q. Research progress on directed object detection based on deep learning. *Image and Signal Processing*, 2024, vol. 13, no. 3, pp. 258–270. DOI: 10.12677/jisp.2024.133022.

3. Abdikan S., Bilgin G., Sanli F. B., Uslu E., Ustuner M. Enhancing land use classification with fusing dual-polarized terrasar-x and multispectral rapideye data. *Journal of Applied Remote Sensing*, 2015, vol. 9, no. 1, p. 096054. DOI: 10.1117/1.JRS.9.096054.

4. Liu F. F., Zhu C. M., Zhao N. N., Wu J. H. Remote sensing small target detection based on multimodal fusion. *Laser & Optoelectronics Progress*, 2024, vol. 61, no. 24, p. 2428010. DOI: 10.3788/LOP241203.

5. Li J., Wei X. M. Research on efficient detection network method for remote sensing images based on self-attention mechanism. *Image and Vision Computing*, 2024, vol. 142, p. 104884. DOI: 10.1016/j.imavis.2023.104884.

6. Liu D., Zhong L., Wu H., Li S., Li Y. Remote sensing image super-resolution reconstruction by fusing multi-scale receptive fields and hybrid transformer. *Scientific Reports*, 2025, vol. 15, p. 2140. DOI: 10.1038/s41598-025-86446-5.

7. Li Z., Wang H., Ma G., Yang W., Ablameyko S. Effective small object detection in remote sensing images based on improved YOLOv8 network. *Nonlinear Phenomena in Complex Systems*, 2024, vol. 27, no. 3, pp. 278–291. DOI: 10.5281/zenodo.13960639.

8. Tan M., Le Q. V. *EfficientNetV2: Smaller models and faster training*, 2021. Available at: https://arxiv.org/abs/2104.00298 (accessed 13.02.2025). DOI: 10.48550/arXiv.2104.00298. (Preprint).

9. Woo S., Park J., Lee J. Y., Kweon I. S. CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.* Springer, Cham, 2018, pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1.

10. Wang A., Chen H., Lin Z., Han J., Ding G. *RepViT: Revisiting mobile CNN from ViT perspective*, 2023. Available at: https://arxiv.org/abs/2307.09283 (accessed 13.02.2025). DOI: 10.48550/arXiv.2307.09283. (Preprint).

11. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., …, Houlsby N. *An image is worth 16×16 words: Transformers for image recognition at scale*, 2020. Available at: https://arxiv.org/abs/2010.11929 (accessed 13.02.2025). DOI: 10.48550/arXiv.2010.11929. (Preprint).

12. Wang X., Zhu D., Yan Y. Towards efficient detection for small objects via attention-guided detection network and data augmentation. *Sensors*, 2022, vol. 22, no. 19, p. 7663. DOI: 10.3390/s22197663.

13. Wang A., Chen H., Liu L. H., Chen K., Lin Z. J., …, Ding G. G. *YOLOv10: Real-time end-to-end object detection*, 2024. Available at: https://doi.org/10.48550/arXiv.2405.14458 (accessed 13.02.2025). DOI: 10.48550/arXiv.2405.14458. (Preprint).

14. Wang J., Xu C., Yang W., Yu L. *A normalized Gaussian Wasserstein distance for tiny object detection*, 2021. Available at: https://doi.org/10.48550/arXiv.2110.13389 (accessed 13.02.2025). DOI: 10.48550/arXiv.2110.13389. (Preprint).

15. Huang H. B., Zhou X. Q., Cao J., He R., Tan T. N. *Vision transformer with super token sampling*, 2022. Available at: https://doi.org/10.48550/arXiv.2211.11167 (accessed 13.02.2025). DOI: 10.48550/arXiv.2211.11167. (Preprint).

16. Wan D. H., Lu R., Tian S., Xu T., Lang X., Ren Z. Mixed local channel attention for object detection. *Engineering Applications of Artificial Intelligence*, 2023, vol. 123, p. 106442. DOI: 10.1016/j.engappai.2023.106442.

17. Wang H., Ablameyko S. Enhancing small object detection in remote sensing images using mixed local channel attention with YOLOv8. *Journal of Computer Technology and Applied Mathematics*, 2024, vol. 1, no. 1, pp. 40–45. DOI: 10.5281/zenodo.10986298.

18. Ding X., Zhang X., Ma N., Han J., Ding G., Sun J. RepVGG: Making VGG-style ConvNets great again. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*, pp. 13 733–13 742.

19. Vasa V. K., Zhu W., Chen X., Qiu P., Dong X., Wang Y. *STA-Unet: Rethink the semantic redundant for medical imaging segmentation*, 2024. Available at: https://arxiv.org/pdf/2410.11578 (accessed 13.02.2025). DOI: 10.48550/arXiv.2410.11578. (Preprint).

20. Yu Z., Huang H., Chen W., Su Y., Liu Y., Wang X. *YOLO-FaceV2: A scale and occlusion aware face detector*, 2022. Available at: https://arxiv.org/pdf/2208.02019v2 (accessed 13.02.2025). DOI: 10.48550/arXiv.2208.02019. (Preprint).

21. Chen H., Shi Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 2020, vol. 12, no. 10, p. 1662. DOI: 10.3390/rs12101662.

Information about the authors

Xianyi Wu, Postgraduate Student of the Faculty of Mechanics and Mathematics of the Belarusian State University.
E-mail: tigerv5872@gmail.com
https://orcid.org/0009-0003-6976-5386

Sergey V. Ablameyko, Acad. of the National Academy of Sciences of Belarus, D. Sc. (Eng.), Prof. of the Faculty of Mechanics and Mathematics of the Belarusian State University.
E-mail: ablameyko@bsu.by
https://orcid.org/0000-0001-9404-1206

Информация об авторах

Ву Сяньи, аспирант механико-математического факультета Белорусского государственного университета.
E-mail: tigerv5872@gmail.com
https://orcid.org/0009-0003-6976-5386

Абламейко Сергей Владимирович, академик НАН Беларуси, доктор технических наук, профессор механико-математического факультета Белорусского государственного университета.
E-mail: ablameyko@bsu.by
https://orcid.org/0009-0003-6976-5386