GMTNet: Dense Object Detection via Global Dynamically Matching Transformer Network

Chaojun Dong, Chengxuan Wang, Yikui Zhai, *Senior Member, IEEE,* Ye Li, Jianhong Zhou, Pasquale Coscia, Angelo Genovese, *Senior Member, IEEE,* Vincenzo Piuri, *Fellow, IEEE,* Fabio Scotti, *Senior Member, IEEE*

***Abstract*—In recent years, object detection models have been extensively applied across various industries, leveraging learned samples to recognize and locate objects. However, industrial environments present unique challenges, including complex backgrounds, dense object distributions, object stacking, and occlusion. To address these challenges, we propose the Global Dynamic Matching Transformer Network (GMTNet). GMTNet partitions images into blocks and employs a sliding window approach to capture information from each block and their interrelationships, mitigating background interference while acquiring global information for dense object recognition. By reweighting key-value pairs in multi-scale feature maps, GMTNet enhances global information relevance and effectively handles occlusion and overlap between objects. Furthermore, we introduce a dynamic sample matching method to tackle the issue of excessive candidate boxes in dense detection tasks. This method adaptively adjusts the number of matched positive samples according to the specific detection task, enabling the model to reduce the learning of irrelevant features and simplify post-processing. Experimental results demonstrate that GMTNet excels in dense detection tasks and outperforms current mainstream algorithms. The code will be available at http://github.com/yikuizhai/GMTNet.**

***Index Terms*—Mobile Window System, Dynamic Sample Matching, Dense Object Detection, Industrial Scenarios.**

I. Introduction

THE rise of artificial intelligence has significantly enhanced industrial productivity by enabling the production of a large number of workpieces in a short time. However, traditional counting methods rely primarily on manual labor, which can be time-consuming and error-prone. Workers may become fatigued and prone to mistakes when faced with the repetitive task of counting similar workpieces for extended periods. Although object detection and recognition technologies are now widely used across various fields and have become key components of computer vision systems [1], conventional detection models struggle to perform well in dense industrial detection tasks, even when they excel in standard conditions [2]. The challenges stem from low visual contrast, uneven lighting, metal surface reflections, and cluttered backgrounds, all of which make detecting and marking workpieces difficult [3]. Additionally, dense detection tasks generate far more candidate boxes than conventional detection, which significantly increases the computational load during the matching process. This large computational demand makes it challenging to establish effective attention and interaction between different levels [4]. Figure 1 shows sample images of dense workpieces in industrial scenarios.

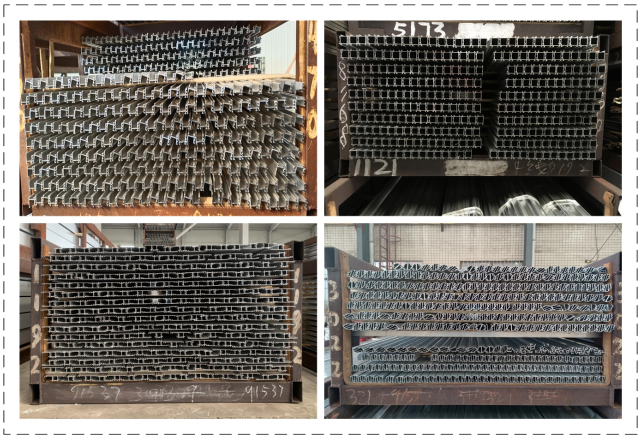


Fig. 1. Here are sample images showcasing dense workpieces in industrial scenes, illustrating typical challenges such as object occlusion, complex backgrounds, and varying lighting conditions.

Conventional detection algorithms, such as Faster R-CNN [5], SSD [6], YOLOV3 [7], RetinaNet [8], and Mask R-CNN [9], rely on generating a large number of candidate boxes for large-scale dense object detection tasks and then assigning effective positive and negative samples. Typically, traditional dense detection methods compare the intersection over union (IoU) between candidate boxes and ground truth boxes to classify samples as positive or negative based on a fixed threshold. This approach often leads to misclassification, especially in cases with ambiguous boundaries. Samples with high IoU but no actual object may be incorrectly labeled as

This study was funded by Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011576), Guangdong Science and Technology Planning Project (No. 2021A0505030080, No. 2021A0505060011), Guangdong Higher Education Innovation and Strengthening School Project (No. 2020ZDZX3031, No. 2022ZDZX1032, No. 2023ZDZX1029), Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), Guangdong Jiangmen Science and Technology Research Project (No. 2220002000246, No. 2023760300070008390). *(Corresponding author: Yikui Zhai.) (Chaojun Dong, Chengxuan Wang, Yikui Zhai, Ye Li and Jianhong Zhou contributed equally to this work.)*

Chaojun Dong, Chengxuan Wang, Yikui Zhai, Ye Li are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, Guangdong 529020, China (e-mail: cjun-dong@163.com; wangcx_UN@163.com; yikuizhai@163.com; leyelee@163.com)

Jianhong Zhou are with the School of Electronics and Information Engineering, South China University Of Technology, Guangzhou, Guangdong 510641, China (e-mail: jackychou_lab@126.com)

Pasquale Coscia, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti are with the Department of Computer Science and the Dipartimento di Informatica, Università Degli Studi di Milano, 20133 Milano, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it; vincenzo.piuri@unimi.it; fabio.scotti@unimi.it)

positive, while low IoU samples containing objects may be mistakenly classified as negative. These classification errors directly impact the model's training, resulting in suboptimal feature learning. In industrial scenarios, the background often contains significant interference, making it difficult to accurately select positive and negative samples. As a result, samples with background information are frequently mislabeled as positive, introducing noise into the training process and slowing down model convergence. Additionally, samples with valuable semantic information may be wrongly assigned as negative, leading to insufficient feature learning and reducing the model's accuracy and robustness in real-world applications. Feature matching can guide the dense sample matching process [10]. In this paper, we propose a method that calculates Top-k using adaptive IoU, allowing the algorithm to dynamically adjust the Top-k value based on different objects. This dynamic Top-k selection effectively reduces the number of candidate boxes, decreases post-processing complexity, and simplifies non-maximum suppression (NMS) processing. Moreover, the complex backgrounds in industrial environments generate more noise than in conventional settings, and dynamic Top-k selection can filter out this noise, improving detection accuracy.

On the other hand, densely placed workpieces tend to merge together, forming new composite shapes consisting of the same or different types of workpieces. We classify these workpieces based on their cut surfaces, which may change when multiple pieces are stacked. To address this challenge, we designed a network architecture that calculates the key, query, and value matrices to detect the head of the network structure. By computing the similarity score between each pair of features, the network captures global information from the feature map and dynamically adjusts the weights based on the correlation of the input features. This allows the model to assign higher weights to important features, enabling it to more accurately distinguish stacked workpieces. In summary, the contributions of this paper are as follows:

- We propose the Global Dynamic Matching Transformer Network (GMTNet), which combines a sliding window mechanism with reweighting techniques to address the challenge of inadequate global information acquisition. GMTNet first partitions the image into patches, then moves an initialized window across these patches with a fixed stride to capture both intra-patch information and inter-patch relationships. The reweighting of key-value pairs in the feature maps further enhances the relevance of global information.
- We propose a dynamic sample matching method designed to minimize the number of irrelevant samples and address the issue of excessive candidate boxes in dense detection tasks. This method adaptively adjusts the number of positive samples based on the specific detection task, enabling the model to effectively reduce the learning of irrelevant features, decrease the number of candidate boxes, and simplify post-processing.
- We propose the Global Scale Fusion (GSF) module, which captures multi-scale feature information and addresses the challenges of occlusion and overlap between objects. The GSF module captures relationships between features at various positions and facilitates interactions between global context information and object-specific features.
- We experimentally validated the effectiveness of our method in detecting dense workpieces in industrial scenarios. The results indicate that our approach surpasses other mainstream algorithms. Additionally, we conducted a comparative analysis of the detection performance between anchor-based and anchor-free methods for this task.

II. RELATED WORKS

Current detection algorithms are primarily divided into Anchor-Based and Anchor-Free approaches. In this chapter, we provide a detailed introduction to detection methods, label assignment strategies, and detectors.

A. Anchor-based and Anchor-free

Anchor-based detectors locate and classify objects by generating predefined anchor boxes on feature maps [11]–[14]. Each anchor box has a different scale and aspect ratio, designed to match objects of varying sizes and shapes. During training, anchor boxes are classified as positive or negative samples based on their Intersection over Union (IoU) with ground truth boxes. Faster R-CNN [5] introduced the Region Proposal Network (RPN) to generate multiple candidate boxes (anchor boxes), each with a fixed aspect ratio and scale. Different-sized regions are mapped onto a fixed-size feature map, and the mapped features are classified and regressed to predict the category and location of the candidate boxes. ATSS [15] generates a set of preset anchor boxes on the feature map, and these anchors are defined by various criteria. The IoU between each anchor box and the ground truth is calculated, and the most informative anchor boxes are selected as positive samples. Dense object detection tasks typically require more anchors to ensure sufficient coverage. However, the large number of densely placed anchors often leads to significant overlap, resulting in increased computational complexity. For each anchor, both classification and regression tasks must be performed, which significantly raises the computational burden, especially in high-density object scenarios.

Anchor-free detectors directly predict the location and size of objects on feature maps, eliminating the need for predefined anchor boxes [16], [17]. These methods predict the object category for each position or pixel. CornerNet [18] uses corner points to achieve object positioning and classification by predicting the upper-left and lower-right corners of each object, which are then used to determine the bounding box. It also incorporates the Feature Pyramid Network (FPN) [19] and Corner Pooling technology to predict corner positions without the need for complex anchor generation and matching processes. DETR [20] revolutionizes traditional object detection methods by replacing the conventional convolutional neural network (CNN) backbone with a transformer structure, eliminating the need for region proposal networks (RPN)

or anchor boxes. Co-DETR [21] is a DETR-based object detection method that aims to enhance the traditional DETR model through collaborative learning, especially in handling complex scenes and improving small object detection performance. It leverages multiple label assignment strategies to supervise several parallel auxiliary heads, which enhance the learning capability of the encoder. The core idea of DINO [22] is to learn powerful image representations using a self-supervised approach without requiring labeled data. It is based on the concept of knowledge distillation and leverages the Transformer architecture to perform visual tasks. Anchor-free methods simplify model design and implementation by predicting the location and size of objects directly, without predefined anchors. This approach better adapts to high-density scenarios where objects are closely packed, as it avoids the limitations of predefined anchors and the need for complex Non-Maximum Suppression (NMS) steps.

Inspired by this difference, we propose dynamically adjusting the value of top-k to allocate positive and negative samples. This approach aims to reduce processing complexity by minimizing the number of candidate boxes. Additionally, precise selection of candidate boxes effectively suppresses irrelevant background information, thereby enhancing the focus on key features.

B. Label assignment methods

In deep learning-based object detection, the distribution of samples has garnered significant attention from researchers. Traditional label assignment methods include Maximum Intersection over Union (MaxIoU) and Generalized IoU (GIoU) [23]. These methods assign positive and negative samples by calculating the overlap between predicted boxes and ground truth boxes. Among them, the MaxIoU strategy assigns the predicted box with the highest IoU value as a positive sample. While this method is simple and effective, it can lead to an uneven distribution of positive samples in dense scenarios. This is particularly problematic when objects are closely spaced or vary significantly in size, as the limited number of high-IoU predicted boxes may result in some potential positive samples being missed. On the other hand, GIoU extends IoU by considering the area difference between the minimum enclosing box of the predicted and ground truth boxes, addressing IoU's limitations in cases of no or low overlap. Although GIoU alleviates some of the shortcomings of IoU to a certain extent, it still faces challenges in dense object detection, especially in scenarios with severe object overlap or extremely small objects. In such cases, GIoU struggles to effectively distinguish objects.

Adjusting the Top-k value for the allocation of positive and negative samples can not only reduce computational costs but also enhance model performance and efficiency. Accurate sample selection is crucial during training as it directly impacts the model's learning ability and the accuracy of object detection. Researchers are actively exploring Top-k methods to adapt to various tasks and scenarios [24], [25]. The primary goal of Top-k selection is to identify the top k candidate boxes most likely to contain the object as positive samples, while the remaining boxes are classified as negative samples.

Traditional Top-k distribution strategies determine positive and negative samples based on fixed IoU thresholds. For example, algorithms such as Faster R-CNN [5], SSD [6], YOLOV3 [7], RetinaNet [8], and Mask R-CNN [9] sort a large number of candidate boxes according to fixed IoU scores relative to each ground truth box. The top K candidate boxes with the highest IoU values are selected as positive samples, while the remaining candidates are considered negative samples. However, fixed IoU thresholds can be limiting when handling objects of varying scales and shapes. To address these limitations, recent advancements include dynamic Top-k allocation. This approach enhances the flexibility and accuracy of sample selection by adaptively adjusting the K value and thresholds. Unlike fixed methods, dynamic Top-k allocation incorporates additional evaluation metrics, such as classification scores and object characteristics, into the sample selection process. This dynamic approach better accommodates variations in object shapes and sizes, improving both the accuracy and robustness of object detection. It not only enhances the model's ability to learn object characteristics but also demonstrates greater adaptability in complex scenes.

C. Object detector

In visual detection tasks, the role of a detector is to automatically identify target objects in images or videos and generate accurate bounding boxes and category labels for these objects. Adaptive Training Sample Selection Head (ATSSHead) [15] is a classification and regression head used for object detection tasks. It is part of the Adaptive Training Sample Selection (ATSS) framework, designed to improve the quality of positive and negative sample selection during training. Unlike traditional fixed IoU threshold-based label assignment, ATSS dynamically determines positive samples based on the statistical characteristics of the IoU distribution for each ground truth box. By utilizing adaptive IoU thresholds for positive sample selection, ATSS is more robust to variations in object scale, shape, and density. By focusing on anchors with the highest IoUs near each ground truth box, ATSS enhances the alignment between predictions and ground truth. It operates in an anchor-based paradigm, where predefined anchor boxes are assigned as positive or negative samples. ATSSHead excels in scenarios with diverse object scales and aspect ratios, often outperforming fixed IoU threshold-based methods such as RetinaNet. Region Proposal Network Head (RPNHead) [5] is a fundamental component of two-stage object detection frameworks such as Faster R-CNN. Its primary function is to propose a set of candidate object regions (region proposals) from an input image. These proposals are then passed to the second-stage detection head for classification and refinement. The main features of RPNHead include anchor-based proposal generation, where predefined anchor boxes of various sizes and aspect ratios are used to propose regions that likely contain objects. For each anchor, RPNHead predicts whether it belongs to the foreground (object) or background (non-object). Simultaneously, it refines the coordinates of the proposed regions to better fit the objects. RPNHead is widely used in scenarios requiring high-quality proposals and in two-stage

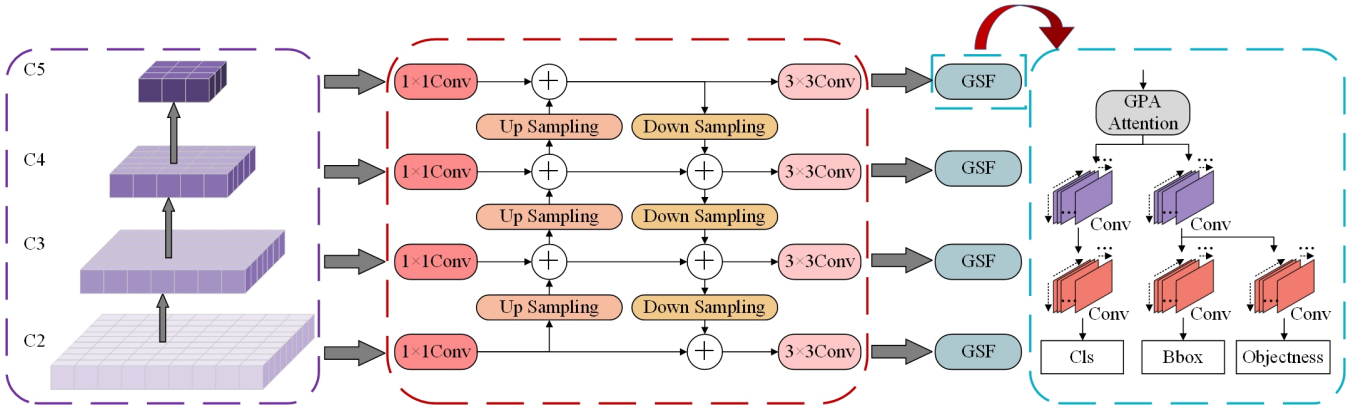


Fig. 2. The RGB image is first divided into multiple small patches, which are processed by the Transformer and transformed into feature vectors via a linear layer. Features C2-C5 are then extracted using a hierarchical structure, where upsampling and downsampling techniques are applied to fuse feature maps across different scales, enabling a more comprehensive feature representation. Following this, the Global Scale Fusion (GSF) module integrates these fused features to perform category prediction, bounding box regression, and objectness prediction through a branching structure, optimizing detection accuracy across varied object sizes and contexts.

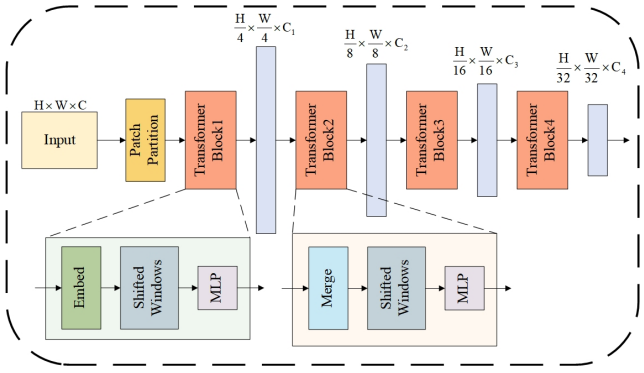


Fig. 3. The sliding window feature extractor has four Transformer layers: an initial Embed module for foundational embeddings, followed by three layers with Merge modules to aggregate features and capture complex patterns, enhancing local and global context understanding.

detectors such as Faster R-CNN, making it suitable for tasks with moderate object density and relatively large objects.

III. METHODS

Many algorithmic models perform well under standard conditions, but their detection accuracy declines significantly in dense detection scenarios. This decline is primarily due to the models' inability to effectively process multi-scale information and handle the extensive occlusion and overlap typical of dense environments, which necessitates robust global information analysis. To address these challenges, we propose the GMTNet architecture. This framework initially divides the image into patches and employs a sliding window mechanism to capture global context. Within the framework, the GSF module enhances global information representation by reweighting the key-value pairs of feature maps.

Dense detection typically generates a substantial number of candidate boxes, which increases post-processing complexity. To mitigate this, we have developed a dynamic sample matching method that adaptively adjusts the K value according to specific detection tasks. This approach significantly reduces the number of candidate boxes, decreases post-processing complexity, and minimizes interference from irrelevant positive samples.

A. GMTNet Frame Structure

To address the issue of losing detailed information in densely distributed objects and the inability to effectively capture global context, we propose the Global Dynamic Sample Matching Framework (GMTNet), which includes the Global Scale Fusion Detection Module (GSF). The overall framework of GMTNet is illustrated in Figure 2. The input RGB image first undergoes Hierarchical Feature Extraction through a feature extraction layer. This hierarchical structure consists of four feature levels, labeled C2, C3, C4, and C5, each representing features at a different scale, ranging from local details (C2) to high-level, abstract features (C5). Inspired by Feature Pyramid Networks (FPN), these multi-scale features (C2-C5) are processed with 1×1 convolutional layers to reduce dimensionality while retaining essential information. Upsampling and downsampling operations are applied to fuse feature maps across different scales, facilitating a comprehensive integration of features and enabling the network to balance local details with global context across feature levels. This multi-scale feature representation supports object detection across various sizes, even in densely packed or overlapping scenes. The Global Scale Fusion (GSF) module is a critical component of GMTNet, tasked with merging the fused features obtained from hierarchical extraction. By aggregating features across scales, the GSF module captures contextual information that spans both local and global perspectives, ensuring that fine details (important for detecting small, densely packed objects) and broader contextual information are preserved. Furthermore, the Global Position Attention (GPA) mechanism dynamically reweights spatial features, allowing the network to focus on the most relevant areas in the scene, thus improving its ability to detect objects accurately across complex backgrounds.

Our framework facilitates the extraction of global information and integration of multi-scale features, combining global

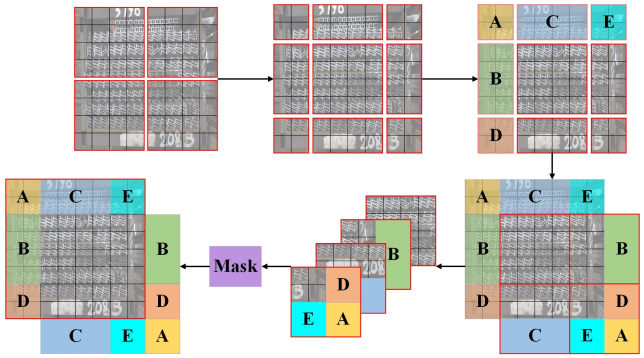


Fig. 4. The feature map is divided into $M \times M$ windows, then slid by $M/2$ pixels to create overlaps for cross-window feature fusion, enhancing global context. Overflowing windows are adjusted by shifting areas to the image edges, and a Mask module isolates regions to prevent unintended connections.

characteristics with local details to enhance the accuracy of dense object detection. During the feature extraction stage, we obtain global context information. Many detection models, such as ResNet [26], are employed for feature extraction. ResNet progressively increases feature abstraction through convolutional and pooling layers. However, for small and overlapping objects in dense scenes, fixed abstractions often fail to capture all details, leading to imbalances in the extracted multi-layered features. To address these limitations, we introduce a transformer-based [27] sliding window feature extractor [28] for capturing global multi-scale information, as depicted in Figure 3.

The Patch Partition module employs a convolutional layer to reduce the channel dimension. In the first layer of the four Transformer layers, the Embed module first segments the image into patches, which are then embedded into feature vectors before being processed by the Shifted Windows module. This module captures cross-window context information, and the resulting data is subjected to non-linear transformation using a Multi-Layer Perceptron (MLP) module. The remaining three layers follow a similar structure: they first process the data through the Merge module, which stitches elements in the row and column directions, and then through the Shifted Windows module and MLP module for information extraction and non-linear transformation. Each image, with dimensions $[H \times W \times C]$, is divided into patches, converting each patch into a feature vector before processing by the Transformer layers. These vectors are then organized into a sequence. The size of the image patches processed by each Transformer layer is $[\frac{H}{4} \times \frac{W}{4} \times C_1]$, $[\frac{H}{8} \times \frac{W}{8} \times C_2]$, $[\frac{H}{16} \times \frac{W}{16} \times C_3]$, and $[\frac{H}{32} \times \frac{W}{32} \times C_4]$. The specific sliding window processing method for images is illustrated in Figure 4. To better capture global information, the input features are initially divided into non-overlapping windows of size $M \times M$. The feature map is then slid by $M/2$ pixels (half the window size) to create new, overlapping windows, facilitating cross-window feature fusion. To address potential overflow of windows beyond the image boundaries, areas A, C, and E are shifted to the bottom of the image, and areas A, B, and D are shifted to the right side. This adjustment ensures uniform computational load across windows. To prevent unintended connections between regions, a Mask module isolates information from different areas. Finally, features C2-C5 are extracted through a hierarchical structure.

Each feature map is processed through a 1×1 convolutional layer to reduce the number of channels while preserving the spatial dimensions. Starting with the highest-level feature map, C5, the feature maps are downsampled sequentially. At each scale, the feature map is fused element-wise with the corresponding lateral connection feature map through upsampling. The fused feature maps are then processed by a 3×3 convolutional layer to further integrate features, enhancing feature fusion and reducing aliasing effects, thereby improving the accuracy of object detection.

The GSF module consists of a global perceptual attention (GPA) module and multiple convolutional layers. Upon receiving the fused multi-scale feature maps, the module first captures global information using the GPA module, thereby enhancing feature representation. The framework of the GPA module is illustrated in Figure 5. To extract features and generate the Query, Key, and Value matrices, the input feature map is initially processed with a convolution operation. A convolutional layer using a 5×5 kernel, with a stride of 2 and padding of 2, processes the input feature map $[C, H, W]$, producing an intermediate feature map with dimensions $[C^{'}, \frac{H}{2}, \frac{W}{2}]$. Subsequently, a convolutional layer with a kernel size of 1 generates the Query, Key, and Value matrices, with dimensions of $[C_q, \frac{H}{2}, \frac{W}{2}]$, $[C_q, \frac{H}{2}, \frac{W}{2}]$ and $[C_v, \frac{H}{2}, \frac{W}{2}]$, respectively. The Query and Key matrices are then flattened into two-dimensional matrices $[C_q, \frac{H}{2} \times \frac{W}{2}]$. The Query matrix is reshaped to $[\frac{H}{2} \times \frac{W}{2}, C_q]$, while the Key matrix is reshaped to $[C_q, \frac{H}{2} \times \frac{W}{2}]$. The dot product of the Query and Key matrices is computed to obtain the attention score matrix A:

$$A = Query \times Key. \tag{1}$$

The attention score matrix A has dimensions $[\frac{H}{2} \times \frac{W}{2}, \frac{H}{2} \times \frac{W}{2}]$, representing the correlation between each position and every other position. The score matrix A is then normalized and scaled using a 1×1 convolutional layer to prevent the values from becoming excessively large.

$$A = \frac{A}{\sqrt{C_q}}. \tag{2}$$

The scaled score matrix A is transformed into an attention weight matrix P by applying the softmax function.

$$P = softmax(A). \tag{3}$$

The Value matrix is reshaped into a two-dimensional matrix with dimensions $[C_v, \frac{H}{2}, \frac{W}{2}]$. The attention weight matrix P is then applied to the Value matrix, producing the weighted output feature O.

$$O = P \times Value. \tag{4}$$

The weighted output feature map O is combined with the original feature map to generate the final output feature map.

The feature maps produced by the GPA module are divided into two branches. In the first branch, a convolutional layer

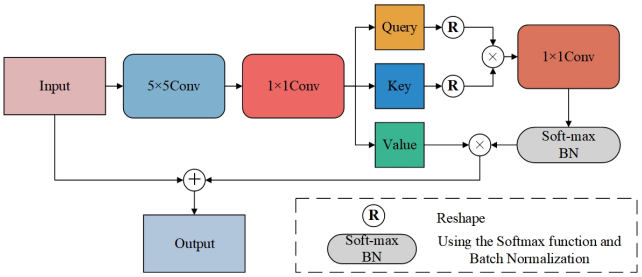


Fig. 5. The input feature map undergoes a 5×5 convolution with stride 2 for downsampling and broader receptive fields. A 1×1 convolution then produces Query, Key, and Value matrices. Flattened and reshaped for matrix multiplication, these matrices enable the model to capture spatial relationships, refining features and integrating local and global contexts effectively.

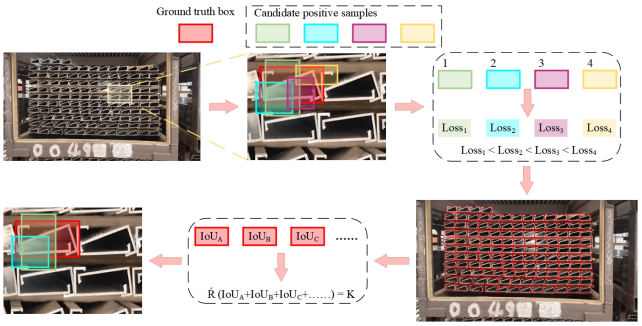


Fig. 6. First, the total loss for all candidate positive samples associated with each ground truth box is computed. Next, the adaptive IoU threshold for all ground truth boxes and their corresponding candidate boxes is determined. The function $\hat{R}()$ is used to sum these adaptive IoU thresholds and round the result to the nearest integer to establish the dynamic value of K, with a minimum value of 1. For each ground truth box, the top K candidate positive samples with the lowest total loss are then selected as the final positive samples.

adjusts the number of channels to match the number of classes, followed by another convolutional layer that generates class predictions (Cls). In the second branch, a shared convolutional layer reduces the feature map to a single channel, which indicates the presence of an object at each position. This branch then passes through separate convolutional layers to produce bounding box regression predictions (Bbox) and objectness scores (Objectness).

B. Dynamic Sample Matching Method

Previously, dense detection algorithms tackled the challenge of counting overlapping objects by avoiding the simultaneous detection of all objects in the image. Instead, they employed an iterative detection process, predicting only a subset of object boxes at a time. However, this approach often resulted in incomplete detection of overlapping or occluded objects. Detecting all objects simultaneously generates a large number of candidate boxes. Relying on a fixed Top-k selection for positive samples means that as the number of candidate boxes increases, computational costs also rise significantly. This includes tasks such as feature extraction, candidate box classification, and regression, ultimately slowing detection speed. Furthermore, the increased number of candidate boxes consumes more memory and storage, requiring more robust hardware resources. Additionally, the surplus of candidate boxes introduces redundancy and low-quality candidates, which complicates further processing and can lead to more false positives and missed detections.

To address this, we propose a Dynamic Sample Matching method that dynamically selects positive samples by adjusting the K value in the GSF module, reducing the number of candidate boxes. The process is illustrated in Figure 6. For each feature layer, the top K predicted boxes with the highest IoU values for all ground truth boxes N are selected as the initial candidate positive samples, denoted as S_1. Next, the classification and regression losses between each candidate positive sample and its corresponding ground truth box are computed. These loss values are then weighted and summed to calculate the total loss. We employ CIoU loss [27], Focal loss [8], and L1 loss. The CIoU loss function, in particular, helps prevent overlapping bounding boxes by providing a more comprehensive measurement, thereby improving the accuracy of bounding box regression. The calculation method is as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^g)}{C^2} + \alpha\upsilon. \tag{5}$$

C represents the diagonal length of the smallest enclosing box that covers both the predicted and ground truth boxes, quantifying the spatial relationship between the boxes. In dense detection tasks, this helps ensure that the size and position of the bounding boxes do not deviate from the true objects, even if the objects have significant variations in shape and size. α focuses on center alignment, and υ focuses on aspect ratio consistency. In dense detection, the balance between these two determines how much attention the model pays to the details. α controls the importance of the center distance term (ρ^2) in the overall loss. A larger α value implies a stronger constraint on center alignment. If there is less overlap between objects, α can be increased to make the model focus more on accurate object localization. υ measures the aspect ratio consistency between the predicted and ground truth boxes. This helps prevent large discrepancies in shape between the predicted and true boxes, especially when the target objects have irregular shapes. Increasing the υ value appropriately helps avoid shape mismatches in object detection. If the objects in the scene have relatively consistent shapes, υ can be lowered accordingly. The $\rho^2(b, b^g)$ represents the squared Euclidean distance between the center points of the predicted box b and the ground truth box b^g. The calculation is as follows:

$$\rho^2(b, b^g) = (x_b - x_g)^2 + (y_b - y_g)^2. \tag{6}$$

Here, (x_b, y_b) and (x_g, y_g) denote the center coordinates of the predicted box b and the ground truth box b^g, respectively.

Focal loss mitigates class imbalance challenges, particularly the disparity between foreground and background classes in object detection. It adjusts the influence of easy-to-classify samples, placing more emphasis on harder examples. The calculation is as follows:

$$L_{FL} = -\alpha(1 - P_t)^{\gamma} log(P_t). \tag{7}$$

In this context, P_t denotes the predicted probability for the object class, α (balancing factor) is used to balance the importance between positive samples (foreground) and negative samples (background). In dense detection scenarios, the background class usually dominates the dataset due to the large number of non-object regions. Setting α close to 0.25–0.5 ensures that positive samples are not overwhelmed by the background during training. γ (modulation factor) adjusts the contribution of easily classified samples to the loss. Larger values more aggressively reduce the loss from easy examples, focusing the model's learning on harder-to-classify examples. If there is significant overlap or occlusion, slightly increasing γ is usually effective, helping the model learn better from challenging instances. Too high a value of γ may result in underrepresentation of easy samples, which could slow down convergence.

For regression tasks, L1 loss is applied by measuring the absolute difference between the predicted and ground truth bounding boxes, which enhances position regression and improves localization precision. The calculation is as follows:

$$L_{L_1} = \sum_{i=1}^{n} |x_i - y_i| . \tag{8}$$

x_i represents the predicted value for the i-th bounding box, and y_i represents the true value for the i-th bounding box.

The loss values are weighted and summed to obtain the total loss TLOSS between the candidate positive sample and the corresponding ground truth box:

$$TLOSS = \lambda_{CIoU} \times L_{CIoU} + \lambda_{CIoU} \times L_{FL} + \lambda_{CIoU} \times L_{L1}. \tag{9}$$

For a single ground truth box, the mean IoU μ_j and standard deviation σ_j of the IoUs for all candidate positive samples are calculated to determine the adaptive IoU threshold T_j for the ground truth box. The calculation formula is as follows:

$$\mu_j = \frac{1}{k} \sum_{i=1}^{k} IoU(B_i, G_j). \tag{10}$$

$$\sigma_j = \sqrt{\frac{1}{k} \sum_{i=1}^{k} (IoU(B_i, G_j) - \mu_j)^2}. \tag{11}$$

$$T_j = \mu_j + \sigma_j. \tag{12}$$

where B_i denotes the i-th candidate box, G_j represents the j-th ground truth box, and k is the number of selected candidate boxes.

The sum of the adaptive IoU thresholds for all ground truth boxes is computed and rounded to the nearest integer to determine the dynamic K value, with a minimum value of 1. Based on the total loss of the preliminary candidate positive samples, the top K candidate boxes with the lowest total loss for each ground truth box are selected as the final positive samples S_2. Subsequently, corresponding labels (including class and regression objects) are assigned to these final positive samples. The remaining predicted boxes are designated as negative samples and assigned appropriate labels.

Algorithm 1 Dynamic Sample Matching Algorithm

Initialize $K \leftarrow$ Predefined constant or dynamic value
Initialize matrix $\mathcal{S}$ for storing top K candidates for each ground truth
for each ground truth box $g \in G$ **do**
    $\mathcal{K} \leftarrow$ Select top K prior boxes based on IoU for ground truth g
    Compute classification loss and regression loss for each candidate
    Total loss $\leftarrow$ Weighted sum of classification and regression loss
    Store $\mathcal{K}$ in $\mathcal{S}$
end for
for each ground truth box $g \in G$ **do**
    Compute mean IoU μ_{IoU} and standard deviation σ_{IoU} of $\mathcal{S}$
    Adaptive IoU threshold $\tau_g \leftarrow \mu_{\text{IoU}} + \sigma_{\text{IoU}}$
end for
$K_d \leftarrow$ Sum of all adaptive IoU thresholds and take integer part
for each ground truth box $g \in G$ **do**
    Final samples $\leftarrow$ Select top K_d candidates with minimum total loss for each ground truth g
end for
$G_t \leftarrow$ Index of final matched ground truth boxes
$P_r \leftarrow$ IoU values of final matched prediction boxes
return P_r, G_t

IV. EXPERIMENTS AND RESULTS

In this section, we present the experimental evaluations conducted on the dense workpiece detection dataset, WPCD. Initially, we performed ablation studies on the GMTNet network, using the optimal results as a benchmark for subsequent experiments. We then carried out ablation and comparative experiments on the proposed GPA module. To further validate the effectiveness of our method, we compared it with current mainstream detection algorithms. Finally, we analyze the experimental results and discuss the concepts of anchor-based and anchor-free approaches.

A. Dataset

Our experiments were conducted on the WPCD [29] dataset, with detailed information provided in Table I. The dataset comprises a total of 1,036 high-resolution images (ranging from 1080 to 4608 pixels) and 121,475 objects across 351 categories. On average, each image contains more than 117 objects, indicating a higher density of object arrangement within each image. The dataset was divided into five parts, with 718 images in the training set serving as key samples. Additionally, based on task difficulty, the WPCD dataset was categorized into different test levels. Level 1 represents the detection of sparsely distributed workpieces with high visibility between them. Level 2 presents a more challenging scenario with denser workpieces, where some may be occluded and visibility is reduced. Level 3 involves the most complex task, where nearly all workpieces are stacked together, further

increasing detection difficulty. Lastly, Level k encompasses all three previous scenarios, simulating the complexity of real-world applications.

TABLE I
WPCD DATASET DETAILS: DISTRIBUTION OF OBJECTS, IMAGES, AND CATEGORIES ACROSS TRAINING AND DENSITY LEVELS

Property	Train	Level1	Level2	Level3	Levelk
Objects	85087	6635	17795	13200	36388
Images	718	120	98	100	318
Categories	199	78	41	40	152

B. Evaluating Indicator

In real-world scenarios, detection accuracy is a priority for workers, so we used the mAP50 metric from COCO as a reference value. During the experimental phase, since Level 1-K represents a range of task scenarios, optimal performance in a specific scenario may depend on different key components. Therefore, it is not feasible to evaluate the model's overall performance based on a single set of data. Accordingly, we divided Level 1-K into four experimental groups (1-4). The model's scenario score (S_c) was evaluated using the following formula:

$$S_c = \sum_{i=1}^{4} \omega_i \frac{S_i - min(S_i)}{max(S_i) - min(S_i)}. \tag{13}$$

In the formula, S_i represents the score achieved by the model in group i, with $max(S_i)$ and $min(S_i)$ denoting the maximum and minimum scores attained by all models within the same group, respectively. The variable ω_i represents the weight assigned to each group, which is determined based on the practical relevance of the Level 1-K scenarios: 2, 3, 3, and 4. The simpler Level 1 scenario is given a lower weight, whereas the Level k scenarios, which more accurately reflect real-world conditions, are assigned higher weights.

C. Experimental Setup

We conducted the experiments using the MMLab framework on an Ubuntu 18.04.6 system with an NVIDIA RTX 3060 GPU. All images were resized to 640×640 pixels before being input into the algorithms. The performance of each algorithm was optimized during the experiments according to the guidelines provided in the respective papers.

D. The Ablation Experiments

1) Ablation Studies of GMTNet: We conducted ablation experiments on the network structure of GMTNet, using the FPN [30] as the neck in all experimental combinations. These experiments evaluated the performance of various backbone and head combinations in dense object detection scenarios. The ablation methods and corresponding experimental results are presented in Table II.

Among all combinations, the Swin Transformer backbone outperforms ResNet in detection accuracy across multiple levels, demonstrating its superior feature extraction capabilities, especially in dense scenes. The Swin Transformer combinations (Groups 4, 5, and 6) exhibit higher detection scores, particularly for Level1 and Levelk, indicating its ability to capture richer hierarchical information. Its hierarchical structure effectively handles objects at various scales, enhancing feature representation and robustness in complex and densely populated scenes. While ResNet performs adequately in simpler scenarios, its convolutional structure may struggle with feature extraction in dense object settings. Specifically, ResNet's limited expressive power at higher feature levels hampers its ability to capture detailed information and context in dense scenes, resulting in inferior detection performance.

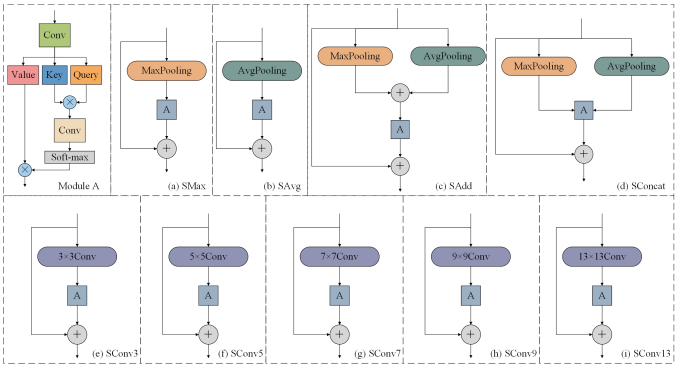


Fig. 7. Different combinations of Global Perceptual Attention (GPA) modules are tested, with consistent components across configurations referred to as Module A. These variations evaluate the impact of pooling operations and receptive field sizes on capturing global and local features in dense detection scenarios, highlighting their effect on scene scores and detection performance.

The different combinations in the table (Groups 1 to 6) represent various Backbone, Neck, and Head configurations. When ResNet [5] is selected as the Backbone (Groups 1-3), using GSF as the detection head (Group 3) yields a score of 7.49 on Sc, significantly outperforming ATSSHead [15] (5.45) and RPNHead (0.11), with notably higher detection accuracy at Level2 and Level3 (81.6 and 78.8). When Swin Transformer is used as the Backbone, the GSF detection head (Group 6) substantially outperforms all other combinations in detection performance at Level2 (85.5) and Levelk (83.0). The combination of Swin Transformer + FPN + GSF achieves a score of 11.51 on Sc, significantly surpassing all other combinations. This score demonstrates that, in dense scenes, GSF is more effective at fusing multi-scale features, enhancing global perception and preserving details in detection. This superior performance is attributed to GSF's advanced design for dense object detection tasks. GSF dynamically adjusts the Top-k parameter with adaptive IoU thresholds, thereby reducing the number of candidate boxes and minimizing the learning of irrelevant features. Additionally, it re-weights features based on their relevance within the input image, emphasizing important features while suppressing irrelevant or noisy information. This approach simplifies the post-processing stage and effectively addresses localization issues in dense object scenarios.

Compared to other combinations such as ATSSHead (Group 4: 55.839G) and RPNHead (Group 5: 66.766G), GSF with Swin Transformer (Group 6) maintained relatively low com-

TABLE II
PERFORMANCE ANALYSIS OF DIFFERENT NETWORK COMPONENT COMBINATIONS IN DENSE SCENE DETECTION

Group	Backbone	Neck	Head	Level1	Level2	Level3	Levelk	S_c	FLOPs(G)	Parameters(M)
1	Resnet [26]	FPN	ATSSHead	86.4	76.1	77.7	78.4	5.45	52.805	32.113
2	Resnet [26]	FPN	RPNHead	83.3	67.2	72.5	77.6	0.11	64.412	41.348
3	Resnet [26]	FPN	GSF	82.9	81.6	78.8	80.6	7.49	32.292	30.705
4	Swintransformer [27]	FPN	ATSSHead	89.3	76.9	**79.0**	79.3	7.68	55.839	35.552
5	Swintransformer [27]	FPN	RPNHead	**89.9**	77.3	77.4	78.9	6.88	66.766	44.746
6(Ours)	**Swintransformer [27]**	**FPN**	**GSF**	88.5	**85.5**	78.8	**83.0**	**11.51**	34.140	33.476

TABLE III
FINE-TUNE THE STRUCTURE OF THE GPA MODULE AND VALIDATE THE IMPACT OF DIFFERENT COMPONENT COMBINATIONS ON ITS PERFORMANCE

Methods	Head	Level1	Level2	Level3	Levelk	S_c	FLOPs(G)	Parameters(M)
baseline	GSF	88.5	85.5	78.8	83.0	8.34	34.140	33.476
+SMax	GSF	90.5	83.2	**79.3**	82.5	5.80	34.501	33.554
+SAvg	GSF	88.4	83.9	79.0	82.4	4.69	34.502	33.554
+SAdd	GSF	**90.9**	84.2	79.2	82.7	7.72	34.502	33.554
+SConcat	GSF	90.5	83.7	78.9	82.5	5.63	34.501	33.554
+SConv3	GSF	86.9	84.9	78.8	82.9	6.46	35.326	34.144
+Sconv5	GSF	90.2	**85.6**	79.1	**83.4**	**11.25**	36.773	35.193
+SConv7	GSF	89.2	83.9	79.2	82.5	5.83	38.943	36.766
+SConv9	GSF	88.4	84.8	77.8	82.2	2.75	41.836	38.863
+SConv13	GSF	89.3	85.3	78.9	83.2	9.36	49.791	44.630

putational costs (FLOPs = 34.140G, Parameters = 33.476M). Despite having lower FLOPs and parameter counts, Group 6 achieved the best overall performance, reflecting the efficiency of the GSF design in balancing computational cost and detection accuracy. Group 6 (Swin Transformer + FPN + GSF) demonstrated outstanding performance on Level2 and Levelk, achieving significant improvements. Swin Transformer consistently outperformed ResNet when paired with any detection head, validating its robustness in dense object detection tasks. GSF exhibited remarkable adaptability and effectiveness, delivering excellent results with both ResNet and Swin Transformer while maintaining computational efficiency.

2) Ablation Studies of GPA: To validate the effectiveness of the GMTNet network in dense workpiece detection, we used the GSF network as a baseline model and conducted experiments with various configurations of the GPA module, as illustrated in Figure 7. The results are detailed in Table III. These experiments demonstrate that introducing different GPA module variants leads to varying degrees of performance improvement.

Compared to the baseline, SMax improves detection accuracy on Level1 and Level3 but reduces performance on Level2 and Levelk. Its overall S_c = 5.80 indicates that it does not consistently improve global detection performance. Similar to SMax, SAvg achieves relatively balanced performance, but the overall scene score is lower (S_c = 4.69), suggesting limited improvement compared to the baseline. The structures of SMax and SAvg are shown in Figure 7(a) and (b), respectively, indicating that relying solely on a single pooling operation does not sufficiently enhance detection performance in dense scenarios. While these structures capture global features and provide significant information, they miss crucial local details. SAdd (Figure 7(c)) achieved the best Level1 accuracy (90.9) among all variants while maintaining competitive scores at other levels. Its scene score (S_c = 7.72) demonstrates strong adaptability to dense scenes. SConcat (Figure 7(d)) delivered competitive results but did not show significant improvement compared to the baseline. These combination methods partially mitigate the limitations of single pooling operations, however, simple concatenation does not fully exploit the complementarity of the two types of features. Due to inadequate feature fusion, the effectiveness of the structures in Figure 7(c) and (d), which incorporate two pooling layers, is limited.

The impact of larger convolution kernels in the GPA module highlights a clear trade-off between computational cost and performance. The structure of SConv3, as shown in Figure 7(e), maintains relatively lower performance across all levels. This illustrates that its small receptive field may fail to capture sufficient global and contextual information in dense scenes, leading to suboptimal detection performance. SConv5 demonstrates the most significant overall performance improvement with an S_c of 11.25, achieving the best scores for Level2 and Levelk. This variant effectively balances accuracy and computational cost, making it a strong candidate for dense detection tasks. Its structure is shown in Figure 7(f), the structure in Figure 7(f) shows that using a 5×5 convolutional kernel optimally balances receptive field size and computational complexity, effectively capturing both object details and global information in dense scenarios. In contrast, the structures in SConv7 (Figure 7(g)), SConv9 (Figure 7(h)) and SConv13 (Figure 7(i)) suffer from excessively large receptive fields, resulting in overly smoothed features and loss of important local information, which adversely impacts detection performance.

The table underscores the importance of balancing detection accuracy and computational efficiency. The SConv5 variant emerges as the optimal choice, delivering the best overall performance (S_c = 11.25) while keeping FLOPs (36.773G) and parameters (35.193M) at manageable levels. Larger kernels

TABLE IV
COMPARE THE GPA MODULE WITH OTHER ATTENTION MECHANISMS TO VALIDATE ITS EFFECTIVENESS IN DENSE INSPECTION TASKS

Methods	Level1	Level2	Level3	Levelk	S_c	FLOPs(G)	Parameters(M)
baseline	88.5	85.5	78.8	83.0	6.63	34.140	33.476
+Cbam [31]	**91.0**	83.2	78.5	82.4	3.50	34.142	33.484
+GatherExcite [32]	89.1	83.9	**79.7**	82.6	6.36	34.141	33.480
+ECANet [33]	89.3	84.4	79.2	82.5	5.64	34.141	33.476
+SENet [34]	88.5	84.3	78.6	81.8	1.63	34.141	33.492
+GPA(Ours)	90.2	**85.6**	79.1	**83.4**	**9.86**	36.773	35.193

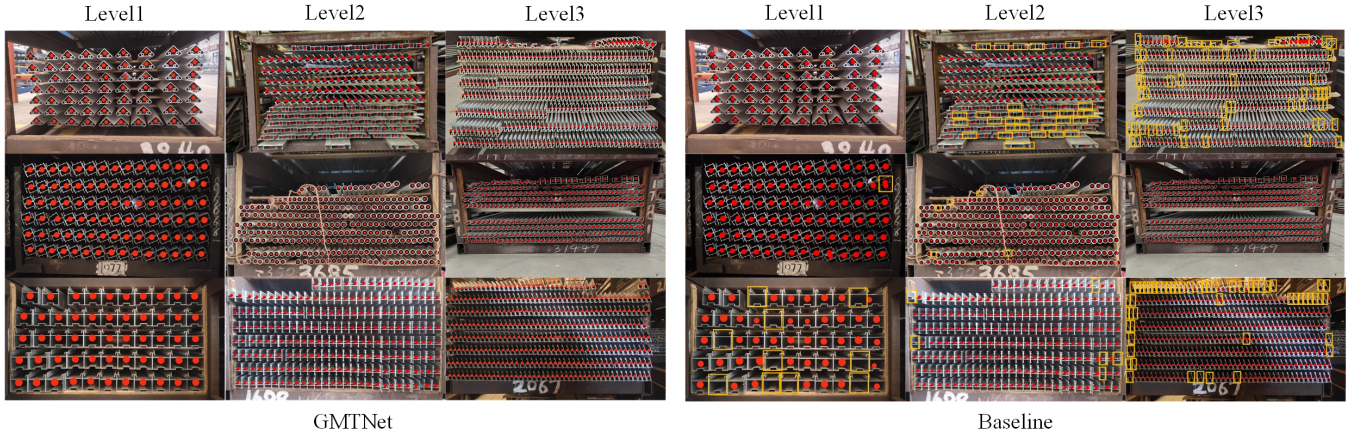


Fig. 8. The detection results of GMTNet and the baseline on level1-k. The red dots in the figure represent the detection results of the algorithm model, while the yellow boxes indicate false detections and missed detections.

TABLE V
COMPARE THE PERFORMANCE OF DYNAMIC TOP-K WITH OTHER LABEL ASSIGNMENT STRATEGIES TO VALIDATE THE EFFECTIVENESS OF DYNAMIC TOP-K IN DENSE DETECTION TASKS

Methods	Level1	Level2	Level3	Levelk
baseline	88.5	85.5	78.8	83.0
+MaxIoUAssigner	90.8	79.1	78.1	80.6
+ATSSAssigner [15]	86.4	76.1	77.7	78.4
+GIoUAssigner [23]	89.9	77.2	78.9	79.6
+Dynamic Top-k(Ours)	**91.0**	**86.4**	**79.3**	**85.8**

like SConv13, while offering slight accuracy improvements, come at the cost of significantly increased computational complexity. Variants with larger convolutional kernels (SConv5, SConv13) tend to perform better on Level2 and Levelk, highlighting their ability to capture global context and enhance feature representation. Smaller kernels or simpler operations (SMax, SAvg) can moderately improve performance but may lack the capacity to effectively handle dense and complex scenes. Overly large kernels (SConv9, SConv13) may lead to overfitting and introduce unnecessary computational overhead, making them less practical for real-world applications.

E. Comparison Experiments

The proposed method, when combined with the GPA module, achieved mAP50 scores of 90.2%, 85.0%, 79.1%, and 83.4% on the WPCD dataset for Level1-k. To validate its effectiveness, we conducted comparative experiments with other modules and mainstream algorithms, using optimal parameters in all cases.

1) Comparative study of attention modules: As shown in Table IV, we incorporated various attention modules—CBAM [31], GatherExcite [32], ECANet [33], and SENet [34]—at the same position within the GSF network. The GPA module significantly outperforms these other attention modules. This suggests that the GPA attention module, by employing dot product operations to enhance the richness and discriminative power of feature representations, more effectively addresses the interference from dense, crowded, and stacked objects. CBAM achieves the highest Level1 score (91.0) among all methods, but its overall scenario score (S_C = 3.50) is relatively low, indicating poor performance in global detection tasks. Gather-Excite performs well, particularly excelling in Level3 (79.7), demonstrating its ability to capture features of stacked or dense objects. However, its scores in Level1 and Levelk are slightly lower compared to GPA. Similar to CBAM, Gather-Excite introduces minimal computational overhead, making it a viable option for low-resource scenarios. ECANet shows slight improvements over the baseline across all levels. Its scores in Level3 (79.2) and Levelk (82.5) suggest decent capability in handling dense scenarios. With a scenario score (S_C = 5.64), ECANet outperforms SENet but does not achieve significant improvements in global performance. Its lower FLOPs and parameter count make it efficient, though not outstanding in terms of accuracy. SENet provides minimal improvements, particularly in Levelk (81.8), and a very low scenario score (S_C = 1.63), indicating that it may not be suitable for dense detection tasks. GPA achieves the best scores

TABLE VI
STUDY THE PERFORMANCE IMPACT OF THE DYNAMIC TOP-K MATCHING METHOD IN DETECTION NETWORKS FOR DENSE DETECTION TASKS

Methods	Using the dynamic Top-k matching method	Level1	Level2	Level3	Levelk	FLOPs(G)	Parameters(M)
baseline	×	88.5	85.5	78.8	83.0	34.140	33.476
	✓	91.0	86.4	79.3	85.8	34.140	33.476
+GPA	×	90.2	85.6	79.1	83.4	36.773	35.193
	✓	**93.1**	**86.1**	**80.3**	**85.6**	36.773	35.193

TABLE VII
PERFORMANCE COMPARISON OF GMTNET WITH MAINSTREAM DETECTION ALGORITHMS ON THE WPCD DATASET

Methods	Level1	Level2	Level3	Levelk	FLOPs(G)	Parameters(M)
YOLOX [35]	69.1	42.0	56.4	51.7	55.839	35.552
YOLO-R [36]	36.7	32.0	37.9	33.4	36.879	25.326
YOLOv7 [37]	17.0	10.0	4.4	25.3	104.55	37.314
Faster RCNN [5]	70.4	41.4	54.3	52.2	64.412	41.348
Mask RCNN [9]	82.7	40.6	50.2	51.9	845.75	35.963
SOLOv2 [38]	79.9	26.2	25.1	37.8	116.19	46.593
BlendMask [39]	66.1	42.3	55.8	51.7	223.49	43.252
BoxInst [40]	82.2	37.5	20.0	29.9	101.63	35.143
CondInst [41]	82.2	36.3	47.1	49.3	89.787	34.164
DETR [20]	66.1	30.5	44.0	41.6	25.15	41.575
VitDet [42]	72.3	42.9	56.4	52.3	285.26	85.439
WPCNet [29]	82.1	47.5	63.4	59.4	52.566	32.293
ATSS [15]	86.4	76.1	77.7	78.4	52.805	32.113
CenterNet [43]	73.1	42.3	56.6	53.2	53.81	32.293
Dab_detr [44]	74.5	43.6	70.5	57.1	29.112	43.702
VFnet [45]	71.0	43.5	55.7	52.0	182.13	32.892
Retinanet [8]	73.3	40.1	52.9	72.3	62.694	37.969
FCOS [46]	71.0	42.5	57.3	53.2	53.823	32.297
YOLOv3 [7]	76.9	42.0	57.1	55.1	46.389	61.949
DINO [22]	90.6	72.6	**80.3**	78.0	80.626	47.702
Deformable_detr [47]	65.4	31.0	51.5	44.4	51.631	40.119
GMTNet(Ours)	**93.1**	**86.1**	**80.3**	**85.6**	36.773	35.193

in Level2 (85.6) and Levelk (83.4), indicating its effectiveness in capturing both global and hierarchical contexts. Although its Level1 score (90.2) is slightly lower than CBAM, its high scores across other levels make it the most balanced and effective method overall. GPA achieves the highest scenario score (S_C = 9.86), confirming its superior adaptability and robustness in dense and challenging detection scenarios. While GPA has higher FLOPs (36.773) and parameters (35.193) compared to other methods, this computational cost is justified by its significant performance gains.

2) Comparative study of sample matching methods: We conducted comparative experiments on different sample matching methods, with results shown in Table V. MaxIoUAssigner performs well in Level1 (90.8) but struggles in Level2 (79.1) and Levelk (80.6), showing limitations in dense detection tasks. ATSSAssigner performs the worst, especially in Level2 (76.1) and Levelk (78.4), indicating it's not suitable for complex scenarios. GIoUAssigner shows slight improvements in Level1 (89.9) and Level3 (78.9), but its performance drops in Level2 (77.2) and Levelk (79.6), making it less competitive. The Dynamic Top-k method outperforms all other methods, achieving the best results across all levels, particularly in Level2 (86.4) and Levelk (85.8), highlighting its adaptability and robustness for dense detection tasks.

Through our experiments, we identified the optimal structure for the GMTNet network. By incorporating the GSF module, GMTNet introduced a dynamic sample matching method that performs dynamic Top-k matching across all samples. This approach reduces the number of candidate boxes, decreases the complexity of the post-processing stage, and simplifies the non-maximum suppression (NMS) process. Additionally, it effectively mitigates noise from complex industrial backgrounds, enhancing detection accuracy. Furthermore, the integration of the GPA module improved mAP50 on Level1-k by 1.7%, 0.1%, 0.3%, and 0.4%, respectively. These improvements demonstrate that the GPA module successfully re-weights features to emphasize important information while suppressing irrelevant data, thereby capturing global contextual information and object relationships more effectively. This addresses challenges such as varying shooting angles, object occlusion, and workpiece stacking.

3) Performance study of the Dynamic Top-k Matching Method: Table VI presents the performance evaluation of various methods, both with and without the dynamic Top-k matching approach, across Levels 1 through k. The table evaluates the impact of the Dynamic Top-k Matching Method on detection performance across different levels and two settings. Introducing the Dynamic Top-k Matching Method into the baseline significantly improves detection across all levels, with notable gains in Level1 (+2.5), Level2 (+0.9), Level3 (+0.5), and Levelk (+2.8). Adding the GPA module further enhances the performance of the baseline, especially in Level1 (+1.7), Level3 (+0.3), and Levelk (+0.4), reflecting its strength in feature representation for dense and complex scenarios.

Combining GPA with the Dynamic Top-k Matching Method yields the best results across all levels, showing significant improvements compared to the setting with only GPA. The integration of the Dynamic Top-k Matching Method amplifies the advantages of GPA.

4) Comparison of GMTNet with other detection methods: We compared the proposed method with other mainstream algorithms on the WPCD dataset. The experimental results, presented in Table VII, indicate that our method achieved mAP50 values of 93.1%, 86.1%, 80.3% and 85.6% on Level1-k, respectively. This performance surpasses that of the other algorithms on Level1-k, further validating the effectiveness of our approach. The visualization comparison of detection results between GMTNet and the baseline is shown in the Figure 8.

F. Discussion

In dense detection tasks, there is a significant performance difference between anchor-based and anchor-free methods. We analyzed the performance data of various methods across different levels of detection tasks (Level1-k).

Anchor-based methods excel in detection tasks across varying densities. These methods use predefined anchor boxes to locate objects, allowing them to cover a range of sizes and aspect ratios. This approach provides greater robustness against significant variations in object scale. For instance, ATSS achieves strong performance at all levels, with scores of 86.4%, 76.1%, 77.7%, and 78.4% for Level1 through Levelk, respectively. Anchor-based methods enhance object-anchor alignment, which improves localization accuracy. Faster R-CNN and RetinaNet demonstrate effective anchor alignment strategies at Level1, with scores of 70.4% and 73.3%, respectively. RetinaNet's performance at Levelk, with a score of 72.3%, highlights its adaptability in complex scenarios.

Anchor-free methods directly regress object bounding boxes, which theoretically allows them to adapt to a broader range of objects. However, the absence of prior information from anchors results in greater performance fluctuations when there is significant variation in object scale. For instance, YOLOX achieves a high performance at Level1 (69.1%) but shows a drop at Level2 and Levelk (42.0% and 51.7%). While anchor-free methods offer higher computational efficiency by avoiding the need to generate numerous anchors, they may struggle with varying object complexities. For example, SOLOv2 performs well at Level1 (79.9%) but performs poorly at Level2 and Level3 (26.2% and 25.1%), indicating difficulties in handling objects of varying complexity. DINO excels at all levels, particularly Level1 and Level3, due to its Transformer-based architecture, which captures both global and local features. However, it has higher computational costs compared to simpler models like YOLOX and YOLO-R. Additionally, anchor-free methods encounter challenges in feature matching due to the lack of anchor references, potentially leading to unstable detection results. DETR demonstrates reasonable performance at Level1 (66.1%) but experiences a decline at Level2 and Level3 (30.5% and 44.0%), underscoring the impact of feature matching issues.

GMTNet achieves outstanding performance across Level1 to Levelk while maintaining relatively low FLOPs (36.773G) and parameters (35.193M). This balance between computational efficiency and performance demonstrates the effectiveness of GMTNet's design in dense detection scenarios. Some mainstream detection algorithms, such as YOLO-R (36.879G, 25.326M) and Dab_detr (29.112G, 43.702M), have comparable FLOPs or parameter counts but exhibit significantly lower performance across all levels, especially in dense scenarios, indicating inefficient utilization of computational resources. GMTNet's lower computational cost makes it a practical choice for industrial scenarios or edge devices, where computational resources may be limited.

V. CONCLUSION AND FOR THE FUTURE WORK

We proposed a global dynamic matching transformer network for dense object detection. This network adjusts the Top-k value dynamically by calculating an adaptive IoU threshold to select appropriate samples. To address challenges such as varying camera angles, potential occlusion, and overlapping target objects, we introduced the GPA module. This module extracts global semantic information from the image by performing dot product operations on the Key, Query, and Value components of the feature map, effectively mitigating these issues.

Our proposed dynamic sample matching method achieved the best performance across all levels, demonstrating its effectiveness in improving both sample matching and feature extraction. By dynamically adjusting the Top-k value through the adaptive IoU threshold and integrating the GPA module, this method significantly enhances performance in dense detection tasks.

In the future, we will further explore integrating the advantages of both anchor-based and anchor-free methods to improve performance in dense detection, especially for complex scenes and multi-scale objects.

REFERENCES

[1] S.-Y. Wang, Z. Qu, and C.-J. Li, "A dense-aware cross-splitnet for object detection and recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 5, pp. 2290–2301, 2022.

[2] F. Ji, Y. Chen, L. Liu, and X.-T. Yuan, "Cross-domain few-shot classification via dense-sparse-dense regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[3] T. Guan, C. Gu, C. Lu, J. Tu, Q. Feng, K. Wu, and X. Guan, "Industrial scene text detection with refined feature-attentive network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6073–6085, 2022.

[4] Z. Li, Y. Liu, B. Li, B. Feng, K. Wu, C. Peng, and W. Hu, "Sdtp: Semantic-aware decoupled transformer pyramid for dense image prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6160–6173, 2022.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[7] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[10] X. Dong, J. Shen, and L. Shao, "Hierarchical superpixel-to-pixel dense matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2518–2526, 2016.

[11] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "Foveabox: Beyound anchor-based object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389–7398, 2020.

[12] H. C. De Vet, R. W. Ostelo, C. B. Terwee, N. Van Der Roer, D. L. Knol, H. Beckerman, M. Boers, and L. M. Bouter, "Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach," *Quality of life research*, vol. 16, pp. 131–142, 2007.

[13] W. Ke, T. Zhang, Z. Huang, Q. Ye, J. Liu, and D. Huang, "Multiple anchor learning for visual object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 206–10 215.

[14] X.-T. Vo and K.-H. Jo, "A review on anchor assignment and sampling heuristics in deep learning-based object detection," *Neurocomputing*, vol. 506, pp. 96–116, 2022.

[15] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.

[16] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[17] M. Zand, A. Etemad, and M. Greenspan, "Objectbox: From centers to boxes for anchor-free object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 390–406.

[18] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[21] Z. Zong, G. Song, and Y. Liu, "Detrs with collaborative hybrid assignments training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 6748–6758.

[22] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[23] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[24] X. Chen, S. Gopi, J. Mao, and J. Schneider, "Optimal instance adaptive algorithm for the top-k ranking problem," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6139–6160, 2018.

[25] A. S. Silberstein, R. Braynard, C. Ellis, K. Munagala, and J. Yang, "A sampling-based approach to optimizing top-k queries in sensor networks," in *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, 2006, pp. 68–68.

[26] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[29] Z. Jiang, Y. Zhai, F. Ke, J. Zhou, A. Genovese, V. Piuri, and F. Scotti, "Learning to count arbitrary industrial manufacturing workpieces," *IEEE Transactions on Industrial Informatics*, 2024.

[30] B. Koonce and B. E. Koonce, *Convolutional neural networks with swift for tensorflow: Image recognition and dataset categorization*. Springer, 2021.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[35] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[36] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," *arXiv preprint arXiv:2105.04206*, 2021.

[37] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464–7475.

[38] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural information processing systems*, vol. 33, pp. 17 721–17 732, 2020.

[39] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8573–8581.

[40] Z. Tian, C. Shen, X. Wang, and H. Chen, "Boxinst: High-performance instance segmentation with box annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5443–5452.

[41] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 282–298.

[42] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *European conference on computer vision*. Springer, 2022, pp. 280–296.

[43] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[44] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329*, 2022.

[45] A. Ahmed, P. Tangri, A. Panda, D. Ramani, and S. Karmakar, "Vfnet: A convolutional architecture for accent classification," in *2019 IEEE 16th India Council International Conference (INDICON)*. IEEE, 2019, pp. 1–4.

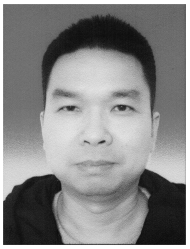
[46] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: A simple and strong anchor-free object detector," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 4, pp. 1922–1933, 2020.

[47] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

Chaojun Dong received the B.S. and M.S. degrees in thermal Engineering from the Central South University of Technology, Hunan,China, in 1990 and 1993,respectively, and the Ph.D. degree in Control Science and Engineering from Xi'an Jiaotong University , Shanxi,China, in 2006.He is a Professor with the Facility of Intelligence Manufacture, Wuyi University. His research interests include intelligent information processing, AI, and Intelligent traffic detection and control.

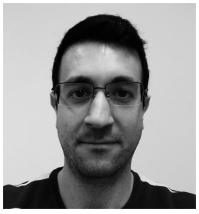
Chengxuan Wang is currently pursuing a master's degree at the School of Electronic and Information Engineering, Wuyi University. His research interests include computer vision, image processing, and object detection.

Yikui Zhai (Senior Member, IEEE) received the B.S. and M.S. degrees from Shantou University, Guangdong, China, in 2004 and 2007 respectively, and the Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013. Since October 2007, he has been working with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, China, where he is a Professor. He has been a Visiting Scholar with the Department of Computer Science, University of Milan, Milan, Italy, since 2016. His research interests include image processing, deep learning, pattern recognition, object detection, Unmanned Aerial Vehicle (UAV) change detection, and selfsupervised learning.



Ye Li currently, he is an engineer in the Intelligent Manufacturing Department of Wuyi University. Published many core papers, with research direction in control engineering and automation.

Jianhong Zhou (Member, IEEE) is currently pursuing the Ph.D. degree in the School of Electronics and Information Engineering, South China University of Technology. He received the B.S. and MS. degree in Wuyi University of Communication Engineering and Information and Communication Engineering in 2020 and 2024. His research interests include computer vision, representational contrastive learning, image processing and object detection.



Pasquale Coscia received the Ph.D. degree in Industrial and Information Engineering from the Università degli Studi di Milano, Italy, in 2019. From 2019 to 2022, he was a post-doctoral researcher at the Università degli Studi di Padova. Dr. Coscia has received the 2021 Best Paper Award of the Elsevier's Image and Vision Computing Journal for his work on human motion prediction. Additionally, he is a member of the program committees for numerous IEEE international conferences and workshops. Since 2023, he has held the position of Co-chair for the Intelligent Measurement Systems Technical Committee (TC-22) within the IEEE Instrumentation and Measurement Society. His main research interests include the areas of applied aspects of computational intelligence for signal and image processing, computer vision, machine learning and probabilistic models applied to multiagents systems.

Angelo Genovese received the Ph.D. degree in computer science from the Università degli Studi di Milano, Crema, Italy, in 2014. He has been a post-doctoral Research Fellow in computer science with the Università degli Studi di Milano since 2014. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters. His current research interests include signal and image processing, three-dimensional reconstruction, computational intelligence techn-ologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems. Dr. Genovese is an Associate Editor of the Journal of Ambient Intelligence and Humanized Computing (Springer).

Vincenzo Piuri (Fellow, IEEE) received the Ph.D. degree in computer engineering from Politecnico di Milano, Milan, Italy, 1989.,He was an Associate Professor at Politecnico di Milano from 1992 to 2000 and a Visiting Professor at the University of Texas at Austin, Austin, TX, USA (summers 1996–1999). He has been a Full Professor in computer engineering since 2000 and the Director of the Department of Information Technology, Università degli Studi di Milano, Milan, from 2007 to 2012. His main research interests are biometrics, signal and image processing, pattern analysis and recognition, theory and industrial applications of neural networks, machine learning, intelligent measurement systems, industrial applications, fault tolerance, digital processing architectures, embedded systems, cryptographic architectures, and arithmetic architectures. He has participated in several national and international research projects funded by the European Union, the Italian Ministry of Research, the National Research Council of Italy, the Italian Space Agency, and industries. Original results have been published in more than 350 papers in international journals, proceedings of international conferences, and book chapters.,Dr. Piuri is a Distinguished Scientist of ACM and a Senior Member of INNS. He has been an Associate Editor of IEEE Transactions on Neural Networks and IEEE Transactions on Instrumentation and Measurement. Dr. Piuri is an ACM Fellow.

Fabio Scotti (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003.,He was an Assistant Professor with the Department of Information Technologies, Università degli Studi di Milano, Milan, from 2002 to 2015, where he was an Associate Professor with the Department of Computer Science from 2015 to 2020. He has been a Full Professor with the Università degli Studi di Milano since 2020. Original results have been published in over 150 papers in international journals, proceedings of international conferences, books, book chapters, and patents. His current research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, 3-D reconstruction, industrial applications, intelligent measurement systems, and high-level system design.,Prof. Scotti is an Associate Editor of the IEEE Transactions on Human–Machine Systems and the IEEE Open Journal of Signal Processing. He is serving as a Book Editor (Area Editor, section Less-Constrained Biometrics) of the Encyclopedia of Cryptography, Security, and Privacy (3rd Edition, Springer). He has been an Associate Editor of the IEEE Transactions on Information Forensics and Security, Soft Computing (Springer) and a Guest Coeditor of the IEEE Transactions on Instrumentation and Measurement.