

Article

Weakly Calibrated Multi-Camera Vision for Safety Distance Estimation Between Power Operation Workers and Energized Equipment

Wei Wei ¹ , Yi Zhang ^{1,*} , Guozheng Peng ¹, Zhengwei Chang ² and Chao Liu ³

¹ Beijing Key Laboratory of Intelligent Cognition and Analysis for Power Systems, China Electric Power Research Institute, Beijing 100192, China; weiwei_ncepu@163.com (W.W.); pengguozheng@epri.sgcc.com.cn (G.P.)

² Power Internet of Things Key Laboratory, State Grid Sichuan Electric Power Research Institute, Chengdu 610041, China; changzw@126.com

³ State Grid Corporation of China, Beijing 100031, China; liuchao@sgcc.com.cn

* Correspondence: 15624951781@163.com; Tel.: +86-15624951781

Abstract

Real-time estimation of the safety distance between power operation workers and energized equipment is essential for preventing electric shock, arc flash injury, and equipment accidents during live or near-energized operations. However, large-scale deployment remains difficult because conventional binocular ranging systems require accurate offline calibration, while monocular vision suffers from metric-scale ambiguity. This paper proposes a weakly calibrated multi-camera vision framework for safety distance estimation in power operation scenes. First, a cross-camera tracking strategy combining mask-guided ReID features, Kalman filtering, spatio-temporal constraints, and Hungarian matching is developed to maintain globally consistent worker identities. Second, a region-wise self-supervised depth estimation model is introduced: photometric consistency is imposed within worker regions, while cross-view feature reprojection is imposed within rigid equipment regions. Intrinsic parameter regularization and scene-inherent dimensional priors are then used to recover metric scale without placing additional calibration targets. Third, the YOLOv8-seg backbone is enhanced with a CNN-Transformer C2f_CT module to improve pixel-level segmentation of workers, energized equipment, and structural components under cluttered backgrounds. Experiments on three representative power operation scenarios show that the proposed method achieves 92.3% MOTA, 94.2% Rank-1 accuracy, 82.0% mAP, 8.7% relative distance error, and 22 FPS end-to-end speed. When using five-frame smoothing and uncertainty threshold, the proposed method can achieve 98.7% warning precision, 0.9% missed alarm rate and 1.3% false alarm rate. The results indicate that the proposed framework provides a practical balance between deployment cost, geometric accuracy, and real-time warning capability.



Academic Editor: Gianpaolo Vitale

Received: 9 June 2026

Revised: 3 July 2026

Accepted: 13 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: near-energized operation; safety distance estimation; weakly calibrated multi-camera vision; self-supervised depth estimation; cross-camera worker tracking; power operation safety

1. Introduction

To maintain power grid reliability and continuity of electricity supply, inspection, maintenance, and upgrading tasks are frequently performed under live or near-energized

conditions [1]. Such operations are often conducted in environments characterized by high voltage, strong current, dense equipment, and limited spatial clearance. Once the distance between a worker, a tool, and an energized component falls below the prescribed safety threshold, the operation may lead to electric shock, arc flash injury, or equipment failure [2,3]. Continuous, accurate, and real-time perception of the spatial relationship between workers and energized equipment is therefore a critical requirement for intelligent safety management at power operation sites.

Existing safety distance sensing techniques include electric field monitoring, BeiDou/ultra-wideband localization, LiDAR ranging, binocular vision, and ordinary camera visual perception [2–5]. Electric-field-based approaches can reflect local electromagnetic variations but are sensitive to equipment configuration, weather, and site layout [2]. Localization-based methods can provide worker coordinates but usually require wearable tags or dedicated base stations. Radar/LiDAR can directly output metric depth without scale ambiguity and maintain stable measurement in live equipment safe distance monitoring [6,7]. Nevertheless, their hardware cost, installation difficulty, registration requirement, and real-time processing burden restrict large-scale deployment in temporary work areas. In contrast, camera-based perception is low-cost, widely deployable, and compatible with existing surveillance infrastructures, making it attractive for scalable safety monitoring in power systems.

Vision-based distance estimation has evolved from early CNN-based monocular depth prediction and end-to-end stereo regression to self-supervised monocular depth learning, deep stereo matching, and multi-camera metric depth estimation [8–13]. Monodepth2 established an effective photometric reconstruction paradigm for self-supervised monocular depth estimation [13], while recent lightweight and Transformer-based methods, such as Lite-Mono, further improve the balance between depth accuracy and inference efficiency [14]. Multi-frame methods, including ManyDepth and ProDepth, exploit temporal information to improve depth consistency, but they may still suffer from dynamic object interference, scale ambiguity, and non-standard scene geometry [15,16]. Recent multi-camera metric depth and self-calibration studies further show that cross-view consistency and online extrinsic estimation are important for scalable depth recovery under weak geometric constraints [17,18]. References [6,19] have demonstrated that camera baseline, angle, number of arrangements, and target distance have an impact on the accuracy of 3D ranging. These advances provide new technical foundations for three-dimensional perception in power operation scenes.

For near-energized safety monitoring, most existing vision-based studies still adopt calibrated binocular stereo matching, semantic segmentation combined with stereo vision, pseudo-LiDAR reconstruction, or image–LiDAR fusion [20]. For example, binocular stereo perception and safety region segmentation have been used to estimate the approach distance of large-scale construction equipment in substations [21–23]. Lightweight feature extraction and binocular matching have also been applied to distance detection between substation workers and energized equipment [24], and improved semantic segmentation networks combined with stereo vision have been investigated for safety distance measurement [25]. Instance segmentation combined with pseudo-LiDAR has been further explored for safe distance monitoring of live equipment [5]. Image–LiDAR fusion methods such as GAFusion [26], ReliFusion [27], and CalibRefine [28] use sparse radar metric prior input to compensate visual feature and realize adaptive weighted depth fusion. More recently, multi-view spatial reasoning has been introduced into construction safety violation recognition, indicating that worker–object interaction and three-dimensional scene context are important for reliable safety monitoring in complex industrial environments [29]. These studies demonstrate the feasibility of vision-based safety distance perception. Nevertheless, many of them still depend on fixed binocular rigs, explicit LiDAR registration, rigorous cal-

ibration, or stable scene geometry, which increases deployment cost and makes cross-scene migration difficult.

Near-energized distance estimation in real power operation scenes remains challenging for three reasons. First, dense equipment, complex textures, and frequent occlusions make pixel-level localization of workers and energized objects difficult. Second, workers often wear visually similar uniforms and move across multiple cameras, causing identity switches and discontinuous trajectories. Third, ordinary cameras usually lack accurate intrinsic and extrinsic parameters as well as hardware synchronization, so depth prediction suffers from metric-scale ambiguity. This paper does not attempt to remove geometric constraints entirely. Instead, it focuses on a weakly calibrated setting in which traditional offline stereo calibration and additional calibration targets are not required, while the rigid structure of in-scene equipment and dimensional priors are used to recover metric safety distances.

This study proposes a weakly calibrated multi-camera visual method for real-time safety distance estimation between power operation workers and energized equipment. First, a region-wise self-supervised depth estimation model combined with photometric consistency loss, cross-view feature reprojection loss, depth smoothness loss and intrinsic parameter regularization loss is established under formal weak calibration constraints. Second, a dual scene prior metric-scale recovery strategy is introduced, utilizing standardized power equipment dimensions and human body height ranges to convert dimensionless relative depth to real-world 3D distances. Third, the C2f_CT enhanced YOLOv8-seg segmentation module and mask-guided spatio-temporal cross-camera tracking module are constructed to improve the tracking accuracy. Fourth, an end-to-end perception and warning pipeline is established by integrating scene element segmentation, cross-camera worker association, region-wise depth estimation, metric-scale recovery, and voltage-threshold-based safety distance assessment, thereby providing a practical visual solution for real-time near-energized operation monitoring.

2. Materials and Methods

2.1. Overall Framework

The proposed method consists of four modules: scene element instance segmentation, cross-camera worker tracking, region-wise self-supervised depth estimation, and metric-scale recovery with safety distance calculation. The segmentation module outputs bounding boxes, categories, and pixel-level masks for workers, energized equipment, and relevant scene elements. The tracking module assigns a globally unique identity to the same worker across multiple cameras by combining motion prediction, appearance embeddings, and spatio-temporal feasibility. The depth module constructs different self-supervised constraints for worker and equipment regions to obtain a stable relative depth map and weak geometric relations. Finally, metric-scale is recovered using standard equipment dimensions and human-scale priors, after which the minimum three-dimensional distance between worker surface points and energized equipment surface points is compared with the safety threshold to generate warning signals. The key idea of the proposed framework is not to eliminate geometric constraints, but to reorganize them according to the physical properties of different scene regions. Worker regions are treated as non-rigid and dynamically changing targets, for which strict stereo geometry is unreliable; therefore, local temporal appearance consistency is used as a weak self-supervised cue. In contrast, energized equipment and structural components are regarded as locally rigid regions, where cross-view feature correspondences can provide more stable geometric constraints for weak extrinsic estimation and relative depth refinement. After obtaining the relative depth, scene-inherent dimensional priors are used to recover the metric scale, and the resulting three-dimensional worker–equipment distance is further converted into a

safety warning decision according to the voltage-level threshold. This design links visual perception, weak geometric reasoning, metric scale recovery, and safety decision-making within a unified pipeline. The overall framework is shown in Figure 1.

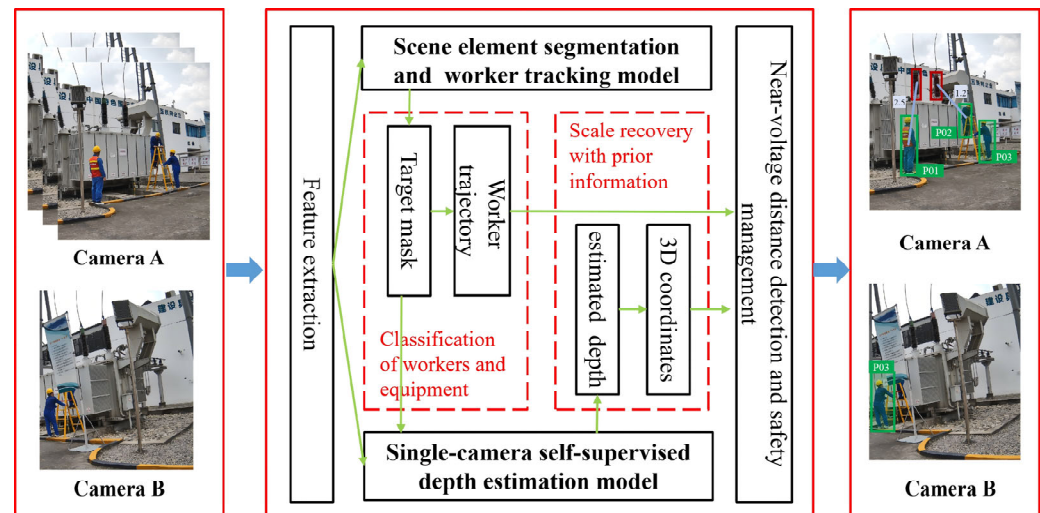


Figure 1. Overall framework of the proposed weakly calibrated multi-camera safety distance estimation method.

The main contributions of this study are summarized as follows:

1. A weakly calibrated multi-camera safety distance estimation framework is developed for near-energized power operation scenarios. Different from conventionally calibrated stereo, self-calibrated stereo, and fully uncalibrated vision, the proposed weakly calibrated framework exploits ordinary surveillance cameras, coarse intrinsic initialization, constrained intrinsic fine-tuning, online weak extrinsic estimation and dimensional priors to balance deployment cost and metric accuracy in temporary or non-standard work areas.
2. A region-decoupled self-supervised depth learning strategy is proposed by explicitly distinguishing non-rigid worker regions from rigid equipment regions. Adjacent frame photometric consistency is imposed on worker masks, whereas cross-view feature reprojection is applied to equipment masks, enabling more reliable relative depth optimization under weak geometric constraints.
3. A scene-prior-guided metric-scale recovery mechanism is introduced to convert dimensionless relative depth into real-world safety distance by using standard equipment dimensions and human-scale ranges, without placing additional calibration targets on-site.
4. A lightweight C2f_CT CNN–Transformer fusion block is introduced for YOLOv8-seg backbone to enhance the segmentation accuracy under cluttered power equipment backgrounds. Based on the pixel-level segmentation results, a mask-guided cross-camera ReID method matching with global Hungarian optimization is designed to reduce identity switch for uniformed power workers.

2.2. Scene Element Segmentation and Cross-Camera Worker Tracking

The first stage aims to obtain stable pixel-level masks of workers, energized equipment, and structural components, and then to maintain worker identity consistency across cameras. YOLOv8-seg is adopted as the base instance segmentation architecture because of its favorable balance between detection accuracy, segmentation capability, and real-time inference efficiency [30]. To improve its ability to model long-range structural relations in power operation scenes, a lightweight Transformer branch is introduced into the C2f

module. The improved multi-scale features are fused through the FPN-PAN neck and then fed to the decoupled detection head and the mask head. For tracking, Kalman filtering is used to maintain short-term trajectories within a single camera, while mask-cropped ReID features, temporal constraints, spatial reachability, and Hungarian matching are used to associate trajectories across cameras.

2.2.1. Context-Aware Feature Extraction Network

The original YOLOv8 backbone consists of alternating convolutional and C2f modules, which can efficiently extract multi-scale semantic features. However, its C2f block is dominated by local convolutions and has limited capacity to represent long-range relationships among towers, conductors, insulator strings, cabinet boundaries, and worker–equipment spatial layouts. In power operation scenes, objects vary substantially in scale, backgrounds are cluttered, and workers and equipment are frequently occluded. Relying only on local texture features may therefore cause missing detections of small components or misclassification of visually similar equipment. To address this problem while preserving inference efficiency, a lightweight Transformer branch is embedded to strengthen global context modeling.

As shown in Figure 2, the proposed C2f_CT replaces the original bottleneck structure inside C2f with a parallel Bottleneck_CT composed of a CNN local branch and a lightweight Transformer global branch. A preceding 3×3 convolution normalizes the input feature and enhances local smoothness. The CNN branch extracts fine-grained local features such as worker contours, metal fitting edges, and equipment textures. The Transformer branch flattens the two-dimensional feature map into a sequence and uses multi-head self-attention to capture long-distance dependencies among conductors, towers, and cabinet structures. The two branches are fused through a learnable residual weighting mechanism:

$$F_{CT} = \alpha F_{CNN} + (1 - \alpha) F_{TR}, \quad \alpha = \sigma(g(F)) \quad (1)$$

where F_{CNN} and F_{TR} denote the features from the CNN branch and the Transformer branch, respectively, and α is a learnable fusion coefficient generated from the input feature F . This design improves scene structure representation without changing the input dimension of the subsequent YOLOv8-seg neck.

2.2.2. Object Detection and Instance Segmentation

The multi-scale features extracted by the improved backbone are fed into the YOLOv8-seg neck, where FPN-PAN fusion is used to integrate semantic and spatial information across scales. The detection head predicts the object category and bounding box, thereby providing spatial constraints for mask generation. The mask head follows the prototype mask and mask coefficient design: a set of shared mask prototypes is first learned, and an instance-specific coefficient vector is then predicted for each detected object. The final mask of instance i is generated by a linear combination of the prototype masks:

$$M_i = \sigma \left(\sum_{k=1}^K c_{ik} P_k \right) \quad (2)$$

where P_k is the k th mask prototype, c_{ik} is the corresponding coefficient of instance i , K is the number of prototypes, and $\sigma(\cdot)$ denotes the sigmoid activation. The resulting masks provide pixel-level inputs for worker region cropping, equipment region feature matching, and subsequent distance calculation.

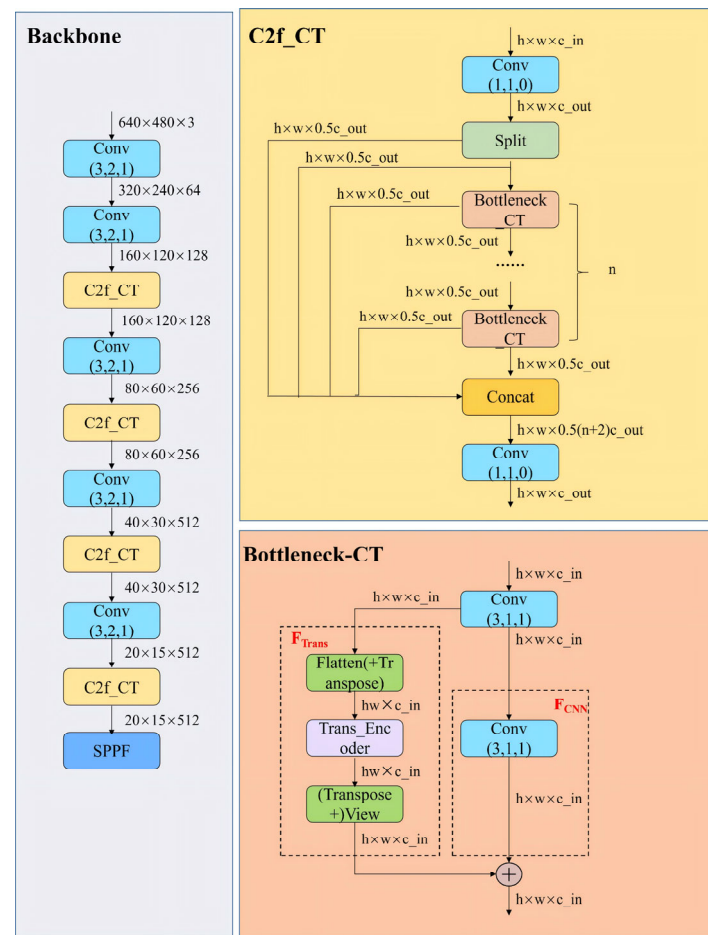


Figure 2. Improved C2f_CT feature extraction module. The module preserves local detail through a CNN branch and introduces global structural context through a lightweight Transformer branch.

2.2.3. Worker Tracking

The tracking stage consists of intra-camera temporal association and cross-camera global association. Within a single camera, the Kalman filter algorithm is used to model the state of each worker trajectory, including target center, scale, aspect ratio, and velocity:

$$\mathbf{x}_t = [u_t, v_t, s_t, r_t, \dot{u}_t, \dot{v}_t, \dot{s}_t]^T \quad (3)$$

where u_t and v_t are the target center coordinates, s_t is the bounding-box scale, r_t is the aspect ratio, and $\dot{u}_t$, $\dot{v}_t$, and $\dot{s}_t$ describe the temporal variations. Prediction and observation updates maintain trajectory continuity under short-term occlusion, camera shake, and missed detections.

For cross-camera association, each detected worker is cropped according to the segmentation mask before ReID feature extraction. This mask-guided crop reduces the interference caused by energized equipment, fences, ground surfaces, and tools. A global trajectory memory stores each historical trajectory with its global ID, ReID feature, timestamp, spatial position, and motion state. The cross-camera matching cost is defined as:

$$C_{ij} = \lambda_a (1 - \cos(\mathbf{f}_i, \mathbf{f}_j)) + \lambda_t \Delta t_{ij} + \lambda_s D_{ij}^S + \lambda_m D_{ij}^M \quad (4)$$

where f_i and f_j are normalized ReID embeddings, Δt_{ij} is the temporal interval, D_{ij}^S represents the spatial reachability cost, and D_{ij}^M is the motion prediction cost. The Hungarian algorithm is then applied to obtain the globally optimal association. This strategy re-

duces identity switches caused by similar uniforms, viewpoint changes, and short-term disappearance across cameras.

2.3. Region-Wise Self-Supervised Depth Estimation

2.3.1. Network Architecture

Power operation scenes exhibit heterogeneous geometric properties. Energized equipment, cabinets, insulators, conductors, and supporting structures are generally static or locally rigid, whereas workers are non-rigid objects with time-varying poses, partial occlusions, and frequent cross-view appearance changes. Applying a single self-supervised depth constraint to all pixels may therefore introduce inconsistent supervision: rigid equipment regions are more suitable for cross-view geometric reprojection, while worker regions are more suitable for local temporal appearance constraints. To address this issue, a region-wise self-supervised depth estimation model is developed under weak calibration. Here, weak calibration refers to a practical setting in which ordinary site cameras are not arranged as a fixed and precisely calibrated stereo rig. The initial intrinsic matrix is constructed from coarse assumptions, including an image center principal point, approximately square pixels, and zero skew, while only a small intrinsic perturbation is optimized under regularization. The cross-camera relation is not obtained from offline calibration, but is estimated from validated correspondences in rigid equipment regions. Therefore, the proposed model retains geometric constraints, but shifts the requirement from precise pre-calibration to online weak geometry estimation using stable scene structures. Similar multi-camera metric depth estimation and self-supervised extrinsic calibration studies have also shown that jointly estimating depth and camera relations can reduce the dependence on accurately pre-calibrated extrinsic parameters in multi-camera perception systems [17,18].

The network first predicts a global relative depth map from the current frame. Instance masks generated by the segmentation module are then used to partition the image into worker regions and equipment regions. For worker regions, adjacent frame photometric consistency is adopted within valid masks to constrain local relative depth variation without assuming strict rigid body motion. For equipment regions, stable feature correspondences across cameras are used to construct a cross-view reprojection constraint, thereby exploiting the rigidity of electrical equipment to estimate weak geometric relations between cameras. Depth smoothness and intrinsic parameter regularization are further introduced to suppress local depth noise and prevent unrealistic camera parameter drift. Through this region-decoupled design, the proposed model improves depth stability in cluttered power operation scenes while avoiding the requirement for precise offline stereo calibration. The architecture is shown in Figure 3.

The depth network follows an encoder–decoder design similar to Monodepth2, which uses photometric reconstruction as a self-supervised signal for monocular depth learning [13]. The input is the current frame from camera A, and the output is a multi-scale relative depth map. The encoder reuses the improved feature extraction network and produces feature maps at $1/2$, $1/4$, $1/8$, $1/16$, and $1/32$ resolutions. The decoder progressively recovers spatial details through upsampling and skip connections and outputs relative depth at four scales. It should be emphasized that the output is dimensionless relative depth; metric distance is obtained only after scale recovery using scene-inherent priors.

For worker regions, this study uses local appearance consistency between adjacent frames to constrain relative depth. Because a worker is a non-rigid object, worker motion is not treated as strict camera motion. Instead, the adjacent frame local appearance change within the worker mask is used as a weak geometric constraint. The reconstruction loss is computed only inside valid worker masks, thereby preventing the static background from dominating the optimization. The pose estimation subnetwork adopts a ResNet18 encoder framework and

a lightweight decoder to output an approximate 6DoF local transformation parameters that describes the projection relationship of the worker region between adjacent frames.

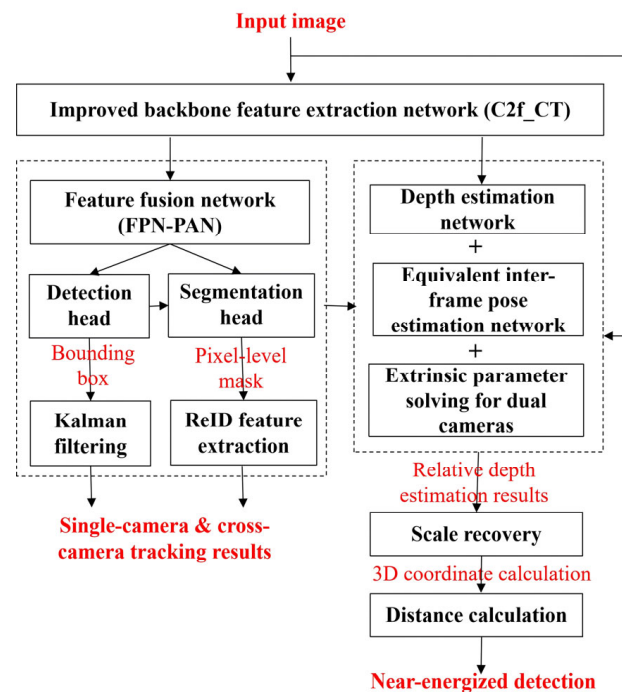


Figure 3. Region-wise self-supervised depth estimation model under weak calibration.

For equipment regions, cross-view feature point reprojection is constructed by exploiting the short-term rigidity of electrical equipment. Since two ordinary cameras are usually installed non-parallelly, directly applying standard stereo disparity estimation is vulnerable to viewpoint difference and uncalibrated camera parameters. Stable features are extracted within the equipment mask, and matched point sets P^A , P^B are used to estimate a weak extrinsic relation between cameras. Specifically, SIFT keypoints and descriptors are extracted from the equipment mask regions in grayscale images; coarse matches are obtained through Lowe's ratio test, and RANSAC-based consensus filtering is used to reject outliers and improve the robustness of cross-view feature matching [31]. The matched points are split into a parameter estimation set and a validation set. The former is used to recover three-dimensional points using current relative depth and to estimate a weak pose transformation from camera A to camera B (rotation matrix R , translation vector t) through scale-free Procrustes alignment, while the latter is used to compute reprojection errors and optimize equipment region depth and geometric parameters.

The Procrustes rigid alignment pipeline [32] is decoupled: only rotation and translation components are solved, while the native scale estimation term is disabled. The global metric scale is fully computed separately by the scene prior module, completely eliminating scale coupling conflicts between geometric parameters and preventing mutual interference between the inherent Procrustes scaling and equipment dimensional priors. The scale-free Procrustes implementation process is as follows.

1. Weighted Centroid Centralization

Compute total weight W and weighted centroids $\bar{P}^A$, $\bar{P}^B$ for two point sets. Decentralize each point by subtracting the centroid to separate translation from rotation and scale parameters, yielding zero-mean point sets $\tilde{P}^A$, $\tilde{P}^B$.

$$\bar{P}^A = \frac{1}{W} \sum_{i=1}^N \omega_i p_i^A \quad (5)$$

$$\bar{P}^B = \frac{1}{W} \sum_{i=1}^N \omega_i p_i^B \quad (6)$$

$$W = \sum_{i=1}^N \omega_i \quad (7)$$

$$\tilde{p}_i^A = p_i^A - \bar{P}^A \quad (8)$$

$$\tilde{p}_i^B = p_i^B - \bar{P}^B \quad (9)$$

where ω_i denotes the i -th set of feature matching confidence weights, normalized by local feature similarity scores. W is the sum of all matching weights.

2. Covariance Matrix Construction and SVD for Optimal Rotation

Construct weighted cross-covariance matrix H from decentralized matched points. Perform singular value decomposition on H to get orthogonal matrices U, V and singular value matrix Σ . Enforce determinant constraint to avoid reflection transformation and solve the optimal orthogonal rotation matrix R .

$$H = \sum_{i=1}^N \omega_i \tilde{p}_i^B \tilde{p}_i^{A T} \quad (10)$$

$$H = U \Sigma V^T \quad (11)$$

$$R = U \cdot \begin{bmatrix} 1 & 0 \\ 0 & \det(UV^T) \end{bmatrix} \cdot V^T \quad (12)$$

3. Closed-Form Estimation of Global Scale Factor

Derive the optimal 2D translation vector t using two centroids, estimated rotation and scale.

$$t = \bar{P}^B - R \bar{P}^A \quad (13)$$

The camera intrinsic parameters are assumed using a weak calibration method proposed in this paper. This method is rigorously distinguished from conventional calibrated stereo, global self-calibration and fully uncalibrated vision systems, with explicit rules for camera parameter classification as follows (Table 1):

Table 1. Comparison of key characteristics of four types of visual calibration schemes.

Scheme	Intrinsic Parameters	Extrinsic Parameters	Calibration Target	Computational Cost
Conventional fully calibrated stereo	Fixed via offline calibration	Fixed via offline calibration	Checkerboard required	Low, with strict deployment restrictions
Self-calibrated multi-camera	Optimized via global bundle adjustment	Optimized via global bundle adjustment	No target required	Extremely high, impractical for on-site deployment
Fully uncalibrated vision	No predefined initial assumptions	Unconstrained	None	Metric scale is completely unobservable, distance measurement fails
Proposed weak calibration	Coarse initialization + regularized fine-tuning	Online solved via rigid equipment features	No extra target needed	Low, well-suited for temporary operation sites

2.3.2. Loss Functions

The worker region photometric reconstruction loss is built using the local transformation predicted by the pose network and the relative depth predicted by the depth network. The worker region at time t is reprojected to the adjacent frame, and the photometric con-

sistency loss is calculated only within valid worker masks. The loss combines L1 difference and structural similarity:

$$L_p = \frac{1}{N_p} \sum_{x \in M_p} \rho(I_t(x), \hat{I}_t(x)) \quad (14)$$

$$\hat{I}_t(x) = I_{t+1} \left(\pi \left(K T_{t \rightarrow t+1} D_t(x) K^{-1} \tilde{x} \right) \right) \quad (15)$$

where $\hat{I}_t$ denotes the reconstructed image at time t , $\pi(\cdot)$ denotes the pixel back-projection and reprojection operator determined by depth, relative transformation, and camera intrinsics; D_t is the relative depth map of camera A at time t ; M_p is the worker mask; N_p is the number of valid pixels in the worker mask; and x is the pixel coordinate.

For equipment regions, the predicted depth is used to recover the three-dimensional coordinates of the matching points. Then, the Procrustes scale alignment method is used to obtain pose transformation parameters based on calculation set feature points. These extrinsic parameters are used to reproject validation set feature points from camera A onto camera B. The equipment feature point reprojection loss is defined as:

$$L_d = \frac{1}{N_d} \sum_{i=1}^{N_d} \| p_i^B - \hat{p}_i^B \|_2 \quad (16)$$

$$\hat{p}_i^B = R p_i^A + t \quad (17)$$

where N_d is the number of validated feature correspondences in the equipment region, p_i^B denotes the observed feature point in camera B, and $\hat{p}_i^B$ denotes the reprojected point in camera B obtained from its matched point p_i^A in camera A. Specifically, $\hat{p}_i^B$ is computed through the transformation parameters between two cameras. The loss therefore penalizes the reprojection discrepancy between the observed feature point and its geometrically projected counterpart in the target camera.

A depth smoothness loss is further applied to suppress local noise while preserving true depth discontinuities near equipment edges and worker contours:

$$L_s = \sum_x \left(|\partial_x D(x)| e^{-|\partial_x I(x)|} + |\partial_y D(x)| e^{-|\partial_y I(x)|} \right) \quad (18)$$

where ∂_x and ∂_y denote horizontal and vertical gradients. Large image gradients reduce the smoothness penalty to preserve depth discontinuities, while homogeneous regions receive stronger smoothness constraints.

The camera intrinsic matrix K contains focal lengths f_x and f_y and principal point coordinates c_x and c_y :

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (19)$$

Because ordinary on-site cameras usually do not provide accurate calibration parameters, the image center is used as the initial principal point and square pixels are assumed to obtain an initial matrix K_0 . A learnable perturbation ΔK is then introduced for small-range adaptive correction:

$$K = K_0 + \Delta K \quad (20)$$

The intrinsic parameter regularization term limits the magnitude of ΔK to prevent unrealistic parameter drift and unstable depth scale:

$$L_k = \| K - K_0 \|_F^2 \quad (21)$$

The overall training objective is the weighted sum of the four loss terms:

$$L = \lambda_p L_p + \lambda_d L_d + \lambda_s L_s + \lambda_k L_k \quad (22)$$

where λ_p , λ_d , λ_s , and λ_k are the weights of the worker photometric reconstruction loss, equipment reprojection loss, depth smoothness loss, and intrinsic regularization loss, respectively.

2.4. Metric Scale Recovery and Safety Distance Calculation

The self-supervised model produces relative depth, which describes the near–far relationship among pixels but cannot be directly used as a metric distance. This also defines the observability boundary of the proposed method. The depth network can provide stable relative depth only under visual and geometric constraints, whereas absolute metric distance becomes observable only when at least one reliable dimensional prior exists in the scene. Such priors may come from standard equipment dimensions, including insulator string length, cabinet width, conductor spacing, or typical metal fitting size. In practical deployment, the dimensional prior is selected according to a reliability ranking strategy rather than being arbitrarily assigned. Rigid equipment size priors are given the highest priority when the equipment category is recognizable, the reference segment is sufficiently visible, and its standard dimension can be obtained from equipment specifications, site records, or operation manuals. Typical examples include cabinet width, insulator string length, conductor spacing, and standard metal fitting dimensions. The reliability of a candidate prior is evaluated according to four factors: category confidence, visible ratio of the reference segment, cross-view feature matching stability, and availability of standard dimensional records. Human height priors are used only as auxiliary cues when no reliable equipment size prior is available, because worker posture and partial occlusion may introduce larger scale uncertainty. If these priors are unavailable or unreliable, the method can still describe relative near–far relationships, but the recovered metric distance should be regarded as uncertain. To obtain real-world safety distances, this study uses scene-inherent priors such as standard dimensions of electrical equipment and human-scale ranges. The proposed method does not require additional calibration boards or known-size reference targets to be placed on-site; however, it still relies on dimensional priors of recognizable in-scene objects to convert relative depth into metric scale.

First, stable equipment with known standard dimensions, such as insulator string length, cabinet width, conductor spacing, or typical metal fitting size, is selected within the equipment mask and acted as the fixed prior information. Its image space span is combined with relative depth to obtain the scale coefficient:

$$s = \frac{L_{real}}{\|X_a^{rel} - X_b^{rel}\|_2} \quad (23)$$

where L_{real} is the known physical length of the selected equipment segment, and X_a^{rel} and X_b^{rel} are its two endpoints reconstructed in the relative depth coordinate system. When a clear equipment scale is unavailable, the height range of the worker can be used as an auxiliary prior for scale correction; this prior is used only for metric conversion and not for identity recognition.

To avoid over-confident scale recovery from unreliable reference objects, each Scale Cue is assigned a reliability score. When the selected equipment segment is heavily occluded, weakly matched across cameras, or inconsistent with standard dimensional records, its reliability is reduced and the recovered metric distance is marked as uncertain. In such cases, the system either switches to an alternative equipment size prior or uses the human-scale prior only as a secondary correction. Therefore, the proposed method does not assume that every detected

object can provide a valid metric reference; instead, metric recovery is performed only when a sufficiently reliable in-scene dimensional prior is available.

After scale recovery, each valid pixel is converted from the image coordinate system to a three-dimensional metric point:

$$X(u, v) = sD(u, v)K^{-1}[u, v, 1]^T \quad (24)$$

The worker surface point set $\mathcal{P}$ and energized equipment surface point set $\mathcal{E}$ are then constructed from their corresponding masks. The safety distance is defined as the minimum Euclidean distance between the two point sets:

$$d = \min_{X_p \in \mathcal{P}, X_e \in \mathcal{E}} \|X_p - X_e\|_2 \quad (25)$$

When d is lower than the safety threshold corresponding to the voltage level, the system outputs a near-energized risk warning. In practical deployment, consecutive frame distance trends and uncertainty thresholds can be further combined to reduce false alarms and improve conservative decision-making.

3. Results

3.1. Data Acquisition and Hardware Configuration

Two ordinary USB high-definition cameras were used to acquire dual-view video with a resolution of 1920×1080 . The cameras were installed at different positions in the operation area and were not arranged in a strictly parallel stereo configuration. For example, camera 1 was deployed on the eastern side of the work area, while camera 2 was placed on the northern side, with an approximate viewing angle difference of 20 degrees. The system did not rely on hardware synchronization; frame-level alignment was implemented through software timestamps. Unlike traditional calibrated binocular systems, the proposed method did not conduct offline precise stereo calibration. Instead, weak geometric relations were estimated using equipment region feature matching and intrinsic parameter regularization.

3.2. Dataset, Ground Truth Distance Annotation, and Evaluation Metrics

The dataset was collected from three representative power operation scenarios, including a 110 kV substation, a transmission line maintenance site, and a distribution station. These scenarios cover both outdoor open-area operations and indoor narrow-space inspection conditions. The outdoor data contain workers operating near insulators, conductors, metal brackets, tower structures, and other energized components, whereas the indoor data contain workers performing inspection tasks near dense cables, switch cabinets, and metallic equipment under relatively weak illumination. A total of 90 high-definition videos were collected, including 20, 30, and 40 videos from the three scenarios, respectively. Each video lasted approximately 25 min and was recorded at 30 FPS with a resolution of 1920×1080 . The raw video data therefore contained approximately 4.05 million frame images, providing sufficient temporal coverage for evaluating worker tracking and safety distance estimation under different operation conditions.

To reduce temporal redundancy among adjacent video frames and control annotation cost, representative frames and worker–equipment instances were sampled for manual annotation. More than 40,000 annotated image–label pairs were constructed for the experiments. The annotations include worker bounding boxes, equipment bounding boxes, instance-level masks, cross-camera worker IDs, and worker–equipment distance labels. The number of workers varied across operation scenes, ranging from one to five workers per video. The equipment categories were not fixed across scenarios and covered approxi-

mately 50–60 types of power operation objects, including insulators, conductors, switch cabinets, metal fittings, supporting structures, tools, and other site-specific components. The detailed content of the dataset is shown in the table below (Table 2).

Table 2. Dataset statistical summary.

Scenario	Videos	Annotated Pairs	Worker IDs	Equipment Classes	Camera View Angle	Resolution	FPS
110 kV substation	20	9200	1–5	56	15–25°	1920 × 1080	30
Transmission line site	30	13,500	1–4	48	20–30°	1920 × 1080	30
Distribution station	40	17,300	1–3	52	10–20°	1920 × 1080	30
Total	90	40,000+	1–5	50–60	10–30°	1920 × 1080	30

The annotated dataset was divided at the video level into training, validation, and test sets at a ratio of 7:2:1. All frames from one continuous operation video are assigned to a single subset (train/val/test = 7:2:1) to eliminate cross-scene data leakage. Accordingly, approximately 28,000, 8000, and 4000 annotated pairs were used for training, validation, and testing, respectively. To evaluate cross-scene generalization and avoid data leakage, videos from the same continuous operation segment were assigned to only one subset, and the test set contained independent videos from different operation scenes and camera viewpoints.

For distance evaluation, a pre-calibrated binocular depth camera is deployed on-site to provide reference distance labels. Frame alignment between our dual surveillance cameras and the depth camera is implemented via software timestamps with synchronization error limited within 10 ms. Invalid depth frames caused by surface reflection, weak texture and heavy occlusion are automatically filtered. All 3D surface points of workers and energized equipment are extracted, and the minimum Euclidean distance is taken as the ground truth safety distance per frame, with centimeter-level baseline measurement uncertainty. This distance refers to the actual worker–equipment separation involved in near-energized operation, rather than the distance between the camera and the worker. Annotations include instance masks, global cross-camera worker IDs and voltage-dependent safety threshold labels.

In frames containing multiple workers and multiple equipment instances, all valid worker–equipment pairs were first identified according to the instance masks and cross-camera IDs. The minimum distance among these valid pairs was then used as the ground truth safety distance, because this value corresponds to the most conservative near-energized risk condition. Although the calibrated binocular depth camera provided the ground truth distance for evaluation, measurement uncertainty may still arise from the effective sensing range of the depth camera, reflective or weak-texture equipment surfaces, partial occlusion of worker body boundaries or equipment edges, and closest-point selection. Under normal visible conditions, the measurement uncertainty was controlled at the centimeter level. These measured ground truth distances were used to calculate the relative distance error of the proposed method.

Several metrics were used to evaluate the system. The segmentation performance was measured using mean average precision (mAP):

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP(q) \quad (26)$$

where Q is the number of object categories and $AP(q)$ is the area under the precision–recall curve of the q th category. Multi-object tracking accuracy (MOTA) was used to assess tracking performance:

$$\text{MOTA} = 1 - \frac{\sum_t (FP_t + FN_t + \text{IDSW}_t)}{\sum_t GT_t} \quad (27)$$

where FP, FN, IDSW, and GT denote false positives, false negatives, identity switches, and ground truth targets, respectively. Rank-1 accuracy was used to evaluate cross-camera identity retrieval.

$$\text{Rank} - 1 = \frac{\text{number of queries with top} - 1 \text{ correct match}}{\text{total number of queries}} \times 100\% \quad (28)$$

The relative distance error (RDE) was defined as

$$\text{RDE} = \frac{|d_{pred} - d_{gt}|}{d_{gt}} \times 100\% \quad (29)$$

where d_{pred} is the distance estimated by the proposed method and d_{gt} is the ground truth distance measured by the calibrated binocular depth camera.

The average absolute error (MAE) measures the average absolute distance deviation.

$$\text{MAE} = \text{mean}(|d_{pred} - d_{gt}|) \quad (30)$$

The root mean square distance error (RMSE) emphasizes larger errors.

$$\text{RMSE} = \text{sqrt}\left(\text{mean}\left(|d_{pred} - d_{gt}|^2\right)\right) \quad (31)$$

95% Confidence Interval (95% CI) means that, under the distribution of test samples, there is a 95% probability that the true population performance metric of the model falls within this interval:

$$95\% \text{CI} = \text{mean}(e) \pm 1.96 \cdot \frac{\text{std}(e)}{\sqrt{N}} \quad (32)$$

$$e = |d_{pred} - d_{gt}| \quad (33)$$

where $\text{mean}(e)$ is the mean value of distance estimated error, and $\text{std}(e)$ is the standard deviation of error. N denotes the total number of valid test samples.

For warning-level evaluation, a predicted warning is regarded as positive when the estimated worker–equipment distance is lower than the voltage-dependent safety threshold. The missed alarm rate, false alarm rate, warning precision, warning recall, and warning latency are computed by comparing the predicted warning state with the ground truth warning state derived from the measured safety distance:

$$\text{missed alarm rate} = \frac{\text{FN}}{\text{TP} + \text{FN}} \times 100\% \quad (34)$$

$$\text{false alarm rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} \times 100\% \quad (35)$$

$$\text{warning precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \quad (36)$$

$$\text{warning recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (37)$$

$$\text{warning latency} = \text{the first alarm time of the system} - \text{the actual time of entering the dangerous state} \quad (38)$$

where TP (true positive) denotes number of frames with real electric hazard and correct warning output; FP (false positive) denotes number of safe frames mistakenly triggered with alarms; FN (false negative) denotes number of hazardous frames without system warning; and TN (true negative) denotes number of safe frames with no predicted warning.

3.3. Training Settings

1. Hyperparameter Settings

Backbone: The feature extraction module adopts C2f_CT improved backbone of conventional YOLOv8-seg and it can output multi-scale feature maps at 1/2, 1/4, 1/8, 1/16, 1/32 downsampling ratios.

Optimizer: AdamW optimizer, base learning rate = 1×10^{-4} , cosine annealing decay over 120 training epochs, batch size = 8, early stopping triggered when validation mAP plateaus for 10 consecutive epochs.

Fixed Loss Weights: $\lambda_p = 1.0$, $\lambda_d = 2.0$, $\lambda_s = 0.1$, $\lambda_k = 0.5$. Equipment reprojection loss is assigned higher weight due to its dominant role in weak extrinsic estimation.

SIFT Matching and RANSAC Hyperparameters: Grayscale SIFT extraction, peak threshold = 0.03, Lowe ratio test threshold = 0.75 for outlier preliminary filtering; RANSAC maximum iterations = 1000, pixel reprojection error inlier threshold = 1.5 px.

Procrustes Alignment: Rigid transform only (rotation + translation, no scaling/shear during extrinsic pose solving; scale solved separately via scene priors).

Mask and Occlusion Gradient Masking: Any instance mask containing fewer than 100 valid foreground pixels is discarded from loss computation; partial occlusion handled via three-frame temporal mask moving average smoothing; all cross-view reprojection gradients are zeroed for worker non-rigid regions to avoid conflicting geometric supervision.

2. Hardware Environment: NVIDIA RTX 4090 GPU, Intel i9 CPU

3. Training Procedures

Forward Pass: Extract instance masks for workers and energized equipment via segmentation head.

Mask Validity Filtering: Pixels covered by heavy occlusion ($\text{IoU} < 0.2$) are masked out and excluded from all loss terms.

Depth Estimation: Produce an estimated multi-scale relative depth map based on the current frame from camera A.

Worker Branch: Predict inter-frame 6DoF local pose, use the assumed intrinsic parameters and estimated depth to compute photometric reconstruction loss only within valid human masks.

Equipment Branch: Extract SIFT keypoints from equipment masks, apply RANSAC outlier rejection, solve weak cross-camera extrinsic via scale-free Procrustes alignment, and calculate cross-view reprojection loss.

Regularization: Compute depth smoothness loss and intrinsic perturbation regularization.

Backward Propagation: Optimize all network weights and learnable intrinsic offsets via weighted total loss. For severe non-rigid human poses with >40% mask deformation, the photometric constraint weight is dynamically attenuated to avoid misleading depth optimization.

3.4. Experimental Results and Comparative Analysis

3.4.1. Image Preprocessing Results

The preprocessing pipeline included three steps: timestamp-based temporal alignment, image normalization, and noise filtering. Temporal alignment synchronized the dual-camera frames with an error lower than 10 ms. Image normalization resized and cropped images from the two cameras to a unified resolution of 640×480 . Noise filtering used a

5×5 Gaussian kernel to suppress video noise while preserving the edges of equipment and workers. To guarantee completely fair quantitative comparison, all models mentioned below are trained or fine-tuned using the above inputs. Figure 4 shows a representative comparison before and after preprocessing.



Figure 4. Comparison of image preprocessing effects: (a) original frame before preprocessing; (b) frame after normalization and noise filtering.

As shown in Table 3, the preprocessing pipeline improved the input quality and enhanced the subsequent tracking and distance estimation performance. Compared with the configuration without preprocessing, the complete preprocessing scheme reduced the relative distance error from 10.1% to 8.7%, increased MOTA from 76.3% to 92.3%, and reduced the model convergence epochs from 125 to 82. Image normalization mainly alleviated resolution and field-of-view inconsistencies among cameras, while noise filtering improved edge stability under low illumination and cluttered backgrounds.

Table 3. Data preprocessing ablation study.

Preprocessing Configuration	RDE (%)	MOTA (%)	Convergence Epochs
No preprocessing	10.1	76.3	125
Image normalization only	9.1	88.5	98
Noise filtering only	9.7	90.7	105
Complete preprocessing	8.7¹	92.3¹	82¹

¹ Bold data represents the optimal result.

3.4.2. Ablation Study of Tracking, Depth Estimation, and Scale Recovery

To avoid confusion between module-level and end-to-end speed, the tracking-module speed and the complete distance estimation speed were recorded separately. Figure 5 compares the frame rate and latency of different tracking algorithms. Under the lightweight input configuration, the proposed tracking module reached 40 FPS with 25 ms single-frame latency. Under the complete video configuration, the segmentation and tracking module achieved an average speed of 30.1 FPS. When depth estimation and scale recovery were further included, the end-to-end near-energized safety distance estimation system reached 22 FPS.

Among conventional tracking methods, DeepSORT and TaMOs achieved certain real-time performance but were less robust to severe occlusion and cross-camera identity association. NOOUGAT and DiffusionMOT introduced more complex semantic fusion mechanisms and therefore incurred greater computational costs, making them more suitable for offline or low-frame-rate analysis. The proposed method maintained high frame rate while improving identity stability in power operation scenes through mask-guided ReID and global cross-camera matching.

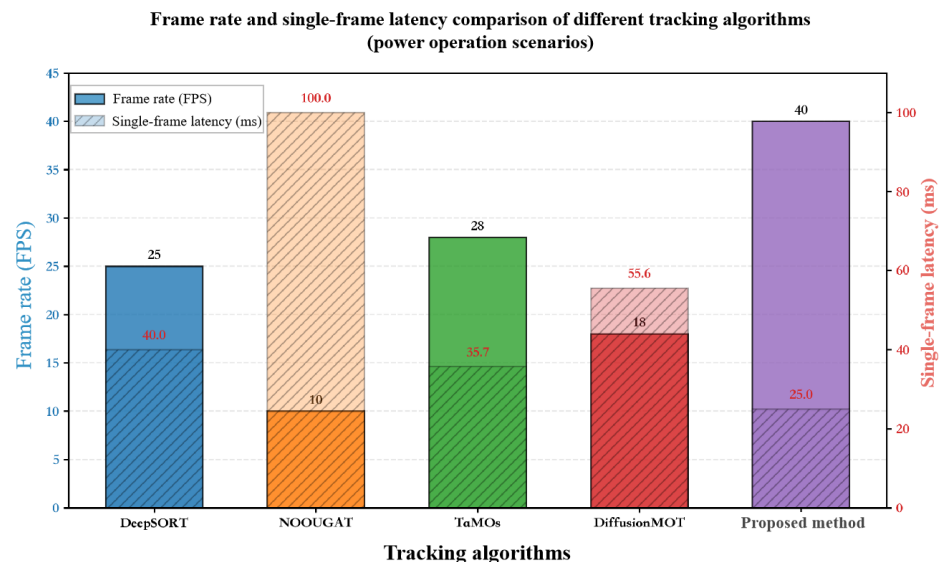


Figure 5. The relationship between trajectory tracking latency and frame rate.

The ablation results in Table 4 show that, starting from the baseline model, progressively adding C2f_CT, mask cropping, and Hungarian matching improved MOTA and mAP overall. It is worth noting that Rank-1 increased from 83.2% to 86.7% after introducing C2f_CT. This suggests that global structural enhancement improves detection and trajectory stability but may dilute fine-grained appearance cues for ReID. After mask cropping and global matching were added, Rank-1 recovered and finally increased to 94.2%. Therefore, the advantage of the proposed method comes from the joint effect of detection, segmentation, appearance embedding, and spatio-temporal matching rather than from a single component.

Table 4. Ablation and comparison results of cross-camera worker tracking.

Method	MOTA (%)	Rank-1 (%)	mAP (%)
Baseline	72.6	83.2	65.2
+C2f_CT	80.5	86.7	71.7
+Mask cropping	85.1	87.0	75.1
+Hungarian matching	88.7	87.2	78.9
Proposed method	92.3	94.2	82.0
DeepSORT	78.4	80.4	62.0
NOOUGAT	89.5	91.3	80.5
TaMOs	87.8	90.6	79.0
DiffusionMOT	91.2	91.9	81.8

Compared with advanced tracking methods, the proposed method improved MOTA, Rank-1, and mAP over DeepSORT by 13.9%, 13.8%, and 20.0%, respectively. Compared with NOOUGAT, the improvements were 2.8%, 2.9%, and 1.5%, respectively. Compared with DiffusionMOT, the gains were 1.1%, 2.3%, and 0.2%, respectively. Considering the real-time warning requirement of power operation sites, the proposed method achieves a favorable balance between accuracy and efficiency.

To evaluate the contribution of each component in the weakly calibrated depth estimation and scale recovery module, an ablation study was conducted by removing or modifying key constraints in the proposed framework. In Table 5, W-Photo, E-Reproj., and K-Reg. denote the worker region photometric loss, equipment region reprojection loss, and intrinsic parameter regularization, respectively. The entries “Yes” and “No” indicate whether the corresponding component is used in each configuration.

Table 5. Ablation results of weakly calibrated depth estimation and scale recovery.

Method	W-Photo	E-Reproj.	K-Reg.	Scale Cue	RDE (%)
w/o worker photometric loss	-	✓ ¹	✓	Equip. + human priors	10.4
w/o equipment reprojection loss	✓	-	✓	Equip. + human priors	13.9
w/o intrinsic regularization	✓	✓	-	Equip. + human priors	11.1
w/o scene dimensional priors	✓	✓	✓	None	N/A ²
Human-scale prior only	✓	✓	✓	human prior	12.5
Equipment size prior only	✓	✓	✓	Equipment prior	9.6
Proposed method	✓	✓	✓	Equip. + human priors	8.7

¹ “✓” represents whether the corresponding component is used in each configuration. ² “N/A” represents “Not Applicable”, which means no corresponding data available.

Table 5 reports the ablation results of the weakly calibrated depth estimation and scale recovery module. In the table, W-Photo, E-Reproj., and K-Reg. denote the worker region photometric loss, equipment region reprojection loss, and intrinsic parameter regularization, respectively. The results show that each component contributes to the final distance estimation performance.

Removing the worker region photometric loss increases the RDE from 8.7% to 10.4%, indicating that temporal appearance consistency within worker masks helps stabilize the relative depth of non-rigid worker regions. When the equipment region reprojection loss is removed, the RDE rises more significantly to 13.9%. This demonstrates that rigid equipment regions provide the most important cross-view geometric constraint for weak calibration, because stable equipment correspondences can better constrain the relative camera relation and equipment region depth. Removing intrinsic regularization also increases the RDE to 11.1%, suggesting that unconstrained intrinsic perturbation may introduce unstable depth-scale variation and degrade distance recovery.

The scale prior ablation further confirms the necessity of scene-dimensional priors. Without scene dimensional priors, the model can still estimate relative depth, but the metric scale is unavailable; therefore, RDE cannot be reliably computed. When only the human-scale prior is used, the RDE is 12.5%, which is higher than that obtained using only the equipment size prior. This is because human apparent height is easily affected by posture, occlusion, and viewpoint changes, whereas the dimensions of power equipment are generally more stable and standardized. Using only equipment size priors reduces the RDE to 9.6%, and the full model further reduces it to 8.7% by jointly using equipment and human-scale priors. These results verify that the proposed method benefits from the complementary effects of worker region temporal supervision, equipment region cross-view reprojection, intrinsic regularization, and scene-prior-guided metric scale recovery.

In order to quantify the influence of prior uncertainty to the distance estimation accuracy, a sensitivity analysis on equipment size errors, human height variation, and partially visible reference objects is conducted. The experimental results are shown in Table 6.

The sensitivity results show that the recovered metric distance changes monotonically with the perturbation of the selected dimensional prior. Moderate equipment size errors lead to bounded degradation, whereas large prior errors or heavily occluded reference segments increase the RDE and warning uncertainty. Human height perturbation causes larger fluctuation than equipment size perturbation because worker apparent height is more sensitive to posture and occlusion. These results indicate that equipment size priors should be preferred when reliable rigid references are visible, while human height priors should be used only as auxiliary Scale Cues.

Table 6. Sensitivity analysis of metric-scale recovery under prior uncertainty.

Setting	Prior Perturbation/Visibility Condition	RDE (%)	MAE (m)	Missed Alarm (%)	False Alarm (%)
Baseline	Original equipment + human priors	8.7	0.092	2.1	3.4
Equipment size prior error	−5%/+5%	9.1	0.096	2.4	3.5
Equipment size prior error	−10%/+10%	9.7	0.104	2.9	4.1
Equipment size prior error	−15%/+15%	11.0	0.115	3.6	5.0
Human height prior error	±5 cm	10.1	0.106	3.0	4.5
Human height prior error	±10 cm	12.5	0.130	4.7	6.1
Reference visibility	Visible ratio > 80%	7.3	0.076	1.2	2.0
Reference visibility	Visible ratio 50–80%	9.5	0.099	2.7	3.8
Reference visibility	Visible ratio < 50%	13.6	0.142	5.3	6.9

3.4.3. Safety Distance Estimation Results

To further evaluate the distance estimation capability of the proposed framework, representative public or reproducible baselines were compared under the same power operation test setting.

1. Instance mask sharing

The proposed YOLOv8-seg segmentation output is reused for all baseline depth branches to eliminate detection performance discrepancy.

2. Metric scale unification

Metric scale unification: Monocular baselines (Monodepth2, ManyDepth) adopt our equipment–human joint scale recovery pipeline to convert relative depth to metric distance.

3. Training schedule

Training schedule: Monodepth2, ManyDepth and PSMNet are fine-tuned on our power operation dataset; calibrated binocular and PnP geometry are traditional geometry algorithms without training process.

Monocular geometric ranging was used as a traditional single-image baseline, where metric distance is inferred from manually defined scene priors or known target dimensions. Monodepth2 and ManyDepth were selected as representative self-supervised monocular video depth estimation methods [13,15]. Monodepth2 mainly learns depth from temporal reconstruction in a single-camera video, whereas ManyDepth further exploits multi-frame information to improve temporal depth consistency. PSMNet was included as a learning-based stereo-matching baseline under rectified stereo input [11]. In addition, a PnP-based geometric baseline with known equipment dimensions was implemented to represent traditional weakly calibrated geometry-driven ranging. For monocular depth estimation methods that output relative depth rather than metric depth, the same scene-level scale alignment protocol was applied before computing RDE. In Table 7, “Scale Cue” denotes the source used to recover metric distance from visual observations, and “Equip.” denotes equipment.

As shown in Table 7, the RDE of calibrated binocular method achieves 4.2%, because it relies on accurately calibrated intrinsic and extrinsic parameters as well as a fixed stereo baseline. However, this method requires strict installation and calibration conditions, making it less flexible for temporary or non-standard power operation sites. Monocular geometric ranging achieves a frame rate of 25 FPS, but its RDE reaches 18.3% because single-image geometry provides insufficient depth constraints and depends strongly on manually defined scene priors. Monodepth2 and ManyDepth reduce the RDE to 14.6% and 12.8%, respectively, indicating that learning-based depth estimation from monocular videos is more effective than purely geometric single-image ranging. Nevertheless, their depth predictions remain scale-ambiguous and require scale alignment before metric distance

calculation, which limits their robustness when equipment layouts, occlusions, and worker poses vary across scenes.

Table 7. Comparison of relative distance errors in power operation scenes.

Method	Input	Calibration Requirement	Scale Cue	RDE (%)	Frame Rate	Deployment Adaptability
Calibrated binocular	Stereo images	Required	Stereo baseline	4.2	18	Medium
Monocular geometric ranging	Single image	No	Scene prior	18.3	25	Low
Monodepth2	Monocular video	No	Scale alignment	14.6	17	Medium
ManyDepth	Monocular video	No	Scale alignment	12.8	11	Medium
PSMNet	Rectified stereo images	Required	Stereo disparity	8.1	2	Medium
PnP + known-size geometry	Single/multi-view images	Weak	Equipment size	13.3	24	Medium
Mechanical LiDAR	LiDAR point cloud + images	Required (LiDAR–camera joint calibration)	Native metric point cloud	2.1 ¹	35	Low
Millimeter Wave Radar	Radar point cloud + images	Required (radar–camera extrinsic calibration)	Native metric ranging	5.8	60 ¹	Medium
CalibRefine	Multi-modal (LiDAR/Radar + multi-camera images)	Weak (targetless online calibration)	Native metric sensor benchmark	3.5	16	Medium
Proposed method	Multi-camera video	Weak	Equip. + human priors	8.7	22	High

¹ Bold data represents the optimal result.

PSMNet obtains an RDE of 8.1%, which is slightly lower than that of the proposed method, because stereo disparity provides strong geometric constraints when rectified stereo images are available. However, its frame rate is only 2 FPS owing to the computational cost of cost-volume construction and 3D convolutional regularization, which makes it less suitable for real-time near-energized warning. The PnP + known-size geometry baseline achieves 24 FPS, but its RDE is 13.3% because its accuracy depends on stable keypoint localization, clear equipment boundaries, and reliable known-size objects.

The mechanical LiDAR method achieves the lowest RDE of 2.1%, benefiting from native metric measurement and stable environmental adaptability. However, such scheme requires expensive hardware and mandatory sensor joint calibration, leading to low or medium deployment adaptability. Millimeter Wave Radar achieves a relatively high frame rate of 60 FPS because the radar hardware has a high refresh rate, but its RDE reaches 5.8%, slightly lower than the mechanical LiDAR. It is easy to install for its small size, but still requires complete extrinsic registration with the camera. There is no need to manually set up calibration targets, relying on inherent features of the scene or prior industrial standard dimensions to complete parameter optimization online. CalibRefine relies on multimodal fusion to achieve high precision (3.5% RDE). It uses inherent features of the scene or prior industrial standard dimensions to complete parameter optimization online, saving the cost of regular offline calibration, but still requiring additional active sensor hardware, which limits its large-scale promotion.

In contrast, the proposed method achieves an RDE of 8.7% while maintaining 22 FPS. Although its ranging error is higher than those of calibrated binocular vision and PSMNet, it substantially reduces the dependence on offline stereo calibration, rectified stereo input, and fixed binocular deployment. Compared with monocular depth estimation and PnP-

based baselines, the proposed method provides a better balance among metric accuracy, real-time performance, and deployment adaptability by jointly using cross-camera weak geometry, region-wise depth constraints, and equipment–human-scale priors.

To comprehensively verify the generalization performance of the proposed weakly calibrated multi-camera distance estimation framework, we split the test dataset into three independent power operation scenarios according to field operation types, and report separated quantitative metrics for each category. The experimental results are shown in Table 8.

Table 8. Quantitative evaluation results of the proposed method across three typical power operation scenarios.

Scenario	No. of Test Pairs	RDE (%)	MAE (m)	RMSE (m)	95% CI (RDE, %)	MOTA (%)	Rank-1 (%)	mAP (%)	FPS
110 kV substation	920	8.3	0.086	0.119	[7.97, 8.63]	93.1	94.8	83.5	22
Transmission line	1350	9.1	0.098	0.134	[8.75, 9.45]	91.7	93.6	80.9	22
Maintenance site	1730	9.7	0.107	0.148	[9.31, 10.09]	91.2	93.1	80.2	22
Overall dataset	4000	8.7	0.092	0.127	[8.36, 9.04]	92.3	94.2	82.0	22

To explicitly clarify the robustness limits and failure modes of the proposed method, we partition all test samples into five typical challenging subsets according to scene space, occlusion degree, camera view angle and risk distance, then count unified RDE, and MAE for each hard case. The experimental results are shown in Table 9.

Table 9. Distance estimation error analysis under challenging power operation conditions.

Case	Grouping Criterion	RDE (%)	MAE (m)	Interpretation	Case
Near-distance warning cases	d_gt close to safety threshold	9.0	0.095	Verify the performance near the danger boundary	Near-distance warning cases
Heavy occlusion	Worker/equipment visible ratio below threshold	13.6	0.142	Verify the impact of occlusion	Heavy occlusion
Indoor scenes	Distribution station	9.7	0.107	Verify the effects of low light and narrow space	Indoor scenes
Outdoor scenes	Substation + transmission line	8.6	0.090	Verify strong light and long-distance effects	Outdoor scenes
Large viewpoint difference	Camera viewpoint difference larger than preset angle	11.0	0.115	Verify cross-perspective robustness	Large viewpoint difference

The per-scenario results show the generalization behavior of the proposed method under different power operation conditions. Outdoor substation and transmission line scenes contain larger depth ranges and stronger illumination variation, while distribution station scenes contain narrower spaces, dense cabinets, and more frequent partial occlusion. Reporting the results separately allows us to identify the conditions under which the proposed weakly calibrated distance estimation remains reliable and the cases where the error increases.

Since the core engineering goal of this system is reliable live operation risk alerting, we further quantify warning indicators including missed alarm rate, false alarm rate, precision, recall and latency based on the full test distance sequence. Four post-processing pipelines are compared to analyze the gain of temporal smoothing and uncertainty filtering on alarm robustness. The experimental results are shown in Table 10.

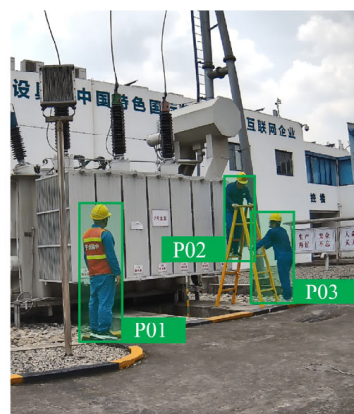
The warning-level evaluation further demonstrates the practical value of the proposed framework. Temporal smoothing reduces false alarms caused by short-term depth fluctuation, while the uncertainty threshold suppresses unreliable warnings under heavy occlusion or unstable scale prior conditions. Although smoothing may slightly increase the warning latency, the latency remains within the real-time warning requirement of the operation site monitoring system.

Table 10. Safety warning performance under different post-processing strategies.

Configuration	Missed Alarm (%)	False Alarm (%)	Warning Precision (%)	Warning Recall (%)	Warning Latency (s)
Raw distance output	2.1	3.4	96.6	97.9	0.00
+3-frame temporal smoothing	1.5	2.2	97.8	98.5	0.10
+5-frame temporal smoothing	1.2	1.7	98.3	98.8	0.17
+5 frame smoothing uncertainty threshold	0.9	1.3	98.7	99.1	0.17

Overall, the proposed method performs well at both cross-camera tracking and safety distance estimation levels. Preprocessing and mask constraints improve the quality of target region input. The C2f_CT module enhances representation capability in complex backgrounds. Global cross-camera matching improves identity continuity, and the region-wise self-supervised depth constraints provide a geometric basis for distance recovery under weak calibration. Compared with calibrated binocular vision, the proposed method sacrifices part of the ranging accuracy but substantially reduces deployment cost. Compared with monocular vision, it markedly decreases the distance error caused by scale ambiguity, making it more suitable for low-cost real-time warning in power operation sites.

Based on the cross-camera tracking results shown in Figure 6, the closest three-dimensional distance between a worker and energized equipment, such as insulators, conductors, or switch cabinets, was further calculated. Figure 7 shows representative distance estimation results. The first row presents dual-view field images, the second row shows relative depth or disparity information, and the third row shows worker distance recognition results. The results demonstrate that the proposed method can continuously track workers and dynamically estimate near-energized distances in complex power operation environments, thereby providing visual perception support for real-time safety warning.



(a1) Tracking result from Camera A in substation scenario



(a2) Tracking result from Camera B in substation scenario

Figure 6. Cont.

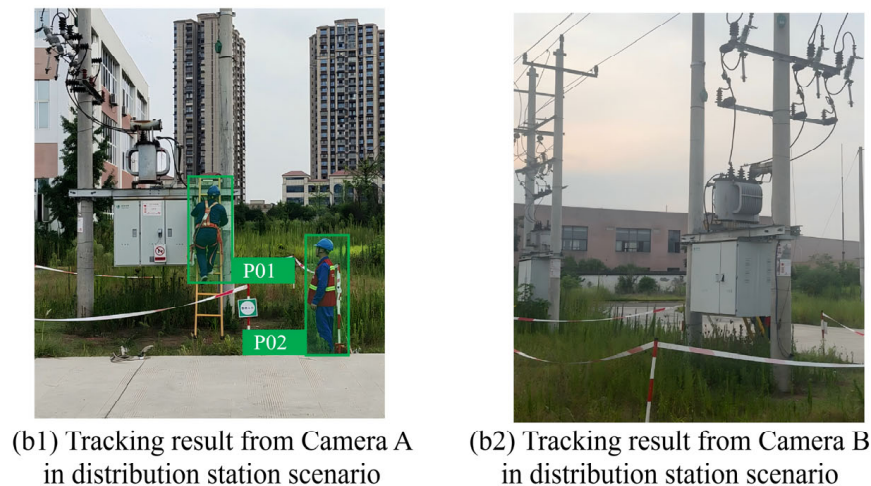


Figure 6. Example of cross-camera worker tracking in power operation scenes. The tracking results from different views provide continuous worker inputs for subsequent minimum distance calculation.

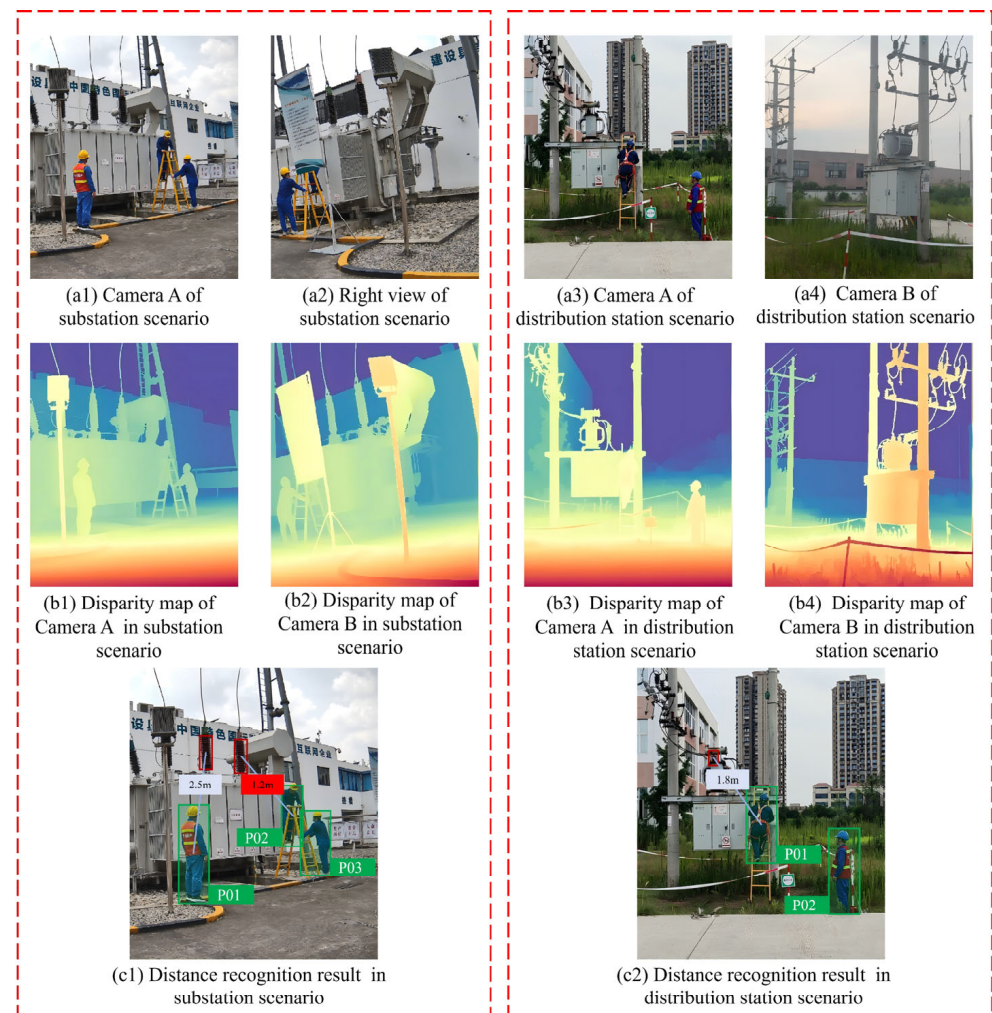


Figure 7. Example of safety distance estimation for workers in power operation scenarios.

4. Discussion

Although the proposed method lowers the deployment threshold of ordinary cameras in power operation sites, several limitations remain. First, the method relies on stable texture or structural features in the equipment region. When energized equipment exhibits severe reflection, weak texture, or heavy occlusion, cross-view feature matching may

degrade, which in turn affects weak extrinsic estimation and depth optimization. Second, metric-scale recovery still depends on in-scene dimensional priors, such as equipment standard sizes or human height ranges. The sensitivity and case-based analyses also reveal the main failure boundaries of the proposed method. The recovered metric scale becomes less reliable when the selected reference equipment is severely occluded, when its standard dimension is ambiguous, or when the available feature correspondences are unstable. Human height priors may also introduce errors under crouching, bending, climbing, or partial body occlusion. Therefore, the proposed method is most suitable for scenes where at least one rigid equipment reference with a reliable standard dimension is visible. In highly occluded scenes or scenes without valid dimensional priors, the system should report increased uncertainty and adopt a conservative warning strategy rather than producing an over-confident distance estimate. If the selected object is partially visible or its standard size is uncertain, scale errors may propagate to safety distance estimation. Third, the worker region self-supervised constraint uses local appearance consistency rather than strict rigid geometric motion. Non-rigid posture changes, fast movement, and severe occlusion may therefore reduce depth reliability. Fourth, although the experiments cover substations, transmission line maintenance sites, and distribution stations, further validation is required under more voltage levels, nighttime illumination, rain, fog, strong backlight, and extreme occlusion conditions.

The camera installation angle also affects the reliability of weakly calibrated depth estimation. When the viewing angle difference between cameras is too small, the parallax is insufficient and the cross-view constraint contributes limited depth information. When the viewing angle difference is too large, the overlapping field of view decreases and stable feature correspondences become difficult to obtain. Therefore, the proposed method is most suitable for scenarios where ordinary cameras have a moderate overlap, visible rigid equipment structures, and at least one reliable dimensional prior. When these conditions are not satisfied, the system should rely more on temporal trends and conservative warning thresholds rather than a single-frame distance estimate.

From an engineering perspective, near-energized safety warning is a high-risk task, and a single-frame distance estimate should not be used as the sole decision basis. In practical deployment, consecutive frame distance trends, estimation uncertainty, and voltage-level safety thresholds should be jointly considered to form conservative warning rules. When model confidence is low or multi-view results are inconsistent, the system should preferentially output a risk prompt rather than silently allow the operation. Such conservative strategies can reduce missed warnings and improve the acceptability of the system in real power operation safety management.

5. Conclusions

This paper presented a weakly calibrated multi-camera vision framework for near-energized safety distance estimation in power operation scenarios. The framework addresses three practical challenges commonly encountered in field deployment: the absence of precise stereo calibration, cross-camera worker movement, and metric-scale ambiguity in ordinary camera perception. By combining region-decoupled self-supervised depth learning, scene-prior-guided scale recovery, and cross-camera worker association, the proposed method establishes an end-to-end pipeline from visual perception to safety distance warning.

First, a scene element segmentation and cross-camera worker tracking framework was developed for complex power operation scenes. The C2f_CT module enhanced both local detail and global structure representation in YOLOv8-seg, while mask-guided ReID features and Hungarian matching improved cross-camera identity association. The method achieved 92.3% MOTA, 94.2% Rank-1 accuracy, and 82.0% mAP on the experimental data.

Second, a region-wise self-supervised depth estimation method was proposed. Different constraints were designed for worker and equipment regions according to their motion characteristics: adjacent frame photometric reconstruction for worker regions and cross-view reprojection for rigid equipment regions. Depth smoothness and intrinsic parameter regularization further improved relative depth stability under weak calibration. Third, a safety distance recovery strategy based on scene-inherent scale priors was introduced. Without placing additional calibration targets on-site, the method used standard equipment dimensions and human-scale priors to convert relative depth into metric three-dimensional distance. In three representative power operation scenarios, the proposed method achieved 8.7% relative distance error and 22 FPS end-to-end processing speed, 98.7% warning precision, 0.9% missed alarm rate and 1.3% false alarm rate with five-frame smoothing and uncertainty threshold.

Future work will expand the dataset to cover more voltage levels, weather conditions, nighttime environments, and worker postures. Distance estimation uncertainty modeling and conservative safety threshold decision mechanisms will also be introduced. In addition, the integration of the proposed method with field video surveillance platforms, work ticket systems, and edge-computing devices will be investigated to support practical deployment.

Author Contributions: Conceptualization, W.W. and Y.Z.; methodology, Z.C. and C.L.; data curation, G.P. and C.L.; validation, Y.Z., Z.C. and G.P.; formal analysis, W.W. and Z.C.; writing—original draft preparation, W.W. and Y.Z.; writing—review and editing, C.L. and G.P.; supervision, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by State Grid Corporation of China Headquarters Project, which is ‘Research and Application of Key Technologies for Explainable Multi-object Recognition of On-site Operational Violations Based on Image-Text Models’, under Grant 5700-202426249A-1-1-ZN.

Data Availability Statement: The power domain dataset used in this study is subject to the data security policies of State Grid Corporation of China. Requests for access may be directed to the corresponding author and will be considered after a formal data sharing agreement has been completed.

Conflicts of Interest: Authors Wei Wei, Yi Zhang and Guozheng Peng were employed by the company China Electric Power Research Institute. Author Zhengwei Chang was employed by the company State Grid Sichuan Electric Power Research Institute. Author Chao Liu was employed by the company State Grid Corporation of China. The authors declare that this study received funding from State Grid Corporation of China. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

References

1. Ma, J.; Ren, B.; Chen, L.; Ma, H.; Zhu, S.; Chen, T. High-precision multi-person pose estimation method for substation near-electricity operations. *Zhejiang Electr. Power* **2024**, *43*, 97–106. [[CrossRef](#)]
2. Zhang, D.; Xu, H.; Liu, J.; Yang, C.; Huang, X.; Zhang, Z.; Jiang, X. Research on the non-contact pollution monitoring method of composite insulator based on space electric field. *Energies* **2021**, *14*, 2116. [[CrossRef](#)]
3. He, M.; Qin, L.; Zhao, F.; Yu, J.; Liu, H.; Wang, Q.; Xu, X.; Liu, K. Intelligent detection algorithm for safety risk management and control in power system on-site operations. *High Volt. Eng.* **2023**, *49*, 2442–2457. [[CrossRef](#)]
4. Ma, F.; Wang, B.; Dong, X.; Dong, X.; Yao, L.; Wang, H. Safety image interpretation in power industry: Basic concepts and technical framework. *Proc. CSEE* **2022**, *42*, 458–475. [[CrossRef](#)]
5. Li, J.; Shuang, F.; Huang, J.; Wang, T.; Hu, S.; Hu, J.; Zheng, H. Safe Distance Monitoring of Live Equipment Based upon Instance Segmentation and Pseudo-LiDAR. *IEEE Trans. Power Deliv.* **2023**, *38*, 2953–2964. [[CrossRef](#)]
6. Lee, S.H.; Civera, J. Robust uncertainty-aware multiview triangulation. *arXiv* **2020**, arXiv:2008.01258.
7. Wang, M.; Wang, B.; Fan, X.; Yin, T.; Ma, H.; Luo, P. Safe Distance Monitoring for Substation Near-Current Operations via Image-LiDAR Cross-Modal Self-Registration. *Electronics* **2026**, *15*, 2321. [[CrossRef](#)]

8. Jiang, J.; Li, Z.; Liu, X. Survey of monocular depth estimation methods based on deep learning. *Chin. J. Comput.* **2022**, *45*, 1276–1307. [\[CrossRef\]](#)
9. Eigen, D.; Puhersch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In Proceedings of the 28th Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1–9.
10. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. GC-Net: End-to-End Learning of Geometry and Context for Deep Stereo Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6681–6690.
11. Chang, J.-R.; Chen, Y.-S. Pyramid Stereo Matching Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5410–5418. [\[CrossRef\]](#)
12. Guo, X.; Yang, K.; Yang, W.; Wang, X.; Li, H. GWCNet: Group-wise Correlation Stereo Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3268–3277.
13. Godard, C.; Mac Aodha, O.; Brostow, G.J. Digging into Self-Supervised Monocular Depth Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3828–3838.
14. Zhang, N.; Nex, F.; Vosselman, G.; Kerle, N. Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 18537–18546.
15. Guizilini, V.; Ambrus, R.; Chen, D.; Zakharov, S.; Gaidon, A. Multi-Frame Self-Supervised Depth with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 160–170.
16. Woo, S.; Lee, W.; Kim, W.J.; Lee, D.; Lee, S. ProDepth: Boosting Self-supervised Multi-frame Monocular Depth with Probabilistic Fusion. In *Computer Vision—ECCV 2024*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2025; Volume 15061, pp. 201–217. [\[CrossRef\]](#)
17. Zou, Y.; Ding, Y.; Qiu, X.; Wang, H.; Zhang, H. M2Depth: Self-Supervised Two-Frame Multi-Camera Metric Depth Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
18. Kanai, T.; Vasiljevic, I.; Guizilini, V.; Gaidon, A.; Ambrus, R. Robust Self-Supervised Extrinsic Self-Calibration. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023. [\[CrossRef\]](#)
19. Yin, Z.; Chen, Y.; She, P.; Liu, C.; He, X.; Lv, S. Modeling the measurement precision of a multi-camera system. *Opt. Lett.* **2025**, *50*, 6489–6492. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Shimizu, T.; Sato, T. Development of Non-Contact Detection of Charged Conductor Using FMCW Millimeter-Wave Radar and Stereo Vision Fusion. *J. Power Energy Eng.* **2026**, *14*, 89–102.
21. Ma, F.; Wang, J.; Jia, R.; Wang, B.; Ma, H. Research on intelligent early warning method for near-electricity safety distance of large construction machinery based on binocular stereo perception and safety region segmentation. *Proc. CSEE* **2025**, *45*, 2554–2567. [\[CrossRef\]](#)
22. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1302–1310.
23. Wang, L.; Wang, B.; Zhang, J.; Ma, H.; Luo, P.; Yin, T. An Intelligent Detection Method for Approach Distances of Large-Scale Construction Equipment in Substations. *Electronics* **2023**, *12*, 3510. [\[CrossRef\]](#)
24. Wang, L.; Wang, B.; Ma, F.; Dong, X.; Yao, L. Research on near-electricity safety distance detection method for substation personnel based on binocular stereo matching and scene element recognition. *Power Syst. Technol.* **2023**, *47*, 1010–1021. [\[CrossRef\]](#)
25. Min, Y.; Wang, X.; Tian, W.; Wang, G. Safety distance measurement method based on improved semantic segmentation network and stereo vision. *Comput. Eng. Appl.* **2025**; online first.
26. Li, X.; Fan, B.; Tian, J.; Fan, H. GAFusion: Adaptive Fusing LiDAR and Camera with Multiple Guidance for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 12456–12465.
27. Wang, T.; Zhang, L.; Liu, Y. Reliability-Driven LiDAR-Camera Fusion for Robust 3D Object Detection. *arXiv* **2025**, arXiv:2502.01856.
28. Cheng, L.; Guo, L.; Zhang, T.; Bang, T.; Harris, A.; Hajj, M.; Cao, S. CalibRefine: Deep Learning-Based Online Automatic Targetless LiDAR–Camera Calibration with Iterative and Attention-Driven Post-Refinement. *IEEE Trans. Instrum. Meas.* **2025**, *75*, 9500418. [\[CrossRef\]](#)
29. Chharia, A.; Ren, T.; Furuhashi, T.; Shimada, K. Safe-Construct: Redefining Construction Safety Violation Recognition as 3D Multi-View Engagement Task. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 11–15 June 2025; pp. 5811–5820. [\[CrossRef\]](#)
30. Jocher, G.; Chaurasia, A.; Qiu, J. What Is YOLOv8: An In-Depth Exploration of the Next-Generation Object Detector. Ultralytics Technical Report. 2023. Available online: <https://docs.ultralytics.com/models/yolov8#supported-tasks-and-modes> (accessed on 10 January 2023).

31. Barath, D.; Mishkin, D.; Matas, J. RANSAC Back to SOTA: A Two-stage Consensus Filtering for Real-time 3D Registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10446–10455.
32. Xia, Z.; Xu, C.; Alahi, A. Fine-Grained Cross-View Localization via Local Feature Matching and Monocular Depth Priors. *arXiv* **2025**, arXiv:2509.09792.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.