Computers and Electrical Engineering

ConsMOT: A Cross-Scale and Re-parameterized Multi-Object Tracker for Overhead Crane Safety Monitoring

--Manuscript Draft--

Manuscript Number:	COMPELECENG-D-25-08607
Article Type:	Research Paper
Section/Category:	Applications of Artificial Intelligence
Keywords:	Smart Construction Site; Multi-Object Tracking; Re-parameterized; Multi-Scale Fusion
Corresponding Author:	Ruixuan Zhang Tianjin University of Science and Technology CHINA
First Author:	Chonghua Zhou
Order of Authors:	Chonghua Zhou
	Ruixuan Zhang
	ningzhi chen
	Haitao He
	Yixi Fu
	Kevin Liu
Abstract:	Ensuring a safe separation distance between workers and heavy machinery is a critical requirement in smart construction sites. Multi-object tracking (MOT) serves as the core perception module for such safety-critical applications, yet existing trackers often struggle with inaccurate detection of small objects in complex overhead crane scenarios. To address these challenges, we propose ConsMOT, a MOT framework tailored for construction-site safety monitoring. A learnable cross-scale aggregation mechanism called Gated Feature Fusion (GFF) is proposed. GFF enables the network to adaptively weight multi-level features, improving robustness under large viewpoint variations and dense human–machine interactions. To compensate for the additional computational cost introduced by the GFF, we further propose RepBlock, a re-parameterizable convolutional structure that algebraically fuses a convolution layer and its subsequent batch-normalization layer into an equivalent form during inference. This transformation reduces latency and memory access without modifying the training-time architecture, making ConsMOT more suitable for edge-side deployment in safety monitoring systems. Extensive experiments on a public tower-crane dataset (CraneView-P) and a private dataset (CraneView-N) show that Gated Feature Fusion produces consistent accuracy gains relative to the baseline, with relative improvements of 0.6% on CraneView-P and 7.1% on CraneView-N. And the ConsMOT attains increase in inference throughput compared to the baseline when re-parameterize the RepBlock in inference-time, while maintaining the accuracy benefits of MOT.

To the Editorial Office, Computers & Electrical Engineering:

We would like to submit our manuscript entitled "ConsMOT: A Cross-Scale and Re-parameterized Multi-Object Tracker for Overhead Crane Safety Monitoring" for consideration in **Computers & Electrical Engineering**. This work proposes a multi-object tracking method tailored for smart construction sites, addressing the critical challenge of maintaining safe separation distances between workers and heavy machinery.

The manuscript proposes two technical innovations. First, Gated Feature Fusion (GFF) enhances cross-scale feature representation to improve the detection and tracking of small objects under high-altitude, top-down views. Second, RepBlock, a re-parameterizable convolutional method, accelerates inference by algebraically fusing convolution and batch-normalization layers without impacting training-time performance. Together, these modules enable real-time, accurate tracking of workers and machinery, a key requirement for edge-deployed safety systems.

ConsMOT was evaluated on both a public tower-crane dataset (CraneView-P) and a private dataset (CraneView-N). The experiments demonstrate consistent accuracy improvements over baseline methods, along with a substantial increase in inference throughput. These results highlight the practical applicability of our approach in real-world smart construction environments.

We believe this manuscript is well-suited for **Computers & Electrical Engineering** due to its focus on efficient computer vision methods for engineering applications, its combination of methodological innovation and real-world deployment, and its relevance to readers interested in intelligent construction site monitoring, real-time systems, and multi-object tracking. We confirm that this manuscript has not been published elsewhere and is not under consideration by any other journal.

Thank you for your consideration. We look forward to your response.

Sincerely,

Chonghua Zhou

EEIS Department, University of Science and Technology of China

Highlights

ConsMOT: A Cross-Scale and Re-parameterized Multi-Object Tracker for Overhead Crane Safety Monitoring

Chonghua Zhou, Ruixuan Zhang, Ningzhi Chen, Haitao He, Yixi Fu, Kevin Liu

- Proposed ConsMOT, a multi-object tracking framework for smart construction sites with overhead crane scenarios.
- Proposed Gated Feature Fusion to improve detection and tracking of small, occluded objects in dense worker–good interactions.
- introduced RepBlock, a re-parameterizable convolution block that accelerates inference without compromising accuracy.
- Achieved consistent improvements in multi-object tracking accuracy on public and private crane-view datasets, with relative gains of 0.6% and 7.1%.

ConsMOT: A Cross-Scale and Re-parameterized Multi-Object Tracker for Overhead Crane Safety Monitoring

Chonghua Zhou[a,c], Ruixuan Zhang[b,*], Ningzhi Chen[b], Haitao He[b], Yixi Fu[b], Kevin Liu[d]

[a] *EEIS Department, University of Science and Technology of China, China*
[b] *College of Electronic Information and Automation, Tianjin University of Science and Technology, China*
[c] *CCCC Mechanical & Electrical Engineering Co. Ltd, China Communications Construction Company, China*
[d] *Western University Department of Computer Science, Canada*

Abstract

Ensuring a safe separation distance between workers and heavy machinery is a critical requirement in smart construction sites. Multi-object tracking (MOT) serves as the core perception module for such safety-critical applications, yet existing trackers often struggle with inaccurate detection of small objects in complex overhead crane scenarios. To address these challenges, we propose ConsMOT, a MOT framework tailored for construction-site safety monitoring. A learnable cross-scale aggregation mechanism called Gated Feature Fusion (GFF) is proposed. GFF enables the network to adaptively weight multi-level features, improving robustness under large viewpoint variations and dense human–machine interactions. To compensate for the additional computational cost introduced by the GFF, we further propose RepBlock, a re-parameterizable convolutional structure that algebraically fuses a convolution layer and its subsequent batch-normalization layer into an equivalent form during inference. This transformation reduces latency and mem-

*Corresponding author.

Email addresses: zhouchonghua@mail.ustc.edu.cn (Chonghua Zhou), zhangrx@tust.edu.cn (Ruixuan Zhang), chenningzhi@mail.tust.edu.cn (Ningzhi Chen), hehaitao@mail.tust.edu.cn (Haitao He), fuyixi@mail.tust.edu.cn (Yixi Fu), kliu469@uwo.ca (Kevin Liu)

ory access without modifying the training-time architecture, making ConsMOT more suitable for edge-side deployment in safety monitoring systems. Extensive experiments on a public tower-crane dataset (CraneView-P) and a private dataset (CraneView-N) show that Gated Feature Fusion produces consistent accuracy gains relative to the baseline, with relative improvements of 0.6% on CraneView-P and 7.1% on CraneView-N. And the ConsMOT attains increase in inference throughput compared to the baseline when re-parameterize the RepBlock in inference-time, while maintaining the accuracy benefits of MOT.

Keywords:
Smart Construction Site, Multi-Object Tracking, Re-parameterized, Multi-Scale Fusion

1. Introduction

In construction site environments, a large number of construction workers and machines operate within shared workspaces, leading to an increased risk of human–machine conflicts and accidents such as collisions and falls from height. The major hazards in the construction industry include falls from height, accidents that were struck-by vehicles, and accidents that occurred between or between vehicles [1]. Some researchers have conducted extensive studies on occupational fatalities in the construction industry, finding that accidents with struck persons are the most common cause of death, followed by caught-in or between incidents, which account for one out of every nine fatal cases [2]. Therefore, achieving efficient perception, dynamic analysis, and risk warning in collaborative scenarios between workers and machines within smart construction site environments has become a key research focus in the field of construction safety management.

In recent years, artificial intelligence technologies, particularly computer vision and deep learning, have provided strong technological support for safety management in smart construction sites. Their applications have expanded from single-dimensional safety monitoring to multiple aspects of auxiliary management tasks, achieving remarkable results [3]. The integration of artificial intelligence and monitoring technologies facilitates precise detection and early warning of potential hazards, thereby substantially enhancing the real-time responsiveness and reliability of construction safety management, while supporting proactive accident prevention and risk mitigation [4]. These systems enable dynamic identification and early warning of

potential risks on construction sites, thereby facilitating the optimization of the working environment, reducing accident rates, and enhancing the overall efficiency of safety management [5]. By deeply mining the latent correlations between worker behavior records and contextual data, some systems [6] facilitate the development of personalized risk control mechanisms, enabling precise identification and proactive prevention of unsafe behaviors. Building on this foundation, an AI-driven occupational safety monitoring framework has been established. By integrating data analytics, machine learning, and Internet of Things technologies with real-time operational data collected from wearable sensors and hazard recognition via computer vision, a new framework [7] enables real-time detection of construction site hazards and early risk prediction. This system effectively prevents safety incidents and equipment failures, identifies unauthorized worker entry into hazardous areas, and comprehensively enhances the level of occupational safety assurance in smart construction sites.

Among the various tasks involved in human–machine safety management, real-time detection of human–machine collision risk represents both the core challenge and a critical research focus. This task is typically modeled as a multi-object detection and tracking problem in video sequences, aiming to localize workers and machinery in real time and determine whether their spatial relationships indicate potential collision risks. From a data perspective, research on worker–machine collision risk detection primarily relies on multi-source video and image data. Different studies have advanced detection accuracy by constructing diverse datasets and integrating specialized technical approaches, thereby providing a solid data foundation for human–machine collision risk analysis. The differentiated datasets mainly include surveillance datasets collected from fixed cameras [8], detection-oriented datasets based on high-resolution image data and deep learning algorithms [9], and multi-dimensional video datasets integrating pose estimation and equipment volume modeling [10]. Building on this foundation, subsequent research has further enriched the dimensionality of datasets, leading to the development of risk monitoring systems based on machine learning techniques. These systems enable automatic identification of human–machine proximity states and provide graded warnings, thereby enhancing the intelligence and effectiveness of safety monitoring on construction sites [11].

Despite significant progress in recent years, existing multi-object tracking (MOT) methods still face several fundamental limitations when applied to overhead crane scenarios. One of these limitations is inaccurate detection

of small objects. Specifically, overhead crane images typically capture workers and machinery from a high-altitude, top-down perspective, resulting in small object sizes, low-resolution human silhouettes, and visually ambiguous boundaries. Many joint detection–tracking frameworks rely on standard multi-scale backbone representations, which tend to underperform when detecting such tiny instances. Inaccurate detection directly propagates downstream, causing fragmented trajectories and elevated ID-switch counts. This challenge underscores the need for MOT frameworks that can robustly detect small objects under overhead viewpoints and maintain high inference efficiency to satisfy real-time deployment demands. To this end, we propose ConsMOT, a multi-object tracking network designed for overhead crane monitoring in smart construction sites. The key contributions of this work are summarized as follows:

- We propose Gated Feature Fusion, a learnable cross-scale aggregation module that enhances representational flexibility and improves detection and tracking performance in dense, dynamic construction scenes.
- We introduce a lightweight inference-time re-parameterization method named RepBlock that merges convolution and batch normalization into a single convolution, offering speedup without affecting training-time behavior or accuracy.
- We validate the effectiveness of ConsMOT on both public and private crane-view datasets, demonstrating consistent accuracy improvements and substantial acceleration in inference throughput.

2. Related Work

2.1. The Current State of Research on Construction Site Safety Management

Research on safety management in smart construction sites has progressed from traditional methods relying on regulatory compliance and manual inspections to advanced methodologies leveraging automated and intelligent analysis through diverse information technologies. This section reviews the latest developments in construction site safety management from three perspectives: sensor-based technologies, video surveillance systems, and intelligent hazard identification driven by vision-language models. Sensor-based

intelligent safety methods [12] were proposed by Ali Rashidi et al. to safeguard the well-being of on-site workers. Real-time monitoring tools leveraging sensor data can proactively identify potential construction hazards, thereby significantly enhancing overall construction site safety. Image recognition–based monitoring methods [13] were developed by Li et al. to enable intelligent safety management on construction sites by monitoring worker presence, personal protective equipment compliance, and potential safety hazards through camera systems. To address the limitations of traditional methods in hazard identification under specific contexts, a construction hazard identification framework based on vision-language models (VLMs) has been proposed [14]. This framework translates safety regulations into contextual queries, enabling VLMs to process visual information and generate compliance-based hazard assessments, thereby reinforcing proactive safety management.

2.2. The Current State of Research on Human–Machine Safety Detection

Mobile viewpoints rely on data acquisition devices with dynamically adjustable positions and angles, enabling comprehensive monitoring of complex, multi-area construction site operations. Research in this area has focused on dataset construction, algorithm adaptation and optimization, and contextualized applications. For instance, a UAV-based collaborative perception method [15] was proposed by Jin et al., accompanied by an experimental dataset comprising approximately 30,000 multi-angle video frames. This method utilizes aerial image-based object detection algorithms to automatically identify construction personnel and equipment, facilitating real-time monitoring of unsafe worker behaviors. Approximately 50,000 aerial image samples were collected by Sourav Kumar et al. using unmanned aerial vehicles to conduct analyses on safety helmet (PPE) wearing status recognition and hazardous area localization accuracy [16]. The study verified the technical feasibility of UAVs in construction safety detection and demonstrated the monitoring advantages of the top-down aerial perspective. The adaptability between algorithms and data was further improved by Xin et al. [17], who integrated the YOLOv8 model with UAV-acquired aerial imagery. This integrated method was designed for specialized detection of personnel and protective equipment in aerial images, effectively enhancing the small-object recognition performance of the traditional YOLO baseline model under top-down viewing perspectives. In studies on the application of crane-mounted cameras, Wang et al. [18] designed an overhead safety monitoring system

mounted on cranes for tower lifting and hoisting operations. This system was developed to explore automated warning mechanisms for collision prevention and personnel safety protection during lifting activities. Construction site image data were collected by L. Joachim et al. using crane-mounted cameras to construct digital elevation models (DEMs) for supporting construction path planning. In addition, multiple SLAM algorithms were quantitatively evaluated in terms of real-time mapping accuracy and operational performance within the crane camera system[19].

Another research branch of mobile viewpoints focuses on helmet-mounted cameras, which enable near-field, first-person data acquisition through worker-worn devices. This method addresses key technical challenges such as close-range hazardous behavior recognition and personal protective equipment (PPE) compliance detection. A multi-task detection framework combining the YOLO model and a spatio-temporal graph convolutional network [20] was proposed by Liu et al., utilizing approximately 3,000 egocentric images as training samples to achieve precise identification of unsafe behaviors, including workers approaching machinery or failing to wear protective equipment. A multi-site mobile vision acquisition system based on GoPro devices was developed by Wang et al. to capture over 50,000 frames of real construction-site videos [21]. The study systematically analyzed data collection standards, annotation protocols, and engineering deployment costs, providing practical insights for the industrial implementation of helmet-mounted mobile vision systems.

Research on fixed-view monitoring primarily relies on pole-mounted cameras, focusing initially on high-pole configurations that offer broad coverage and strong temporal continuity across medium- to large-scale construction zones. To address the high false-alarm rates often observed during long-term monitoring, Syed Farhan Alam Zaidi et al. [22] were the first to integrate temporal sequence fusion analysis into conventional object detection algorithms. This method effectively mitigated false detections caused by transient occlusions or illumination variations, significantly enhancing model stability in long-duration video surveillance. A near real-time 3D reconstruction method [23] was proposed by Sun et al., leveraging the wide coverage capability of high-pole cameras to dynamically capture and reflect construction site morphology and progress. This method provides crucial technical support for spatially aware safety monitoring and data visualization in construction environments.

In fixed-view monitoring, low-pole cameras are employed to observe lo-

calized work areas, offering high target resolution and rapid response capabilities. Research reported in [24] indicates that integrating low-pole cameras with multiple complementary technologies can further improve data processing efficiency and real-time performance, rendering this method particularly well-suited for localized monitoring in complex construction site environments. To address high-risk scenarios such as falls from height and flying debris, a construction site falling object and debris detection system [25] was designed by Zheng et al. This system leverages the high-detail capture capability of low-pole cameras and employs a synergy of camera imaging and visual recognition algorithms to achieve precise localization of vertically moving objects. As a result, the system effectively enhances early warning capabilities for fall and debris incidents, thereby strengthening personnel safety within the work area.

In summary, mobile vision systems are characterized by diverse data types, moderate data volumes with high-quality annotations, and the incorporation of spatial awareness, multi-task loss functions, and temporal modeling into baseline models such as YOLOv5/v8 and Faster R-CNN, facilitating real-time risk detection in complex and dynamic environments. Evidently, both mobile and fixed viewpoints have achieved substantial advancements in human–machine safety detection. Nonetheless, as highlighted by Tian et al.[26], fixed-view research continues to present critical avenues for further exploration, including the optimization of camera sensor layouts, enhanced adaptability to dynamic changes in construction sites, and the deeper integration of multi-source data.

3. Method

The construction environment introduces unique perception challenges including workers appear as low-resolution objects, materials and machinery exhibit diverse spatial scales, and the top-down viewpoint produces large intra-scene variance in object geometry. To address these issues, we proposes a multi-object tracking method for construction site named ConsMOT. which integrates adaptive feature fusion and re-parameterization modules to improve the MOT accuracy of workers and goods. The overall pipeline follows the FairMOT but introduces two key modifications. The architecture of our ConsMOT is illustrated in Fig.1.

Specifically, Section 3.1 introduces a cross-layer feature fusion module that integrates multi-scale information through a learnable gating strategy.

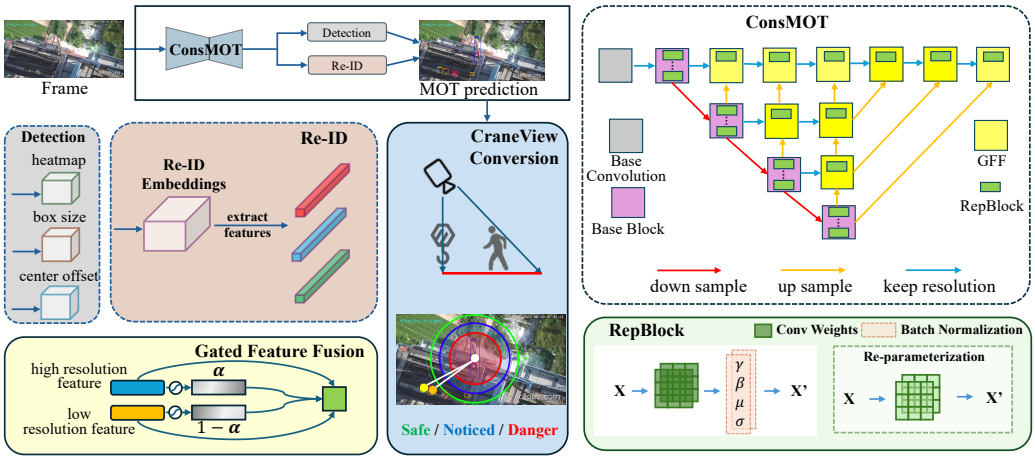


Figure 1: Overall framework of the proposed ConsMOT

Section 3.2 introduces the structural re-parameterization method of the backbone network, which folds convolutional layer and batch normoalization layer into a single equivalent kernel to enable efficient inference. Section 3.3 describes the joint optimization of loss functions, which adaptively balances the detection and re-identification objectives.

Overall, the proposed method achieves a favorable trade-off between accuracy, real-time performance, and resource efficiency, providing a practical solution for multi-object tracking in construction site environments.

3.1. Gated Feature Fusion

In overhead crane monitoring, the spatial footprint of workers and small equipment is drastically reduced due to extreme camera height. As a result, the effectiveness of small-object detection strongly depends on the quality of high-resolution feature maps. FairMOT employs an iterative deep aggregation module to aggregate features across multiple scales. However, the iterative deep aggregation module uses a non-adaptive cross-scale summation operation, that is $F = \mathbf{x}_1 + \mathbf{x}_2$.

where $\mathbf{x}_1$ and $\mathbf{x}_2 \in (0, 1)^{H \times W \times C}$ are the high-resolution feature map and the upsampled low-resolution counterpart. While efficient, this operation implicitly assumes that all spatial scales contribute equally across all scenarios. In construction environments—where workers, materials, and machinery occupy extremely different spatial extents—such fixed fusion can introduce

noise or oversmooth subtle patterns needed for precise small-object localization.

To overcome these limitations, we introduce Gated Feature Fusion (GFF), a learnable method that adaptively controls the contribution of each feature scale, enabling the model to prioritize fine-grained spatial cues when necessary. Given two input feature maps $\mathbf{x}_1, \mathbf{x}_2$, GFF predicts a gating factor $\boldsymbol{\alpha}$ through a lightweight transformation consisting of simple convolution and sigmoid function. Formally,

$$\begin{aligned} \boldsymbol{\alpha} &= GFF(\mathbf{x}_1, \mathbf{x}_2) && (1) \\ &= \sigma_2(f_2(\sigma_1(f_1(\mathbf{x}_1, \mathbf{x}_2, \theta_1)), \theta_2)), && (2) \end{aligned}$$

where σ_1, σ_2 denote Relu and Sigmoid functions. The θ_1, θ_2 denote the convolutional weights. The gating factor forms a soft spatially varying interpolation mask that modulates the fusion process. The final fused representation is computed as

$$F = \boldsymbol{\alpha} \odot \mathbf{x}_1 + (1 - \boldsymbol{\alpha}) \odot \mathbf{x}_2, \tag{3}$$

where $\odot$ denotes element-wise multiplication.

This formulation allows the fusion weights to adapt to both spatial context and channel semantics. In regions corresponding to workers, where object sizes are extremely small and sensitive to fine-scale distortions, the gating tensor can naturally favor $\mathbf{x}_1$, preserving high-resolution details crucial for accurate detection and tracking. Conversely, in areas occupied by large machinery or background structures, deeper and coarser semantic cues in $\mathbf{x}_2$ can be emphasized to stabilize the representation. The soft and continuous nature of $\boldsymbol{\alpha}$ also provides a smooth mechanism for balancing fine- and coarse-scale information in areas where workers and machinery interact closely, a frequent and challenging situation in real construction environments.

In summary, the Gated Feature Fusion module replaces the heuristic and rigid feature summation of FairMOT with an adaptive and learnable alternative. By tailoring the fusion weights to the spatial and semantic characteristics of crane-view imagery, GFF yields more discriminative multi-scale features, improves the detectability of small human objects, and enhances the downstream association performance of the MOT pipeline.

3.2. RepBlock

During inference, the computational graph of multi-object tracking method is often dominated by convolution and normalization layers, whose repeated execution introduces nontrivial latency and memory access overhead. Although batch normalization (BN) stabilizes optimization and accelerates convergence during training, its presence in the inference pipeline is unnecessary, as its affine transformation can be analytically absorbed into the preceding convolution. To reduce redundant computation while maintaining architectural compatibility with common training setups, we adopt a lightweight re-parameterization strategy that transforms each convolution and BN pair into an equivalent single convolution kernel before deployment.

Let a convolutional layer be parameterized by weights $W \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k \times k}$ and bias b, followed by a BN layer with parameters $(\gamma, \beta, \mu, \sigma^2)$. During inference, the BN transformation applied to an input activation x is

$$\text{BN}(y) = \gamma \frac{y - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta, \tag{4}$$

where $y = W \times x + b$. Since both convolution and BN are affine operators, they can be merged into a single equivalent convolution with merged kernel W' and bias b'. Formally,

$$W' = \frac{\gamma}{\sqrt{\sigma^2 + \epsilon}} W, \quad b' = \beta + \frac{\gamma}{\sqrt{\sigma^2 + \epsilon}} (b - \mu). \tag{5}$$

After this transformation, the inference-time computation reduces to $y' = W' \times x + b'$, which produces outputs identical to the original Conv–BN sequence, thereby guaranteeing representational equivalence without altering the forward behavior of the model. This exact correspondence ensures that training dynamics, regularization effects, and representational capacity remain unchanged, while the deployed model benefits from a strictly simplified computational graph.

Compared with more complex re-parameterization method integrates multiple convolutional branches into a unified structure [27], our approach embraces minimalism. No auxiliary branches or structural modifications are introduced during training; the architecture seen by the optimizer is identical to the baseline model except for the post-training conversion step. This design choice is advantageous in settings where architectural stability is needed—for example, when ablation studies must remain comparable or when the backbone must remain compatible with existing MOT tracking heads and asso-

ciation modules. The simplicity of the transformation also avoids additional regularization terms or training heuristics that might complicate hyperparameter tuning.

The practical benefits become notable in resource-constrained deployments. Removing BN layers decreases the number of memory reads and writes, as BN involves fetching mean and variance statistics as well as performing element-wise arithmetic. The kernel fusion reduces the number of launched CUDA kernels during inference, further improving utilization on edge devices commonly used in smart construction-site monitoring systems. Although the gain for a single layer may appear modest, applying the conversion across the entire network yields measurable improvements in end-to-end throughput. In our experiments, the re-parameterized network achieves a substantial increase in inference speed—while preserving all accuracy benefits attained during training—highlighting that even modest architectural simplifications can play a non-negligible role in practical MOT deployments.

3.3. Loss Functions

ConsMOT's dual-branch loss integrates anchor-free detection precision and identity-aware embedding discrimination into a unified optimization framework. The joint objective enhances tracking robustness in crowded and occluded scenes, ensuring accurate detection and reliable identity consistency for multi-object tracking.

3.3.1. Detection Loss

The detection branch follows the anchor-free design based on CenterNet, employing three parallel prediction heads to estimate object centers, bounding box sizes, and center offsets. The detection loss is composed of the heatmap loss $\mathcal{L}_{hm}$ and the bounding box regression loss $\mathcal{L}_{reg}$, expressed as:

$$\mathcal{L}_{det} = \mathcal{L}_{hm} + \mathcal{L}_{reg}. \tag{6}$$

The heatmap head predicts the probability of each pixel being an object center. A Gaussian kernel centered on each ground-truth object center is used to generate target heatmaps. To handle the class imbalance between foreground and background pixels, the focal loss [28] is adopted:

$$L_{hm} = -\frac{1}{N}\sum_{x,y}\begin{cases}(1-\hat{M}_{xy})^{\alpha}\log(\hat{M}_{xy}), & M_{xy}=1\\ (1-M_{xy})^{\beta}(\hat{M}_{xy})^{\alpha}\log(1-\hat{M}_{xy}), & otherwise,\end{cases} \tag{7}$$

where M and $\hat{M}$ denote the ground-truth and predicted heatmaps, respectively, and α, β are balancing parameters.

The box offset and size heads regress the bounding box position and scale around each detected center. To mitigate quantization errors from down-sampling, an L_1 loss is applied to both size and offset predictions:

$$L_{reg} = \sum_{k=1}^{N} |o_k - \hat{o}_k|_1 + \lambda_s |s_k - \hat{s}_k|_1, \tag{8}$$

where $\hat{s}_i$ are the ground-truth offset and size vectors, $\hat{o}_i$ and $\hat{s}_i$ are their predicted counterparts, and λ_s is a scaling coefficient. This detection loss effectively ensures precise localization and size estimation of objects in complex tracking scenarios.

3.3.2. Re-Identification Loss

The Re-ID branch learns discriminative appearance embeddings that differentiate object identities. Each object instance is regarded as a unique class in the training set. Given an object feature vector $E_{x,y}$ extracted from the predicted object center, it is mapped through a fully connected layer followed by a softmax classifier to produce identity probabilities. The cross-entropy loss is then computed as:

$$L_{re} = -\sum_{k=1}^{N} \sum_{c=1}^{C} G_k(c) \log(P_k(c)), \tag{9}$$

where $\mathcal{Y}(k)$ is the one-hot ground-truth label for identity k, $p(k)$ is the predicted probability, and K is the number of identities in the training dataset. This loss enforces feature consistency across frames for the same identity while maximizing inter-class separability.

3.3.3. Total Loss

To achieve balanced training of the detection and Re-ID tasks, FairMOT adopts an uncertainty-based weighting strategy to automatically adjust the relative contribution of each task:

$$L = \frac{1}{2}\left(\frac{1}{e^{\alpha}} L_{det} + \alpha\right) + \frac{1}{2}\left(\frac{1}{e^{\beta}} L_{re} + \beta\right), \tag{10}$$

where α and β are learnable parameters. This formulation effectively mitigates task competition and ensures fair optimization between spatial detection and identity learning.

3.4. Distance Calibration

Ensuring reliable human–machine safety assessment requires measuring the actual physical distance between workers and the suspended material rather than relying solely on pixel-level separations. Three spatial zones are defined around the hoisted goods to characterize different risk levels. A working zone D3 with a radius of 7m. A noticed zone D2 with a radius of 5m, and a danger zone D1 with a radius of 2m. Accurate zone assignment therefore depends on converting the image-space distance to the corresponding real-world distance. However, the scale in the CraneView dataset is not fixed. The crane height varies considerably across frames, causing the pixel-to-meter ratio to change dynamically. A direct use of static calibration would inevitably introduce systematic errors and undermine the reliability of risk estimation.

To address this challenge, we estimate a frame-wise scale factor by exploiting the statistical regularity of worker detection boxes. Under the top-down view of tower-crane cameras, the size of a worker's detection box is weakly perspective-dependent and consistently approximates worker shoulder width when represented by the larger side of the box. Formally, let the detected worker bounding boxes in the current frame be $\{B_i\}_{i=1}^{N}$, and let $s_i = \max(h_i, w_i)$ denote the larger dimension of the i-th box, where h_i and w_i represent its height and width in pixels. The average apparent shoulder width is then computed as

$$\bar{s} = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{11}$$

Assuming an average worker height of 1.7m and a canonical shoulder–height ratio of approximately 0.26, the expected real shoulder width is estimated as $\ell_{\text{shoulder}} = 1.7 \times 0.26 \approx 0.442\,\text{m}$. This anthropometric prior provides a stable physical reference that is insensitive to clothing variations and body posture. The dynamic pixel-to-meter scale factor for the current frame is therefore $\beta = \frac{\ell_{\text{shoulder}}}{\bar{s}}$, which converts any pixel distance d_{pix} to its real-world counterpart $d_{\text{real}} = \beta\, d_{\text{pix}}$.

By applying this conversion, the distance between each worker and the hoisted goods can be evaluated in meters and compared against the safety thresholds of D1, D2, and D3. This adaptive scaling strategy ensures that the safety-zone assignment remains accurate despite variations in crane elevation, camera pitch, and worker position within the field of view. Consequently, the final human–machine interaction assessment reflects the true spatial configuration on the construction site, enabling reliable and interpretable collision-risk analysis.

4. Experiment

4.1. Dataset

The image data used for model training and testing in this study were collected from two datasets. The first public dataset as shown in Fig.2(a), CraneView-P (public), consists of images captured at a construction site in Dalian, China, using a vertically downward-facing camera mounted on the top of a tower crane [29]. Six crane operation scenes were selected, resulting in 13,155 annotated images.

The second private dataset as shown in Fig.2(b), named CraneView-N (new), follows the same data acquisition strategy as CraneView-P, employing a top-mounted downward-looking camera on the tower crane. It contains 4,000 annotated images covering four core crane operation scenes. Building upon the scene specificity of the CraneView-P (public) dataset, CraneView-N achieves significant improvements in multiple aspects. First, the annotation categories are more comprehensive, not only covering the core objects present in the original dataset but also introducing additional key elements in construction scenes, such as a greater number of workers and various types of lifted materials. Second, the annotation granularity is refined, providing more accurate bounding-box localization and clearer category differentiation, which enables the dataset to better capture the complex interactions among objects during crane operations. Furthermore, the diversity of scenes is enhanced by incorporating data collected under varying illumination conditions and construction stages, thereby improving the dataset's generalization ability and adaptability to real-world scenarios.

To ensure temporal consistency between the training and testing phases, the collected image data were divided into training and testing subsets in chronological order, with the testing data positioned after the training data.

(a) Images from CraneView-P (b) Images from CraneView-N

Figure 2: Samples from CraneView-P (public) and CraneView-N (new)

Considering the relatively limited dataset size and following the related research setting[29], the division ratio was set to 9:1, where the last portion of each video sequence was used for testing.

4.2. Experiment Settings

The experiments were conducted under the following hardware and software environments: the system was equipped with an Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz and an NVIDIA Quadro RTX 4000 GPU (8 GB memory). The software environment consisted of Ubuntu 20.04.6 LTS as the operating system and PyTorch 1.13.1 with CUDA 11.7 as the deep learning framework.

The Adam optimizer was used for parameter optimization, and the model was trained for 60 epochs. The batch size was set to 4, and the initial learning rate was set to 7×10^{-5}. The learning rate was reduced by a factor of 0.1 after every 10 epochs, in order to ensure more stable convergence during the later stages of training.

Furthermore, a pretrained model on the COCO dataset [30] was adopted to initialize the network parameters, enabling transfer learning to accelerate convergence and enhance feature representation. This strategy effectively

improves detection and tracking performance in crane operation scenarios with limited domain-specific data.

4.3. Metircs

In multi-object tracking (MOT) tasks, the evaluation of model performance generally requires a comprehensive consideration of both detection accuracy and tracking stability. In this study, several widely used MOT metrics defined by the MOTChallenge benchmark are adopted, including Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), the number of Identity Switches (IDSW), False Positives (FP), False Negatives (FN), and the proportion of Mostly Tracked (MT) objects. These metrics collectively reflect the performance of the model in detection and data association, providing a multifaceted assessment of its tracking capability in real-world scenarios.

MOTA measures the overall tracking accuracy by jointly considering false positives, false negatives, and identity switches. MOTA is defined as:

$$MOTA = 1 - \frac{\mathrm{FN} + \mathrm{FP} + \mathrm{IDSW}}{\mathrm{GT}}, \tag{12}$$

where FN denotes the number of false negatives, FP represents false positives, IDSW indicates the number of identity switches, and GT is the total number of ground-truth objects. A higher MOTA value implies better overall tracking accuracy and is often regarded as the most comprehensive metric for evaluating multi-object tracking performance.

MOTP evaluates the spatial overlap between predicted bounding boxes and ground-truth annotations, reflecting the model's localization precision. MOTP is computed as:

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_{t} c_t}, \tag{13}$$

where $d_{i,t}$ represents the distance (commonly measured by the Intersection over Union (IoU) or the center-point distance) between the predicted and ground-truth boxes of the i-th object in frame t, and c_t denotes the number of matches in frame t. A higher MOTP value indicates more accurate bounding-box localization and stronger spatial precision.

IDSW refers to the number of instances in which the same object is incorrectly assigned different identities during tracking. Such switches typically occur when objects are temporarily occluded or when multiple visually similar objects intersect. A lower IDSW count indicates that the model can

maintain consistent object identities, thereby achieving better temporal continuity.

FP are detections produced in the absence of a true object, whereas FN represent instances where true objects are missed by the detector. FP and FN respectively correspond to the precision and recall of the detection system. A high number of FPs results in false alarms, while excessive FNs lead to missed detections; both types of errors significantly degrade the MOTA score.

MT measures the long-term tracking capability of the algorithm, defined as the proportion of ground-truth trajectories successfully tracked for more than 80% of their lifespan. A higher MT value indicates that the system maintains more continuous object trajectories, effectively reducing track fragmentation and loss.

4.4. Experiments on CraneView-P

4.4.1. Comparison Experiments

Based on the comparison on the CraneView-P dataset in Tab. 1 and Tab. 2, it can be observed that the proposed ConsMOT achieves stable and practically meaningful improvements in overall multi-object tracking performance. In terms of global metrics, ConsMOT outperforms the baseline FairMOT in both MOTA and mAP, corresponding to an overall error rate reduction of 19%. This indicates that, while maintaining high detection accuracy, ConsMOT can further suppress various tracking-related errors, thereby improving the overall tracking quality on long video sequences.

From the perspective of error composition, the two models exhibit different trade-offs between false positives and false negatives. Compared with the baseline, ConsMOT slightly increases the number of false positives but substantially reduces the number of false negatives, with the latter decreasing by more than one third. Given that MOTA is more sensitive to missed detections, this strategy of accepting a small increase in false alarms leads to higher overall tracking performance. In construction monitoring scenarios, missed detections are typically more critical and less tolerable than false alarms, making this error distribution more valuable in real-world applications.

At the category level, workers are inherently more challenging due to frequent occlusions, large pose variations, and complex interactions with the surrounding environment. For the workers, ConsMOT yields a slight improvement in MOTA and effectively alleviates long-term target loss and trajectory interruption, with most of the gain attributed to the reduction in

missed detections. IDSW for the workers are completely eliminated, indicating more stable identity preservation and inter-frame association, which effectively mitigates trajectory fragmentation. For the goods, where objects exhibit relatively stable appearance and simple geometric structure, both methods maintain strong performance in terms of false positive control and identity association. Building upon this, ConsMOT further improves recall, leading to slight gains in both MOTA and mAP, and pushing the performance for this class close to saturation.

Regarding spatial localization accuracy, ConsMOT achieves slightly higher MOTP values than the baseline. According to the MOTChallenge definition, this can be interpreted as a mild relaxation in bounding-box alignment precision. However, combined with the aforementioned improvements in MOTA and the substantial reduction in missed detections, it can be inferred that the model intentionally trades strict bounding-box alignment for higher recall and more stable trajectories.

Table 1: Performance of FairMOT on the CraneView-P Dataset.

Class	MT	FP	FN	IDSW	MOTA (%)	MOTP	mAP (%)
Worker	9	26	51	2	95.20	0.215	97.80
Goods	6	0	12	0	99.10	0.138	99.08
Overall	15	26	63	2	96.90	0.181	98.44

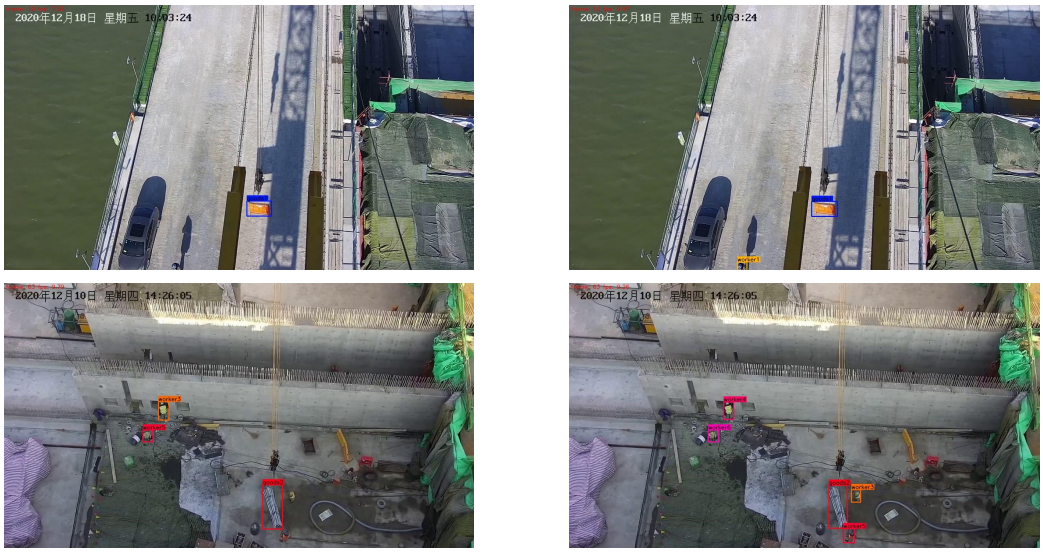
Table 2: Performance of ConsMOT on the CraneView-P Dataset.

Class	MT	FP	FN	IDSW	MOTA (%)	MOTP	mAP (%)
Worker	9	38	35	0	**95.60**	**0.229**	97.64
Goods	6	0	7	0	**99.50**	**0.164**	**99.47**
Overall	15	38	42	0	**97.50**	**0.196**	**98.55**

4.4.2. Visualization Analysis

The visualization results of the ConsMOT framework for worker-good detection and tracking, compared to FairMOT, on the public dataset CraneView-P are shown in Figure 3. The workers are outlined in yellow, orange, red, and magenta. The goods are outlined in blue and red. Each detected object is assigned a unique identity to enable cross-frame association.

To evaluate the models' capability in detecting heterogeneous objects on construction sites, we first compare their performance in scenes containing



(a) MOT results of baseline (b) MOT results of ConsMOT

Figure 3: Multi-Object Tracking and Detection results on CraneView-P.

both hoisted materials and nearby workers. In the upper figure of Fig.3(a), FairMOT successfully identifies the suspended good but entirely fails to detect the workers present in the scene, revealing its difficulty in maintaining sensitivity across different object categories. In contrast, the upper figure of Fig.3(b) shows that ConsMOT is able to detect both the hoisted material and the worker (worker1) simultaneously. This more complete detection outcome indicates that ConsMOT provides improved multi-class responsiveness and mitigates the category imbalance issue that commonly arises in construction-site monitoring tasks.

A second comparison focuses on model robustness under complex, high-density environments. The lower figure of Fig.3(a) depicts a cluttered construction zone with four workers distributed at varying distances and levels of occlusion. FairMOT detects only two workers, demonstrating a noticeable degradation in detection reliability when object density increases and background interference becomes more pronounced. In contrast, the lower figure of Fig.3(a) illustrates that ConsMOT accurately identifies all four workers as well as the hoisted goods, while maintaining consistent identity assignments throughout the scene. These results collectively confirm that ConsMOT exhibits superior stability and robustness in large-scale multi-object environ-

ments, particularly where occlusion, motion complexity, and heterogeneous object types jointly challenge the detection process.

Overall, the experimental results demonstrate that the proposed ConsMOT framework significantly improves detection accuracy and tracking stability in complex construction environments, compared to FairMOT, enabling real-time and accurate monitoring of human-machine interactions.

4.4.3. Ablation Study

To comprehensively assess the effectiveness of the proposed ConsMOT, we conduct the ablation study to quantify how each module contributes to MOT and to evaluate whether these improvements can be achieved without sacrificing computational efficiency.

The ablation study results of MOT are shown in Tab. 3. The baseline FairMOT provides a relatively balanced performance across both goods and worker sequences, reflecting a well-rounded detection–tracking pipeline under the baseline design constraints. While RepBlocks are re-parameterized, the goods sequence exhibits noticeably higher tracking and detection performance, whereas improvements in the more challenging worker sequence remain modest and occasionally fluctuate. This divergence suggests that RepBlock primarily strengthens the model's capacity to handle objects with simpler and more consistent appearance patterns, while its effect on worker-related variations—such as articulated motion, occlusion, and fine-grained shape changes—is more limited. In contrast, GFF shifts the trade-off toward enriched feature representation: the worker sequence benefits more from its enhanced modeling of subtle appearance cues, especially in cases where pose variability and partial occlusion demand stronger discriminative features. However, the intensified representations may also introduce slight instability in association, indicating that representational strength alone does not guarantee optimal tracking robustness in dynamic scenes. When RepBlock and GFF are combined, their complementary properties become evident. The goods sequence approaches saturation performance while the worker sequence shows improvements over the baseline and RepBlock-only variants, indicating that the combination achieves a more desirable balance between robustness and fine-grained discriminability. Overall, Tab. 3 shows that RepBlock primarily benefits stable-appearance objects, GFF enhances detection in visually complex human-motion scenarios, and the joint configuration harmonizes these strengths to provide consistently strong results across both sequence types.

Table 3: Ablation Study of RepBlock and GFF on the CraneView-P Dataset.

GFF	RepBlock	Class	MOTA (%)	MAP (%)
	w/o. re-parameterizing	Goods	99.10	99.08
		Worker	95.20	97.80
		Overall	96.90	98.44
	w/. re-parameterizing	Goods	99.60	99.62
		Worker	94.70	97.30
		Overall	97.10	98.46
✓	w/o. re-parameterizing	Goods	98.90	98.93
		Worker	94.00	98.65
		Overall	96.50	**98.79**
✓	w/. re-parameterizing	Goods	99.50	99.47
		Worker	95.60	97.64
		Overall	**97.50**	98.55

The computational and efficiency analysis in Tab. 4 further contextualizes these accuracy gains from a deployment standpoint. Across all variants, parameter counts remain nearly unchanged, confirming that the added modules do not inflate model size. Instead, their impact is more visible in runtime behavior. RepBlock with re-parameterizing significantly accelerates inference while preserving model compactness, demonstrating the effectiveness of structural re-parameterization in improving throughput. The GFF, while beneficial for feature expressiveness, introduces more complex operations that reduce frame-rate performance despite having similar parameter size.

Table 4: Performance Comparison of FairMOT with Ablation of RepBlock and GFF Enhancements.

GFF	RepBlock	Total Parameters	FPS	Speedup
	w/o. re-parameterizing	20.35 M	15.05	x1
	w/. re-parameterizing	20.34 M	21.06	x1.40 (40%)
✓	w/o. re-parameterizing	20.62 M	14.01	x0.93 (-7%)
✓	w/. re-parameterizing	20.61 M	19.12	x1.36 (36%)

The combined configuration achieves an advantageous trade-off, which retains GFF's representational strength while recovering much of the lost inference speed through RepBlock's acceleration effects. This balance is par-

ticularly valuable for smart construction applications, where both accurate tracking and real-time responsiveness are essential.

4.4.4. Collision Detection

The visualization of consecutive frames for the human–machine collision warning process based on the CraneView-P dataset is presented in Fig.4. In the figure, the hoisted object goods1 is taken as the center, where the red circle D1 denotes the danger zone, the blue circle D2 denotes the warning zone, and the green circle D3 together with its exterior region represents the safe zone. The worker labeled worker1 is the monitored subject. The entire sequence spans from 0 s to 14 s and covers eight key time instants (a–h).

At the initial frame at 0 s (Fig.4(a)), worker1 is located between the blue D2 and the green D3 regions, remaining in a relatively safe position and only triggering low- or medium-level proximity warnings. At this moment, goods1 is suspended near the center of the working area and remains almost stationary.

In the subsequent frames at 2 s, 4 s, and 6 s (Fig.4(b)–(d)), worker1 gradually moves along the passage toward the hoisted object. The bounding box of worker1 shifts from the inner edge of D3 into D2 and further approaches the boundary of D1. By 6 s, worker1 has already reached the edge of the danger zone, and the actual distance between the worker and the object continues to decrease. Accordingly, the risk level issued by the system escalates from a warning state to a high-risk state, indicating a potential collision hazard.

In the following frames at 8 s, 10 s, and 12 s (Fig.4(e)–(g)), worker1 clearly enters the red danger zone D1 and at one point moves to a position directly beneath goods1, where the trajectory of the worker highly overlaps with the hazardous motion range of the object. During this period, the system maintains the highest alarm level and, according to predefined rules, suggests emergency braking or personnel evacuation. Meanwhile, goods1 consistently remains at the center of the three concentric regions, and the system is able to stably track the position of worker1 and dynamically adjust the warning level under the scenario where the object is nearly static and the worker continuously approaches.

In the final frame at 14 s (Fig.4(h)), worker1 still stays inside the red danger zone D1 and remains close to goods1. The system continues to output the highest alert level, indicating an extremely high collision risk. At this time, the trajectory of worker1 is still highly overlapped with the hazardous range of the object, and the warning system maintains a persistent high-risk

response.

Overall, this sequence from 0 s to 14 s clearly demonstrates the proposed distance-based human–machine collision warning mechanism: as worker1 moves from the safe zone (D3) into the warning zone (D2) and further intrudes into the danger zone (D1), the system progressively increases the risk level according to the degree of approach. When the worker remains in the danger zone for an extended period, the system sustains a high-risk warning, thereby validating the proposed method's high sensitivity to proximity risk and its stable dynamic warning capability in real-world lifting operations.

4.5. Experiments on CraneView-N

4.5.1. Comparison Experiments

The comparative analysis of the multi-object tracking performance between FairMOT and ConsMOT on the CraneView-N are shown in Tab.5 and Tab.6.

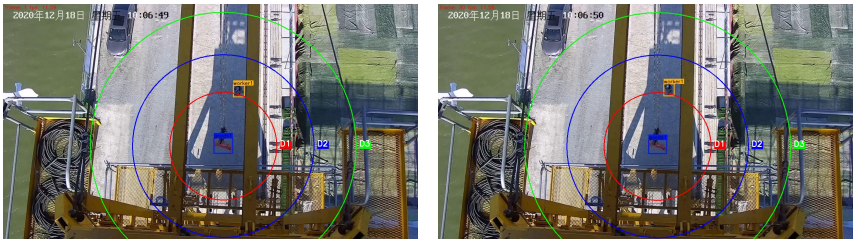
Table 5: Tracking and detection performance of FairMOT on the CraneView-N dataset.

Class	MT	FP	FN	IDSW	MOTA (%)	MOTP	mAP (%)
Worker	12	81	149	40	78.80	0.226	87.48
Goods	1	0	2	0	98.00	0.216	97.98
Overall	13	81	151	40	88.40	0.221	92.73

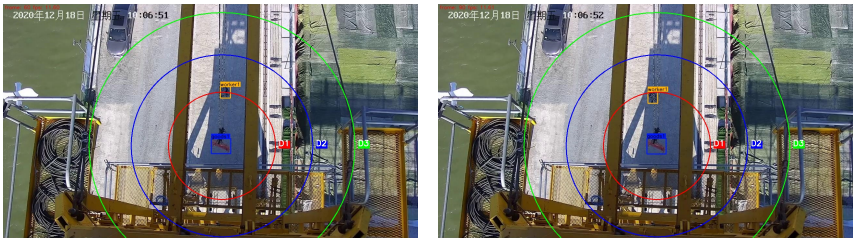
Table 6: Tracking and detection performance of ConsMOT on the CraneView-N dataset.

Class	MT	FP	FN	IDSW	MOTA (%)	MOTP	mAP (%)
Worker	15	28	55	38	**90.50**	**0.227**	**95.36**
Goods	1	0	1	0	**99.00**	**0.219**	**98.99**
Overall	16	28	56	38	**94.70**	**0.223**	**97.17**

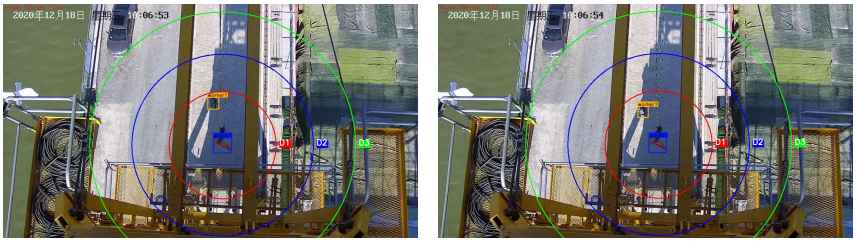
Based on the comparison between Tab. 5 and Tab. 6, it is evident that ConsMOT achieves more comprehensive and substantial performance improvements over FairMOT on the CraneView-N dataset, particularly in complex scenarios. From an overall perspective, ConsMOT delivers consistent gains across both tracking accuracy and detection precision, indicating stronger capabilities in object recognition quality, trajectory stability, and temporal consistency.



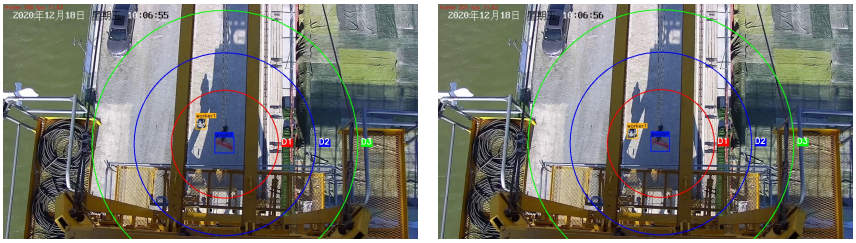
(a) 0s (b) 2s



(c) 4s (d) 6s



(e) 8s (f) 10s



(g) 12s (h) 14s

Figure 4: Collision Warning for CraneView-P Dataset.

The trends observed in error composition further demonstrate that ConsMOT is able to simultaneously reduce false positives and false negatives while also lowering the number of identity switches. This breaks the trade-off observed on the CraneView-P dataset, where reducing false negatives often came at the expense of increased false positives. The concurrent reduction of multiple error types reflects fundamental improvements in classification reliability, recall, and the stability of long-term tracking. Given the high sensitivity of MOTA to false negatives and identity switches, such multidimensional error reduction directly contributes to the overall accuracy improvement.

ConsMOT achieves particularly notable gains for the workers, including an increased number of successfully tracked trajectories, fewer interruptions during detection and tracking, and improved identity consistency. For the Goods class, both models already achieve high performance, yet ConsMOT further reduces key errors and pushes the results closer to saturation.

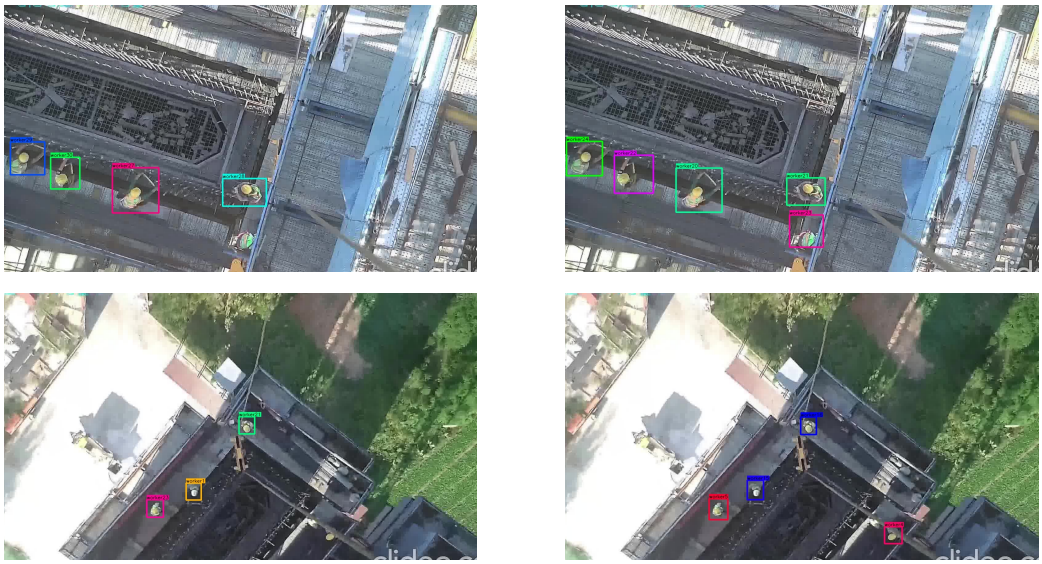
In terms of spatial localization accuracy, the difference between the two models is minimal. ConsMOT maintains nearly the same bounding-box alignment precision while simultaneously achieving higher recall and temporal consistency. This slight trade-off is particularly suitable for real-world construction monitoring, where severe occlusions and visually complex conditions are common. Prioritizing continuity and completeness of tracking over marginal changes in localization accuracy leads to a more resilient and practical tracking system.

4.5.2. Visualization

The visualization results of proposed ConsMOT on the CraneView-N dataset are shown in Figure 5. Different bounding box colors represent different tracking IDs to achieve temporal association across frames.

In regions where multiple workers gather, FairMOT is able to detect most objects and preserve their identities across frames, yet it occasionally exhibits identity switches or fragmented tracks when individuals appear in close proximity. ConsMOT reduces these errors by sustaining more stable identity assignments and maintaining coherent temporal associations, showing improved discriminability when objects move within dense clusters.

In areas containing steel structures and visually cluttered backgrounds, uneven illumination and frequent partial occlusions introduce further difficulty. FairMOT can still identify most workers, but its trajectories may drift or collapse under abrupt lighting changes or when a worker is partially obscured by structural elements. ConsMOT maintains greater resilience in



(a) MOT results of baseline (b) MOT results of ConsMOT

Figure 5: Multi-Object Tracking and Detection results on CraneView-N.

these situations, which separates multiple workers more reliably in heavily obstructed regions and preserves their identity continuity even when only limited appearance cues are available.

4.5.3. Ablation Study

The ablation experiment results on CraneView-N presented in Tab. 7 provide further insights into the performance trade-offs of different dataset. The baseline FairMOT demonstrates highly reliable performance in scenarios with structured visual patterns, particularly for objects with stable appearance such as goods, where the tracking and detection metrics are already close to saturation. In contrast, the worker remains significantly more challenging due to issues.

Introducing RepBlock leads to notable improvements in overall tracking quality, despite its virtually unchanged computational footprint. The enhancement is particularly evident in the more complex worker scenario, where both detection recall and association stability show substantial gains. Moreover, the considerable increase in inference speed highlights its advantage for real-time deployment.

The GFF, designed to enrich fine-grained feature representation, exhibits

Table 7: Ablation study of RepBlock and GFF on the CraneView-N.

GFF	RepBlock	Class	MOTA (%)	MAP (%)
	w/o. re-parameterizing	Goods	98.00	97.98
		Worker	78.80	87.48
		Overall	88.40	92.73
	w/. re-parameterizing	Goods	97.00	96.97
		Worker	83.30	90.16
		Overall	90.10	93.56
✓	w/o. re-parameterizing	Goods	94.90	94.95
		Worker	82.50	90.74
		Overall	88.70	92.84
✓	w/. re-parameterizing	Goods	99.00	98.99
		Worker	90.50	95.36
		Overall	**94.70**	**97.17**

performance aligned with its relatively computational demand. It delivers meaningful improvements in the challenging worker category, reflecting its strong representation learning capacity.

When RepBlock and GFF are combined, the resulting configuration demonstrates clear complementarity. The efficiency benefits introduced by RepBlock offset the computational overhead of GFF, allowing the model to achieve both higher accuracy and competitive speed. This configuration attains the best overall tracking and detection performance among all ablation variants. The consistency and magnitude of these improvements indicate that the combined model excels in both detection reliability and temporal association stability.

4.5.4. Loss Analyzation

The variation of different loss components during the training process of the ConsMOT model on the CraneView-N dataset are shown in Fig.6. The monitored losses include the total loss and its sub-components, that is heatmap loss L_{hm}, offset loss L_{reg}, and identity loss L_{re}.

The heatmap loss decreases rapidly during the early training stages and gradually converges to stable values, indicating efficient feature learning and stable model optimization. Although the offset and identity losses exhibit slight fluctuations, they maintain an overall downward trend, suggesting that ConsMOT effectively balances detection precision and identity consistency

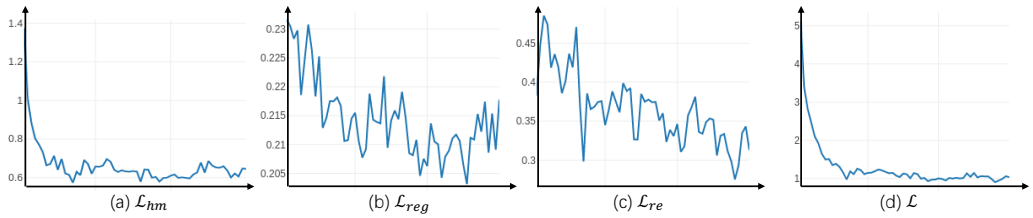


Figure 6: Loss visualization of ConsMOT training on CraneView-N.

during joint training.

These results demonstrate that ConsMOT achieves stable convergence and reliable feature representation learning for MOT tasks in complex construction-site environments.

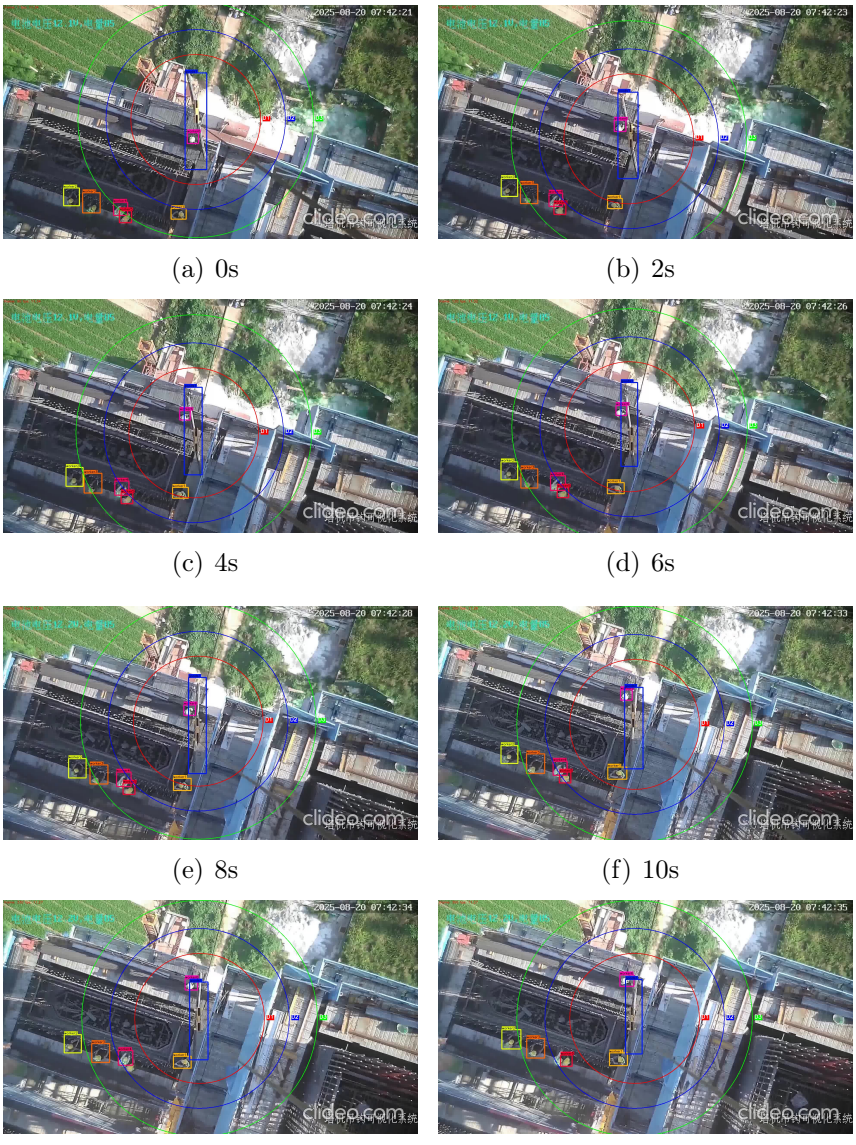
4.5.5. Collision Detection

The visualization results of consecutive frames for the human–machine collision warning process based on the CraneView-N dataset are shown in Fig.7. The visualization highlights how risk levels evolve as workers move relative to the hoisted good.

At the beginning of the sequence, the worker marked in purple (worker6) is positioned extremely close to the good within the red danger zone D1, establishing an immediate high-risk condition. In contrast, the workers marked in light orange (worker1), pink (worker4), red (worker5), dark orange (worker3), and yellow (worker2) remain inside or outside the green safe zone D3, maintaining low-risk states that require only routine monitoring.

As the scene progresses from Fig. 7(b) to 7(d), the system captures a gradual escalation of risk primarily caused by the motion of the light-orange worker1. This worker repeatedly transitions from the green D3 region into the blue warning zone D2, which represents an intermediate-risk state. When entering D2, the system issues mid-level warnings because the worker is approaching trajectories that may intersect with the moving good. In parallel, the purple worker6 continues inside D1, sustaining persistent high-risk alarms. Meanwhile, the remaining workers—yellow, dark orange, pink and red—remain distributed across low-risk areas, confirming that the system maintains selective attention toward only those individuals whose motion patterns warrant it.

In the later frames Fig. 7(e) to 7(g), the spatial dynamics become more complex. The purple worker6 stays inside D1, and the warning system con-



(a) 0s (b) 2s

(c) 4s (d) 6s

(e) 8s (f) 10s

(g) 12s (h) 14s

Figure 7: Collision Warning for CraneView-N Dataset.

tinues to trigger high-intensity alerts and recommended emergency responses. The light-orange worker1 repeatedly moves along the boundary between D2 and D1, increasing the likelihood of dangerous proximity. The red worker5 also shows a brief approach toward D1, entering a warning state, while the yellow and dark-orange workers maintain safe distances outside D3. These transitions demonstrate that the system can sustain stable identification of high-risk trajectories while simultaneously updating intermediate-risk decisions as workers move across different spatial buffers.

By the final stage Fig. 7(h), the purple worker6 remains fully inside D1, and the system outputs the highest alert level to indicate extreme collision risk. The light-orange worker1 and red worker5 remain close to the boundary of D1 and require intensified monitoring, whereas the yellow and dark-orange workers stay outside D3 with minimal risk. This final configuration reflects a well-maintained risk hierarchy under continuous movement.

Overall, the complete sequence illustrates the spatiotemporal reasoning of the proposed distance-based warning framework. The system distinguishes safe, warning and dangerous states using green D3, blue D2, and red D1 zones and maintains consistent high-risk tracking of the purple worker6 while dynamically adjusting warnings for approaching workers such as the light-orange worker1. These results confirm that the system is capable of early risk identification, graded alert escalation and sustained high-risk monitoring in complex and dynamic construction environments.

5. Conclusion

In this work, we proposed ConsMOT, a multi-object tracking method specifically designed for overhead crane monitoring in smart construction sites. To address the challenges posed by small object detection under high-altitude top-down views, we proposed Gated Feature Fusion (GFF), a learnable cross-scale aggregation module that adaptively weighs multi-level features to improve robustness in dense worker–good interactions. To mitigate the computational overhead introduced by GFF, we further introduced RepBlock, a re-parameterizable convolutional method that fuses convolution and batch normalization layers at inference time, enabling faster processing without affecting training-time accuracy.

Despite its encouraging results, ConsMOT still relies on high-quality detections as input, and its performance may degrade in extremely cluttered scenes or under severe illumination changes. Moreover, the current distance

estimation assumes camera calibration stability, which might limit its transferability across sites with varying camera setups.

Future work will explore integrating adaptive anthropometric modeling, incorporating temporal attention mechanisms to further enhance trajectory consistency, and extending ConsMOT to multi-camera setups to achieve fully comprehensive safety monitoring across large-scale construction sites.

Acknowledgments

This work was supported by National Key R&D Program of China under Contract 2022ZD0119802. This work reflects only the author's views and the funders are not responsible for any use that may be made of the information it contains. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

[1] D. Almaskati, S. Kermanshachi, A. Pamidimukkala, K. Loganathan, Z. Yin, A review on construction safety: hazards, mitigation strategies, and impacted sectors, Buildings 14 (2) (2024) 526.

[2] A. Darda'u Rafindadi, B. Kado, A. M. Gora, I. B. Dalha, S. I. Haruna, Y. E. Ibrahim, O. Ahmed Shabbir, Caught-in/between accidents in the construction industry: A systematic review, Safety 11 (1) (2025) 12.

[3] C. Wang, The role of artificial intelligence in construction management: A case study of smart worksite systems.

[4] A. R. Mohammed, S. S. Ram, M. I. Ahmed, S. A. Kamran, Remote monitoring of construction sites using ai and drones (2024).

[5] S. B. Rathod, R. A. Mahajan, P. A. Khadkikar, H. R. Vyawahare, P. R. Patil, Improving workplace safety with ai-powered predictive analytics: enhancing workplace security, in: AI Tools and Applications for Women's Safety, IGI Global Scientific Publishing, 2024, pp. 232–249.

[6] Q. Fang, D. Castro-Lacouture, C. Li, Smart safety: big data–enabled system for analysis and management of unsafe behavior by construction workers, Journal of Management in Engineering 40 (1) (2024) 04023053.

[7] C. O. Ozobu, F. E. Adikwu, N. Odujobi, F. Onyekwe, E. O. Nwulu, Advancing occupational safety with ai-powered monitoring systems: A conceptual framework for hazard detection and exposure control, World Journal of Innovation and Modern Technology 9 (1) (2025) 186–213.

[8] J. Seong, H.-s. Kim, H.-J. Jung, The detection system for a danger state of a collision between construction equipment and workers using fixed cctv on construction sites, Sensors 23 (20) (2023) 8371.

[9] Y.-S. Lee, D.-K. Kim, J.-H. Kim, Deep-learning-based anti-collision system for construction equipment operators, Sustainability 15 (23) (2023) 16163.

[10] Y.-S. Shin, J. Kim, A vision-based collision monitoring system for proximity of construction workers to trucks enhanced by posture-dependent perception and truck bodies' occupied space, Sustainability 14 (13) (2022) 7934.

[11] Y. Ding, X. Luo, Monocular 2d camera-based proximity monitoring for human-machine collision warning on construction sites, arXiv preprint arXiv:2305.17931 (2023).

[12] A. Rashidi, G. L. Woon, M. Dasandara, M. Bazghaleh, P. Pasbakhsh, Smart personal protective equipment for intelligent construction safety monitoring, Smart and Sustainable Built Environment 14 (3) (2025) 835–858.

[13] J. Lee, S. Lee, Construction site safety management: a computer vision and deep learning approach, Sensors 23 (2) (2023) 944.

[14] M. Adil, G. Lee, V. A. Gonzalez, Q. Mei, Using vision language models for safety hazard identification in construction, arXiv preprint arXiv:2504.09083 (2025).

[15] K. Kim, S. Kim, D. Shchur, A uas-based work zone safety monitoring system by integrating internal traffic control plan (itcp) and automated object detection in game engine environment, Automation in Construction 128 (2021) 103736.

[16] S. Kumar, M. Poyyamozhi, B. Murugesan, N. Rajamanickam, R. Alroobaea, W. Nureldeen, Investigation of unsafe construction site conditions using deep learning algorithms using unmanned aerial vehicles, Sensors 24 (20) (2024) 6737.

[17] X. Jiao, C. Li, X. Zhang, J. Fan, Z. Cai, Z. Zhou, Y. Wang, Detection method for safety helmet wearing on construction sites based on uav images and yolov8, Buildings 15 (3) (2025) 354.

[18] Y. Wang, Y. H. Ng, H. Liang, C.-W. Chang, H. Chen, Bird's-eye view safety monitoring for the construction top under the tower crane, arXiv preprint arXiv:2506.18938 (2025).

[19] L. Joachim, W. Zhang, N. Haala, U. Soergel, Evaluation of the quality of real-time mapping with crane cameras and visual slam algorithms, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 43 (2022) 545–552.

[20] Z. Liu, J. Xu, C. W. K. Suen, M. Chen, Z. Zou, Y. Shi, Egocentric camera-based method for detecting static hazardous objects on construction sites, Automation in Construction 172 (2025) 106048.

[21] S. Wang, Automated non-ppe detection on construction sites using yolov10 and transformer architectures for surveillance and body worn cameras with benchmark datasets, Scientific Reports 15 (1) (2025) 27043.

[22] S. F. A. Zaidi, J. Yang, M. S. Abbas, R. Hussain, D. Lee, C. Park, Vision-based construction safety monitoring utilizing temporal analysis to reduce false alarms, Buildings 14 (6) (2024) 1878.

[23] A. Sun, X. An, P. Li, M. Lv, W. Liu, Near real-time 3d reconstruction of construction sites based on surveillance cameras, Buildings 15 (4) (2025) 567.

[24] M. A. Musarat, A. M. Khan, W. S. Alaloul, N. Blas, S. Ayub, Automated monitoring innovations for efficient and safe construction practices, Results in Engineering 22 (2024) 102057.

[25] S. Zheng, M. Soy, J. Lee, Falling objects/debris detection system using surveillance camera at construction site, in: ISARC. Proceedings of the

International Symposium on Automation and Robotics in Construction, Vol. 42, IAARC Publications, 2025, pp. 603–608.

[26] W. Tian, H. Li, H. Zhu, Y. Wang, X. Liu, R. Yang, Y. Xie, M. Zhang, J. Zhu, X. Wang, A review of smart camera sensor placement in construction, Buildings 14 (12) (2024) 3930.

[27] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, Repvgg: Making vgg-style convnets great again, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13733–13742.

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

[29] M. Zhang, S. Ge, Vision and trajectory–based dynamic collision prewarning mechanism for tower cranes, Journal of Construction Engineering and Management 148 (7) (2022) 04022057.

[30] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision (ECCV), 2014.

Declaration of interests

☐The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Chonghua Zhou reports financial support was provided by National Key R&D Program of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Title: ConsMOT: A Cross-Scale and Re-parameterized Multi-Object Tracker for Overhead Crane Safety Monitoring

Chonghua Zhou
organization: EEIS Department, University of Science and Technology of China，CCCC Mechanical & Electrical Engineering Co. Ltd, China Communications Construction Company;
email: zhouchonghua@mail.ustc.edu.cn;
contribution: Data collection, conducted main experiments, and drafted the manuscript;
Ruixuan Zhang
organization: College of Electronic Information and Automation, Tianjin University of Science and Technology;
email: zhangrx@tust.edu.cn;
contribution: Led the model design, conducted main experiments, and drafted the manuscript.
Ningzhi Chen
organization: College of Electronic Information and Automation, Tianjin University of Science and Technology;
email: chenningzhi@mail.tust.edu.cn;
contribution: Conducted experimental verification and comparison of the methods;
Haitao He
organization: College of Electronic Information and Automation, Tianjin University of Science and Technology;
email: hehaitao@mail.tust.edu.cn;
contribution: Conducted experimental verification and comparison of the methods;
Yixi Fu
organization: College of Electronic Information and Automation, Tianjin University of Science and Technology;
email: fuyixi@mail.tust.edu.cn;
contribution: Drafted the manuscript;
Kevin Liu
organization: Western University Department of Computer Science, Canada;
email: kliu469@uwo.ca;
contribution: Annotate the dataset.