CALIBFREE: SELF-SUPERVISED FEATURE DISENTANGLEMENT FOR CALIBRATION-FREE MULTI-CAMERA MULTI-OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-camera multi-object tracking (MCMOT) faces significant challenges in maintaining consistent object identities across varying camera perspectives, particularly when precise calibration and extensive annotations are required. In this paper, we present CALIBFREE, a self-supervised representation learning framework that does not need any calibration or manual labeling for the MCMOT task. By disentangling view-agnostic and view-specific features through single-view distillation and cross-view reconstruction, our method adapts to complex, dynamic scenarios with minimal overhead. Experiments on the MMP-MvMHAT dataset show a 3% improvement in overall accuracy and a 7. 5% increase in the average F1 score over state-of-the-art approaches, confirming the effectiveness of our calibration-free design. Moreover, on the more diverse MvMHAT dataset, our approach demonstrates superior over-time tracking and strong cross-view performance, highlighting its adaptability to a wide range of camera configurations.

1 INTRODUCTION

Multiple Object Tracking (MOT) is an essential problem in computer vision, aiming to identify and track multiple objects within video streams. While single-camera tracking has been extensively studied Cao et al. (2023); Wojke et al. (2017); Cai et al. (2022); Meinhardt et al. (2022); Zeng et al. (2022); Zhang et al. (2023), the importance of Multi-Camera Multi-Object Tracking (MCMOT) continues to grow with the rising applications of multi-camera systems in surveillance, smart cities, and autonomous vehicles Gilbert & Bowden (2006); He et al. (2020); You & Jiang (2020); Cheng et al. (2023); Zhang et al. (2022a); Gu et al. (2023). MCMOT aims to maintain consistent object identities across multiple camera views, addressing inherent challenges such as viewpoint variation, occlusions, and synchronization issues, as illustrated in Figure 1. By integrating diverse viewpoints, MCMOT can offer improved tracking robustness, enhanced scene understanding, and fewer blind spots compared to single-camera methods Han et al. (2020; 2021).

Despite these advantages, achieving effective MCMOT remains challenging He et al. (2020); You & Jiang (2020); Chen et al. (2014). A primary difficulty arises from significant variations in object appearance and motion across different camera views, making reliable object re-identification (ReID) nontrivial. Moreover, many MCMOT methods Ristani et al. (2016b); Maksai et al. (2017); Tesfaye et al. (2019); He et al. (2020); You & Jiang (2020); Cheng et al. (2023) rely on calibrated camera setups or large-scale annotations. Even minor camera shifts—such as relocating a camera or changing its angle—can break calibration, causing immediate performance declines until the system is recalibrated and annotated data are recollected. Similarly, transitioning to a new scene often necessitates gathering a fresh dataset, performing calibration, and retraining the model. As camera networks expand or reconfigure, the associated computational overhead grows, making frequent recalibration and reannotation both costly and impractical in real-world applications.

Main Results: To address these limitations, we propose a self-supervised learning framework specifically designed for multi-camera setups with overlapping fields of view. Our method avoids explicit calibration and reduces the need for annotations by leveraging data-driven representation learning. In particular, we present a disentangled feature learning strategy that separates view-agnostic and view-specific features through single-view distillation and cross-view reconstruction.

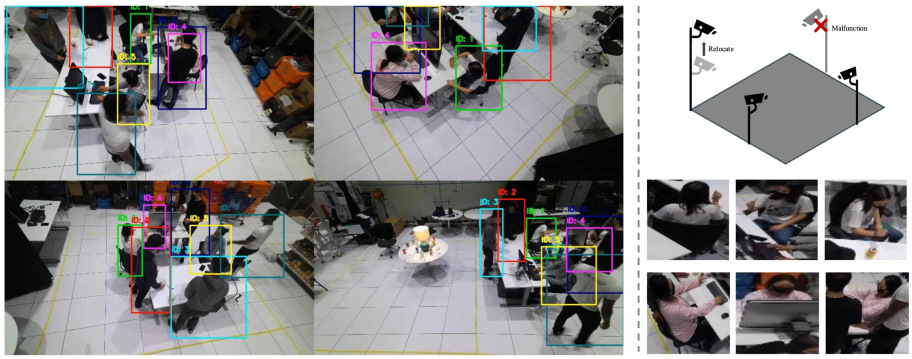

Figure 1: **Multi-Camera Multi-Object Tracking (MCMOT) setup.** *Left:* Multi-view scenes with individuals tracked across overlapping camera views, each assigned a unique color-coded bounding box. *Top right:* Flexible camera configurations illustrating variable camera numbers and placements across scenarios, or due to factors like malfunction or relocation. *Bottom right:* Examples of appearance variations for individuals across different viewpoints, highlighting the challenge of maintaining consistent identity association in multi-view tracking.

This approach mitigates viewpoint-based discrepancies and improves cross-view tracking without costly manual calibration or any labeling. Our contributions can be summarized as follows:

1. We propose a self-supervised representation learning framework for MCMOT, effectively reducing reliance on both manual annotations and camera calibration.
2. We introduce a disentangled feature learning strategy via single-view distillation and cross-view reconstruction, enhancing robustness against viewpoint variations.
3. We empirically validate our method on two challenging MCMOT datasets: (1) MMP-MvMHAT, featuring densely placed indoor cameras that capture crowded, occluded environments, where our method surpasses state-of-the-art baselines by 3% in overall accuracy and 7.5% in average F1 score. (2) MvMHAT, containing both indoor and outdoor scenes with sparser camera coverage and reduced overlapping fields of view, where our approach likewise demonstrates superior over-time tracking and strong cross-view performance, underscoring its adaptability to diverse real-world scenarios.

2 Related Work

2.1 Single-Camera Multi-Object Tracking

Single-camera multi-object tracking (MOT) has been extensively studied, with the tracking-by-detection paradigm being the most widely adopted Cao et al. (2023); Leal-Taixé et al. (2016); Schulter et al. (2017); Wojke et al. (2017). In this framework, object detectors Duan et al. (2019); Girshick (2015); Ge et al. (2021) identify objects in each frame, and temporal associations are made using methods like the Kalman Filter Welch (1995) and the Hungarian Matching algorithm Kuhn (1955). Deep appearance features further improve association accuracy Chu & Ling (2019); Xu et al. (2019; 2020). End-to-end approaches such as MOTR Zeng et al. (2022), MOTRv2 Zhang et al. (2023), and TrackFormer Meinhardt et al. (2022) leverage query-based object detection to perform long-term tracking without manual association rules Carion et al. (2020). TransTrack Sun et al. (2020) and P3Aformer Zhao et al. (2022) improve efficiency using location-based cost matrices. However, single-camera MOT struggles with occlusions and complex interactions due to limited viewpoints, motivating research in multi-camera multi-object tracking (MCMOT).

2.2 Multi-Camera Multi-Object Tracking

Multi-camera multi-object tracking (MCMOT) has gained growing attention for its complexity and broad applications. Existing methods fall into three main categories: distributed, global, and end-to-end. *Distributed methods* perform tracking independently within each camera, followed by cross-view association Gilbert & Bowden (2006); Prosser et al. (2008); Cai & Medioni (2014); Chen et al.

(2014). Techniques such as hierarchical clustering Murtagh & Contreras (2012) and non-negative matrix factorization (NMF) Wang & Zhang (2012) are used to merge intra-camera tracklets, though they often assume ideal conditions not met in dynamic environments. *Global methods* detect individuals across views and associate them directly to build tracklets Ristani et al. (2016b); Maksai et al. (2017); Tesfaye et al. (2019). TRACTA He et al. (2020) and DMCT You & Jiang (2020) utilize perspective models and occupancy heatmaps, while ReST Cheng et al. (2023) employs a reconfigurable graph for robust association. *End-to-end methods* like MUTR3D Zhang et al. (2022a), PF-Track Pang et al. (2023), and ViP3D Gu et al. (2023) are designed for 3D tracking tasks such as autonomous driving. MCTR Niculescu-Mizil et al. (2024) proposes a calibration-free framework using track embeddings, but it depends on labeled data and fixed camera setups, limiting flexibility. Our work focuses on overlapping-camera scenarios, where shared fields of view offer appearance and geometric constraints for trajectory linking. In contrast, non-overlapping setups Javed et al. (2005); Tesfaye et al. (2017); Chilgunde et al. (2004) face challenges such as time delays and lack of spatial correspondence.

2.3 Self-Supervised Learning

Self-supervised learning (SSL) has become a powerful paradigm in computer vision Doersch et al. (2015); Noroozi & Favaro (2016) and multi-modal tasks Akbari et al. (2021); Wang et al. (2021), enabling robust representation learning without labeled data. Pretext tasks like context prediction Doersch et al. (2015); Pathak et al. (2016), jigsaw puzzles Noroozi & Favaro (2016); Kim et al. (2018b), and colorization Larsson et al. (2017) have shown effectiveness for image-level learning. For videos, pace prediction Wang et al. (2020) and space-time cube puzzles Kim et al. (2018a) help capture temporal dynamics. More recent techniques such as contrastive learning Chen et al. (2020); He et al. (2019); van den Oord et al. (2018) and masked autoencoding He et al. (2021); Huang et al. (2022) are effective across both image and video domains. In multi-object tracking, self-supervised methods use spatial-temporal consistency to reinforce object identity. Strategies include cross-input consistency Bastani et al. (2021), cycle-consistency Yin et al. (2023), and path-consistency Lu et al. (2024). In MCMOT, recent approaches like MvMHAT Feng et al. (2024) and MvMHAT++ Gan et al. (2021) leverage consistency-based tasks such as symmetric-consistency (SymC) and transitive-consistency (TrsC), though they rely on CNN-based features. In contrast, our method adopts masked autoencoding He et al. (2021), which is more compatible with transformer architectures and enables richer, more adaptable representation learning in complex MCMOT settings.

3 Method

In this section, we present the details of our proposed approach, **CALIBFREE**. We begin by formulating the problem, followed by a description of our algorithm, and conclude with how the generated features are used during inference.

3.1 Problem Formulation

Multi-camera multi-object tracking (MCMOT) aims to track all subjects across synchronized video streams from V cameras and associate identities across views. This can be formulated as a spatio-temporal association problem with two objectives:

- **Intra-camera tracking:** Given detections $D_t^v = \{D_i^v \mid i = 1, 2, \ldots, N_t^v\}$ at frame t in view v, associate them over time to form tracklets τ_t^v, as in single-camera MOT.
- **Cross-view matching:** Match detections $\bar{D}_t = \{\bar{D}_t^1, \bar{D}_t^2, \ldots, \bar{D}_t^V\}$ across views at time t that belong to the same subject.

Like single-camera methods (e.g., DeepSORT Wojke et al. (2017), ByteTrack Zhang et al. (2022b)), MCMOT relies on robust feature representations to ensure reliable association. These features must remain consistent across time and camera viewpoints while being discriminative enough to separate different identities.

Given all detections at time t, $D_t = \{D_{t,1}^1, \ldots, D_{t,N_t^V}^V\}$, the goal is to extract two types of features for each detection $D_{t,i}^v$:

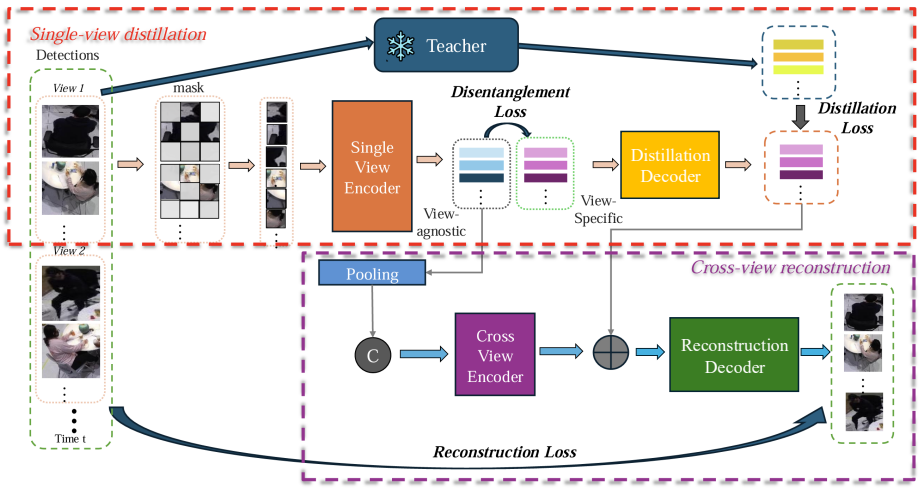

Figure 2: **Overview of CALIBFREE.** The method includes single-view distillation, feature disentanglement, and cross-view reconstruction. In single-view distillation (red box), masked detections are encoded, and feature quality is supervised by a teacher model using distillation loss. The disentanglement module splits features into view-agnostic and view-specific parts. For cross-view reconstruction (purple box), pooled view-agnostic features are processed to reconstruct masked patches across views, optimized with reconstruction loss.

- **View-agnostic features** (f_a): Capture identity-preserving cues (e.g., silhouette, body shape, pose) for cross-view matching.
- **View-specific features** (f_s): Encode appearance-specific details (e.g., clothing, texture) useful for temporal tracking within a view.

These features support both within-view and cross-view association, enabling robust identity continuity across space and time in uncalibrated multi-camera environments.

3.2 CALIBFREE

Masked autoencoders He et al. (2021) have proven effective in learning visual semantics, generating high-quality features from images. CALIBFREE builds on this framework to improve representations for detections of different persons, see Figure 2. Unlike traditional methods that reconstruct from partial observations within the same image and view, CALIBFREE introduces a cross-view reconstruction task, enabling reconstruction from observations across different views using view-agnostic features. Furthermore, it incorporates a distillation process from large models to refine the learning of view-specific features.

Pre-processing. At each timestep t, V frames are captured from V cameras. An off-the-shelf detector is applied to each frame to generate bounding boxes for all visible persons. The detected regions are cropped and resized to a uniform size (H, W). Since the number of detections can vary across views, the maximum number of detections N is used as a preset. For views with fewer detections, zero tensors of size (H, W, C) are added to represent missing detections. The resulting input is $D_t = \{D^j_{t,i} \in \mathbb{R}^{V\times N\times H\times W\times C} \mid i = 1, 2, \ldots, N; j = 1, 2, \ldots, V\}$, which consolidates all detections from all views at time t.

Masking. Each detection is divided into non-overlapping patches, $P = \{P_i \mid P_i \in \mathbb{R}^{C\times h\times w}\}_{i=1}^{M}$, where $M = \frac{H}{h} \times \frac{W}{w}$ is the total number of patches. These patches are converted into a sequence of tokens, $K = \{K_i \mid K_i \in \mathbb{R}^{E}\}_{i=1}^{M}$, using patch embedding and positional encoding. A subset of tokens $K^{vis} \subset K$ (e.g., 25%) is randomly sampled without replacement, and the remaining tokens are masked, following a "random masking" strategy. Although various masking strategies exist, Random masking is chosen for its simplicity and ease of implementation without requiring additional inputs.

The same mask is applied across all detections D_t to ensure consistency between views. This shared mask preserves positional encoding and prevents disruptions in cross-view reconstruction, which relies on consistent masking across views, as discussed later.

Single View Encoder. The single-view encoder Φ_{sve} is a standard Vision Transformer (ViT Dosovitskiy et al. (2021)) applied to the M^{vis} visible, unmasked patch tokens $K^{vis} \subset K$. Unlike conventional masked autoencoders, the encoder processes all unmasked tokens from detections within each view, enabling multi-head self-attention across patches in a single view. This setup captures variations between different detections, with consistent masked token positions enhancing cross-detection learning.

Positional embeddings for the patch tokens are generated using a sinusoidal function across all detection patches within a view. This ensures that while unmasked tokens may occupy the same positions across detections because of consistent mask, their positional embeddings remain distinct.

The encoder outputs features split evenly into view-agnostic features f_a (first half) and view-specific features f_s (second half), as follows: $f_a, f_s = \Phi_{sve}(K^{vis}), \quad f_a, f_s \in \mathbb{R}^{M^{vis} \times \frac{E}{2}}$

Distillation Decoder. We project the view-specific encoder features to the decoder width E_d with a linear layer and concatenate learned embeddings for the masked positions to form a length-M token sequence. This sequence is fed to a shallow ViT decoder Φ_{distill}. Positional embeddings are added to *all* tokens so that masked tokens retain their spatial coordinates.

The decoder outputs per-patch features for the entire detection, $\hat{f}_s \in \mathbb{R}^{M \times E_d}$. In parallel, the corresponding unmasked crop is processed by a pretrained teacher to obtain patch-level targets. Before computing the distillation loss, a linear head is used to align the student features $f_{student}$ to the teacher feature space $f_{teacher}$.

We use the publicly released ViT-L MAE model He et al. (2021) pretrained on ImageNet-1K (self-supervised). Its single-view pretraining emphasizes view-specific cues—e.g., color, fine textures, and local details—while contributing less to view-agnostic properties such as aspect ratio, coarse silhouette, or pose. This makes it a good supervisor for the view-specific branch (via patch-level distillation) while leaving the cross-view branch to learn identity-consistent signals across cameras.

Cross View Encoder. The view-agnostic features are passed through a pooling layer to combine patch information, producing a single view-agnostic embedding per detection. Note that no information is mixed across cameras at this stage—only patches within the same detection are combined. All embeddings from each view are then projected into the cross-view encoder dimension, E_d, and sent to a shallow ViT-based cross-view encoder. Multi-head self-attention is applied across these embeddings to capture differences between views. The output feature $\hat{f}_a \in \mathbb{R}^{E_d}$ is learnt through all the views, representing the high-level semantic features that are universal across views.

Reconstruction Decoder. The view-agnostic feature $\hat{f}_a$ is combined with the view-specific feature $\hat{f}_s$ for each patch, creating an enriched representation that captures both cross-view consistency and camera-specific details. These combined features are fed into the reconstruction decoder, which reconstructs the original image by predicting pixel values for each masked patch.

During decoding, each output vector from the decoder represents the pixel values of a specific patch, effectively reconstructing masked areas. The decoder's final layer uses a linear projection to match the total pixel count per patch, preserving each patch's spatial structure. After projection, the output is reshaped to form a coherent, reconstructed image, closely resembling the original input.

Loss. To ensure robust training, we employ a combination of three losses:

Disentanglement Loss: This normalized mutual information (NMI) loss measures the independence between view-agnostic and view-specific features, quantifying how much information about one feature set is shared by the other:

$$L_{\text{disentangle}} = NMI(f_{\text{a}}, f_{\text{s}})$$

Minimizing $L_{\text{disentangle}}$ enhances feature disentanglement by reducing shared information between the two feature sets.

Distillation Loss: This loss facilitates knowledge transfer from a larger teacher model pretrained on a different dataset. Given potential domain differences, Smooth L1 Loss is used to mitigate the

impact of outliers:

$$L_{\text{distillation}} = \text{SmoothL1}(f_{\text{student}}, f_{\text{teacher}})$$

Reconstruction Loss: This loss calculates the mean squared error (MSE) between the reconstructed and original images in pixel space, applied only to masked patches:

$$L_{\text{reconstruction}} = \text{MSELoss}(f^{\text{masked}}_{\text{reconstructed}}, f^{\text{masked}}_{\text{original}})$$

The overall loss function combines these components:

$$\text{Loss} = L_{\text{disentangle}} + L_{\text{distillation}} + L_{\text{reconstruction}}$$

3.3 Inference

A key advantage of **CALIBFREE** is its independence from camera calibration and human annotations. While both the single-view encoder/decoder and the cross-view encoder are used during self-supervised training, only the single-view encoder is needed at inference.

During inference, all patches are passed (unmasked) through the single-view encoder to generate feature embeddings. These features are average-pooled across patches to produce a single embedding per detection, which is then split into view-agnostic and view-specific components. For single-camera tracking, we integrate the view-specific features into DeepSORT Wojke et al. (2017) for within-camera association, using Kalman filtering to refine tracks. For cross-camera matching, we use the view-agnostic features to compute the association matrix, without applying any Kalman filter.

4 Results

Due to page limits, we include dataset, evaluation metrics, implementation details and more results in the Appendix.

4.1 Main Results and Analysis

4.1.1 Baseline Methods

We compare our method against state-of-the-art approaches in Tables 1 and 2, where the best results are highlighted and second-best underlined.

For single-camera (over-time) tracking, we include four representative MOT methods: Tracktor++ Bergmann et al. (2019), CenterTrack Zhou et al. (2020), TraDeS Wu et al. (2021), and TrackFormer Meinhardt et al. (2022). Since these do not support cross-view tracking, we assign ground-truth IDs upon first appearance in each camera and apply each method independently within views.

For MCMOT, we evaluate DeepCC Ristani & Tomasi (2018) and SVT Dong et al. (2021), with DeepCC leveraging an off-the-shelf ReID model Zhong et al. (2017) for cross-view association. We also include MvMHAT and its extension MvMHAT++, two self-supervised methods that require no fine-tuning, with MvMHAT++ introducing an additional training stage. All MCMOT methods are evaluated on the MMP-MvMHAT dataset using ground-truth bounding boxes for consistency, except that previous self-supervised methods are also tested with YOLOX-generated detections.

Detector-based results using YOLOv7 and YOLOX are shown in Table 4. For MvMHAT, we use a Detectron2 Wu et al. (2019) detector; among available models, we select the ResNet-50 variant that achieves MOTA closest to the original MvMHAT paper for fair comparison.

Lastly, we do not re-train supervised baselines on our dataset, as they require labeled data—unlike our self-supervised approach—ensuring a fair comparison in terms of generalization and calibration-free capability.

4.1.2 Results on MMP-MvMHAT

Over-Time Tracking: Table 1 reports CALIBFREE's performance on the indoor-focused MMP-MvMHAT dataset, where many subjects exhibit limited motion (e.g., seated office workers). Despite

Methods	*Over-time Tracking* IDP	IDR	IDF1	MOTA	HOTA	*Cross-View Tracking* AIDP	AIDR	AIDF1	MHAA	*Overall* *A*	*F*
supervised											
Tracktor++Bergmann et al. (2019)	67	56	61	67.2	46.2	62	23.2	33.8	19.1	43.1	47.4
CenterTrackZhou et al. (2020)	35.9	24.1	28.8	48.2	27.1	29.3	3.9	6.8	3.1	25.7	17.8
TraDeSWu et al. (2021)	59.7	50.1	54.5	66.1	42.7	54.5	17.3	26.2	13.3	39.7	40.4
TrackFormerMeinhardt et al. (2022)	41.8	28.6	34	46.7	30.2	39.9	5.3	9.4	3.2	25	21.7
DeepCCRistani & Tomasi (2018)	51.6	52.5	52.1	92.5	59.3	42.7	23.4	30.2	19.7	56.1	41.2
SVTDong et al. (2021)	63.1	63.4	63.3	96.7	68.8	53.8	33.4	41.2	29.8	63.1	52.3
self-supervised											
MvMHAT(YOLOX)Gan et al. (2021)	51.1	53.2	52.7	82.1	47.2	30.4	17.1	23.2	14.1	48.1	37.9
MvMHAT(GT)Gan et al. (2021)	58.6	58.8	58.7	93.7	65	35.4	21.2	26.5	20.3	57	42.6
MvMHAT++(YOLOX)Feng et al. (2024)	59.3	60.5	60.1	82.2	52.1	45.7	34.1	40.5	28.9	55.5	50.3
MvMHAT++(GT)Feng et al. (2024)	67.1	67.6	67.3	95	70.2	**62.1**	42.5	50.4	40.6	67.8	58.9
CALIBFREE(YOLOX)	81.3	77.1	79.1	82.5	59.2	52.3	48.4	50.3	34.4	58.4	64.7
CALIBFREE(GT)	**82.2**	**78**	**80**	**97.6**	**75.7**	55	**50.9**	**52.8**	**44.2**	**70.8**	**66.4**

Table 1: **Results on MMP-MvMHAT.** CALIBFREE surpasses both supervised and self-supervised methods across key metrics, demonstrating robust identity tracking in over-time and cross-view scenarios.

Methods	*Over-time Tracking* IDP	IDR	IDF1	MOTA	HOTA	*Cross-View Tracking* AIDP	AIDR	AIDF1	MHAA	*Overall* *A*	*F*
supervised											
Tracktor++Bergmann et al. (2019)	54.2	40.1	46.1	66.5	42.8	34.3	14.6	20.5	37.1	51.8	33.3
CenterTrackZhou et al. (2020)	44.3	33.5	38.1	63.5	37.8	29.7	9.1	13.9	34.1	48.8	26.0
TraDeSWu et al. (2021)	46.7	43.2	44.9	69.5	42.9	32.4	14.0	19.6	36.0	52.8	32.2
TrackFormerMeinhardt et al. (2022)	52.3	47.2	49.6	**70.4**	47.3	47.8	23.2	31.3	40.2	55.3	40.4
DeepCCRistani & Tomasi (2018)	44.7	44.2	44.4	63.9	41.1	57.9	34.8	43.4	43.8	53.9	43.9
SVTDong et al. (2021)	47.9	47.2	47.6	65.4	43.1	61.7	45.7	52.5	50.4	56.9	50.0
self-supervised											
MvMHATGan et al. (2021)	53.1	52.0	52.5	64.7	47.9	53.0	46.4	49.5	51.7	58.2	51.0
MvMHAT++Feng et al. (2024)	58.5	57.4	57.9	66.3	51.8	**63.8**	56.0	**59.6**	**59.7**	**63.0**	**58.8**
CALIBFREE	**59.1**	**58.4**	**58.7**	60.4	**52.0**	58.9	**56.2**	57.4	57.0	58.4	58.1

Table 2: **Results on MvMHAT.** CALIBFREE achieves best or second best results across most key metrics.

the simplicity of such motion—often inflating IDF1 for other methods—CALIBFREE achieves an IDF1 of 80.0, demonstrating strong identity continuity under occlusion and crowding. Its HOTA score of 75.7 reflects balanced accuracy in detection and association, minimizing ID switches and ensuring stable long-term tracking.

Cross-View Tracking: CALIBFREE achieves an AIDF1 of 52.8 and MHAA of 44.2, outperforming all baselines. While its AIDP is slightly lower than MvMHAT++, CALIBFREE achieves higher AIDR, indicating better recall of cross-camera matches. This reflects its ability to capture identity-consistent features under challenging viewpoint changes and appearance similarity. Compared to supervised baselines like DeepCC and SVT—which require manual annotations—CALIBFREE delivers stronger association without labels. Center-based trackers (e.g., CenterTrack) lack robust appearance modeling and accumulate ID switches in crowded scenes, while TrackFormer can produce mismatches when its detection step underperforms.

Sensitivity to Bounding Box Quality: As shown in Table 1, CALIBFREE is more robust to noisy bounding boxes than prior self-supervised methods using YOLOX detections. While others suffer significant performance drops, CALIBFREE maintains ID-related metrics with minimal degradation, highlighting the resilience of its learned features to imperfect detections.

Overall: CALIBFREE surpasses self-supervised baselines (MvMHAT, MvMHAT++) in both accuracy (70.8) and F1 score (66.4), demonstrating the advantage of its disentangled features for both temporal and spatial consistency. Although motion in MMP-MvMHAT is simpler, the cluttered indoor layout introduces frequent identity ambiguities, which CALIBFREE handles effectively.

4.1.3 Results on MvMHAT

Over-Time Tracking. Table 2 presents results on the MvMHAT dataset, which includes both indoor and outdoor scenes with sparsely placed cameras and minimal overlap. Single-camera methods like TrackFormer yield competitive MOTA (e.g., 70.4), largely reflecting detector quality. However, CALIBFREE achieves higher ID-based metrics (IDP, IDR, IDF1) and HOTA, indicating better temporal identity consistency.

Cross-View Tracking. Single-view trackers struggle when ID switches occur, propagating errors across views and degrading AIDF1. That said, their cross-view performance improves slightly over MMP-MvMHAT due to reduced overlap and fewer direct transitions. CALIBFREE outperforms multi-view trackers like DeepCC and SVT across all cross-view metrics. MvMHAT++ achieves

Methods	*Over-time Tracking*					*Cross-View Tracking*				*Overall*	
	IDP	IDR	IDF1	MOTA	HOTA	AIDP	AIDR	AIDF1	MHAA	*A*	*F*
ViT-L (teacher)	82.1	77.8	79.9	97.5	75.6	47.9	44.4	46.1	36.7	67.1	63
Distillation only	78.4	67.8	72.7	97.2	68.4	47.5	40.1	43.5	34.4	65.8	58.1
Reconstruction only	81.2	77.1	79.1	97.5	75.2	52.3	48.4	50.3	41.8	69.7	64.7
CALIBFREE(ours)	**82.2**	**78**	**80**	**97.6**	**75.7**	**55**	**50.9**	**52.8**	**44.2**	**70.8**	**66.4**

Table 3: **Ablation studies of CALIBFREE variations.** The full model, combining distillation, reconstruction, and feature disentanglement, achieves the best performance across all tracking metrics.

Detector Pretrain	Detector Inference	*Over-time Tracking*					*Cross-View Tracking*				*Overall*	
		IDP	IDR	IDF1	MOTA	HOTA	AIDP	AIDR	AIDF1	MHAA	*A*	*F*
YOLO7	YOLO7	59.9	57.5	58.6	44.9	47.7	47.2	44.3	45.7	16.4	30.7	52.2
Groundtruth	YOLO7	59.8	57.4	58.5	44.9	47.7	49.3	46.1	47.6	18.3	31.6	53.1
YOLOX	YOLOX	81.3	77.1	79.1	82.5	59.2	52.3	48.4	50.3	34.4	58.4	64.7
Groundtruth	YOLOX	81	76.8	78.9	82.5	59	54	50.1	52	37.5	60	65.5
Groundtruth	**Groundtruth**	**82.2**	**78**	**80**	**97.6**	**75.7**	**55**	**50.9**	**52.8**	**44.2**	**70.8**	**66.4**

Table 4: **Ablation study on detector choice during pretraining and inference.** Results show that CALIBFREE maintains high ID association consistency, with inference detector choice impacting tracking accuracy more than pretraining.

higher precision and AIDF1 in this setting, which we attribute to two factors: (1) CALIBFREE uses a less accurate detector (lower MOTA), reducing cross-view consistency; and (2) MvMHAT++ benefits from a second-stage training step specifically tailored for cross-view association, offering an advantage in sparsely overlapped environments.

Overall Performance. CALIBFREE demonstrates strong gains over most single-camera and multi-view baselines in overall accuracy (A) and F1 score (F). By disentangling view-specific and view-agnostic features, it maintains identity across views without calibration. Although MvMHAT++ excels in some cross-view metrics, CALIBFREE's unified, annotation-free framework delivers robust and generalizable performance under real-world challenges like occlusions, sparse views, and subject similarity.

4.2 Ablation Studies

4.2.1 Effect of Distillation and Reconstruction

Table 3 presents an ablation study comparing four CALIBFREE variants. The *ViT-L (teacher)* setting uses features from a pretrained ViT-L Dosovitskiy et al. (2021) directly, without training a student. *Distillation only* trains a student using only the distillation loss $\mathcal{L}_{\text{distill}}$, while *Reconstruction only* trains a student from scratch using only $\mathcal{L}_{\text{recon}}$. Our *full model* combines $\mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{disent}}$ and outputs disentangled view-specific and view-agnostic features.

The ViT-L teacher achieves strong over-time tracking (IDF1: 79.9, MOTA: 97.5) but limited cross-view performance (AIDF1: 46.1, MHAA: 36.7). Distillation alone underperforms due to reduced model capacity and masked inputs (AIDF1: 43.5). Reconstruction alone slightly lowers over-time performance (IDF1: 79.1) but improves cross-view accuracy (AIDF1: 50.3), highlighting the importance of spatial reconstruction for multi-view consistency. The full CALIBFREE model achieves the best overall results (F: 66.4, Accuracy: 70.8), validating the importance of all components for robust uncalibrated tracking.

4.2.2 Effect of Detector Choice

Table 4 compares three detector configurations—ground truth, YOLOX Ge et al. (2021), and YOLOv7 Wang et al. (2023)—used during both pretraining and inference. CALIBFREE demonstrates strong robustness to detector choice during pretraining; however, inference quality has a more substantial effect. Switching from YOLOX to the less accurate YOLOv7 results in noticeable performance drops, highlighting the importance of reliable detections—a challenge common to all tracking methods. Notably, models pretrained with ground truth and inferred using YOLOX achieve ID-based metrics comparable to those with ground-truth inference, demonstrating CALIBFREE's adaptability when the inference detector maintains reasonable accuracy. Moreover, as shown in Table 1, CALIBFREE is significantly more resilient to bounding box imperfections compared to previous self-supervised methods.

Teacher Model	Student Model	*Over-time Tracking* IDP	IDR	IDF1	MOTA	HOTA	*Cross-View Tracking* AIDP	AIDR	AIDF1	MHAA	*Overall* *A*	*F*
ViT-L	**ViT-B**	**82.2**	**78**	**80**	**97.6**	**75.7**	**55**	**50.9**	**52.8**	**44.2**	**70.8**	**66.4**
ViT-B	ViT-B	82.0	77.8	79.8	97.5	75.5	52.1	47.8	49.7	42.0	69.8	64.8
ViT-B	ViT-S	77.6	75.8	76.7	97.2	71.8	50.3	45.4	47.7	40.4	68.8	62.2

Table 5: **Ablation studies of different models.** Using ViT-L as teacher and ViT-B as student achieves best results.

Mask ratio	*Over-time Tracking* IDP	IDR	IDF1	MOTA	HOTA	*Cross-View Tracking* AIDP	AIDR	AIDF1	MHAA	*Overall* *A*	*F*
0.9	**82.3**	78	**80.1**	97.6	**75.8**	53.8	48.7	51.1	43.0	70.3	65.6
0.75	82.2	**78**	80	**97.6**	75.7	55	**50.9**	**52.8**	**44.2**	**70.8**	**66.4**
0.5	82.0	77.9	79.8	97.5	75.6	**55.1**	50.1	52.4	43.9	70.7	66.1

Table 6: **Ablation studies of mask ratios.** 0.75 achieves the best balance between over-time and cross-view tracking.

4.2.3 Impact of Teacher and Student Model Sizes

Table 5 investigates three teacher–student setups: ViT-L→ViT-B, ViT-B→ViT-B, and ViT-B→ViT-S. A larger teacher (ViT-L) improves cross-view performance (AIDF1: 49.7→52.8, MHAA: 42.0→44.2) due to richer representations feeding into the cross-view encoder. Over-time metrics (IDF1, MOTA) remain mostly unchanged, suggesting that temporal continuity is less sensitive to teacher size. On the other hand, reducing the student to ViT-S lowers both over-time (IDF1: 76.7) and cross-view (AIDF1: 47.7) performance, indicating insufficient capacity for robust identity modeling. Although smaller students are more efficient, this comes at the cost of accuracy—especially in complex multi-view scenarios.

4.2.4 Effect of Mask Ratios in Pretraining

We examine the impact of mask ratios (0.5, 0.75, 0.9) in Table 6. High masking (e.g., 0.9) harms cross-view performance (AIDF1: 51.1, MHAA: 43.0), suggesting that excessive masking limits the model's ability to learn spatially consistent features. Over-time metrics (IDF1, MOTA) remain stable, as temporal tracking depends less on detailed spatial information. While both 0.5 and 0.75 achieve comparable cross-view results (AIDF1: 52.4 vs. 52.8), the 0.75 setting reduces token usage, offering better efficiency. Thus, a 0.75 mask ratio strikes the best balance between accuracy and computational cost for both temporal and cross-view tracking.

5 Conclusion, Limitation, and Future Work

We have introduced **CALIBFREE**, a self-supervised multi-camera multi-object tracking (MCMOT) method that achieves state-of-the-art performance without relying on camera calibration or manual annotations. By disentangling view-agnostic and view-specific features, supported by cross-view reconstruction and knowledge distillation, CALIBFREE robustly handles complex identity associations across time and views. Experiments on the MMP-MvMHAT and MvMHAT datasets underscore its strong adaptability in both over-time and cross-view tracking.

Despite these advances, our approach remains limited by its exclusive use of RGB features, thereby overlooking valuable geometric relationships across views. Although learning geometric associations without camera parameters is nontrivial, incorporating such information could enhance identity consistency and cross-view associations. In future work, we aim to explore self-supervised methods for integrating geometric cues, enabling CALIBFREE to better leverage spatial relationships between detections.

Ethics Statement. This work uses publicly available datasets (MvMHAT, MMP-MvMHAT) with no personally identifiable information or human-subject interaction. We adhere to the ICLR Code of Ethics. While multi-camera tracking may raise privacy concerns, our contributions are intended for academic research and do not enable identification beyond provided annotations.

Reproducibility Statement. Model architecture, objectives, training schedules, and evaluation protocols are specified in Secs. 3–4 and the appendix; ablations (Sec. 4.4) support design choices.

Datasets and preprocessing steps are documented, and we will release code, pretrained weights, and evaluation scripts to reproduce all reported results upon publication.

REFERENCES

https://iccv2021-mmp.github.io/.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24206–24221. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cb3213ada48302953cb0f166464ab356-Paper.pdf.

Favyen Bastani, Songtao He, and Samuel Madden. Self-supervised multi-object tracking with cross-input consistency. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13695–13706. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/71e09b16e21f7b6919bbfc43f6a5b2f0-Paper.pdf.

Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 941–951, 2019. URL https://api.semanticscholar.org/CorpusID:76665153.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8090–8100, 2022.

Yinghao Cai and Gerard Medioni. Exploring context information for inter-camera multiple target tracking. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 761–768. IEEE, 2014.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9686–9696, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020. URL https://api.semanticscholar.org/CorpusID:211096730.

Xiaotang Chen, Kaiqi Huang, and Tieniu Tan. Object tracking across non-overlapping views by learning inter-camera transfer models. *Pattern Recognition*, 47(3):1126–1137, 2014.

Cheng-Che Cheng, Min-Xuan Qiu, Chen-Kuo Chiang, and Shang-Hong Lai. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10051–10060, 2023.

Amit Chilgunde, Pankaj Kumar, Surendra Ranganath, and Weimin Huang. Multi-camera target tracking in blind regions of cameras with non-overlapping fields of view. In *BMVC*, pp. 1–10. Citeseer, 2004.

Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6172–6181, 2019.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Junting Dong, Qi Fang, Wen Biao Jiang, Yurou Yang, Qi-Xing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:6981–6992, 2021. URL https://api.semanticscholar.org/CorpusID:236159249.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6569–6578, 2019.

Wei Feng, Feifan Wang, Ruize Han, Yiyang Gan, Zekun Qian, Junhui Hou, and Song Wang. Unveiling the power of self-supervision for multi-view multi-human association and tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, 2008. doi: 10.1109/TPAMI.2007.1174.

Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. URL https://api.semanticscholar.org/CorpusID:239011901.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Andrew Gilbert and Richard Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II 9*, pp. 125–136. Springer, 2006.

R Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.

Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5496–5506, 2023.

Ruize Han, Wei Feng, Jiewen Zhao, Zicheng Niu, Yujun Zhang, Liang Wan, and Song Wang. Complementary-view multiple human tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10917–10924, 2020.

Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5225–5242, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019. URL https://api.semanticscholar.org/CorpusID:207930212.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2021. URL https://api.semanticscholar.org/CorpusID:243985980.

Yuhang He, Xing Wei, Xiaopeng Hong, Weiwei Shi, and Yihong Gong. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE Transactions on Image Processing*, 29:5191–5205, 2020.

Zhicheng Huang, Xiaojie Jin, Cheng Lu, Qibin Hou, Mingg-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:2506–2517, 2022. URL https://api.semanticscholar.org/CorpusID:251105242.

Omar Javed, Khurram Shafique, and Mubarak Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pp. 26–33. IEEE, 2005.

Dahun Kim, Donghyeon Cho, and In-So Kweon. Self-supervised video representation learning with space-time cubic puzzles. *ArXiv*, abs/1811.09795, 2018a. URL https://api.semanticscholar.org/CorpusID:53762354.

Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 793–802. IEEE, 2018b.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6874–6883, 2017.

Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 33–40, 2016.

Zijia Lu, Bing Shuai, Yanbei Chen, Zhenlin Xu, and Davide Modolo. Self-supervised multi-object tracking with path consistency. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19016–19026, 2024. URL https://api.semanticscholar.org/CorpusID:269005039.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021.

Andrii Maksai, Xinchao Wang, Francois Fleuret, and Pascal Fua. Non-markovian globally consistent multi-object tracking. In *Proceedings of the IEEE international conference on computer vision*, pp. 2544–2554, 2017.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8844–8854, 2022.

Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

Alexandru Niculescu-Mizil, Deep Patel, and Iain Melvin. Mctr: Multi camera tracking transformer. *arXiv preprint arXiv:2408.13243*, 2024.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *ArXiv*, abs/1603.09246, 2016. URL https://api.semanticscholar.org/CorpusID:187547.

Ziqi Pang, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, and Yu-Xiong Wang. Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17928–17938, 2023.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Bryan James Prosser, Shaogang Gong, and Tao Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, volume 8, pp. 164–1. Leeds, UK, 2008.

Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6036–6046, 2018. URL https://api.semanticscholar.org/CorpusID:4462331.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pp. 17–35. Springer, 2016a.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pp. 17–35. Springer, 2016b.

Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6951–6960, 2017.

Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, and Mubarak Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv preprint arXiv:1706.06196*, 2017.

Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, and Mubarak Shah. Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets. *International Journal of Computer Vision*, 127:1303–1320, 2019.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. URL https://api.semanticscholar.org/CorpusID:49670925.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 504–521, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58520-4.

Luyu Wang, Pauline Luc, Adrià Recasens, Jean-Baptiste Alayrac, and Aäron van den Oord. Multimodal self-supervised learning of general audio representations. *ArXiv*, abs/2104.12807, 2021. URL https://api.semanticscholar.org/CorpusID:233407605.

Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6):1336–1353, 2012.

G Welch. An introduction to the kalman filter. 1995.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649. IEEE, 2017.

Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12347–12356, 2021. URL https://api.semanticscholar.org/CorpusID:232240682.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3988–3998, 2019.

Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. How to train your deep multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6787–6796, 2020.

Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yuanhang Yin, Yang Hua, Tao Song, Ruhui Ma, and Haibing Guan. Self-supervised multi-object tracking with cycle-consistency. In *Conference on Multimedia Modeling*, 2023. URL https://api.semanticscholar.org/CorpusID:257986445.

Quanzeng You and Hao Jiang. Real-time 3d deep multi-camera tracking. *arXiv preprint arXiv:2003.11753*, 2020.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pp. 659–675. Springer, 2022.

Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4537–4546, 2022a.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. 2022b.

Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22056–22065, 2023.

Zelin Zhao, Ze Wu, Yueqing Zhuang, Boxun Li, and Jiaya Jia. Tracking objects as pixel-wise distributions. In *European Conference on Computer Vision*, pp. 76–94. Springer, 2022.

Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. *CoRR*, abs/1711.10295, 2017. URL http://arxiv.org/abs/1711.10295.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ArXiv*, abs/2004.01177, 2020. URL https://api.semanticscholar.org/CorpusID:214775104.

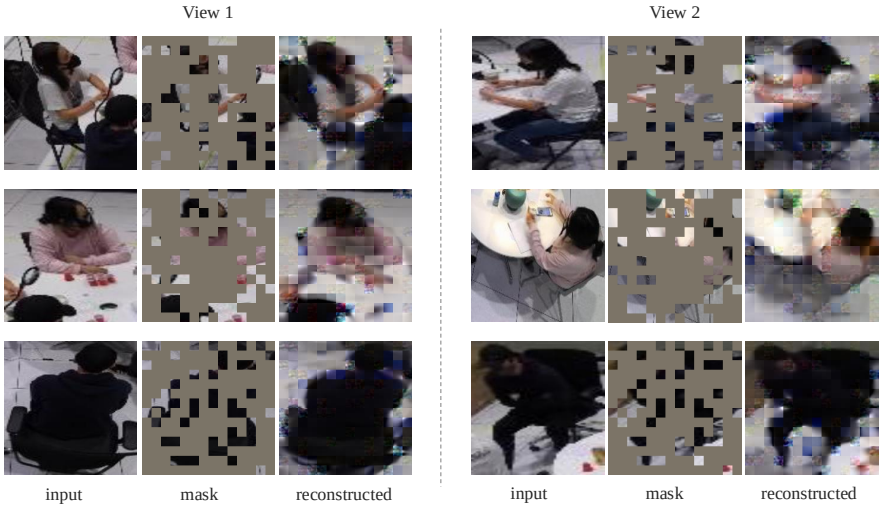

Figure 3: **Cross-view reconstruction results.** The input images are the original scenes, while the masked images indicate regions removed for reconstruction. CALIBFREE effectively reconstructs these masked regions, capturing identity-preserving details across viewpoints even when large portions are obscured.

A Key Takeaways and Calibration Independence

Calibration-free design. Our approach never accesses camera intrinsics, extrinsics, or homographies at any stage of training or inference. As a result, deployment to a new camera network requires no calibration effort, and the method remains robust even if cameras are re-positioned or experience drift over time (Section. 1, Figure. 1).

Disentangled feature learning. A central insight is that explicitly disentangling two complementary embeddings—learned via a masked auto-encoder without any camera metadata—is both feasible and beneficial. The *view-specific* branch preserves appearance nuances tied to a particular camera, while the *view-agnostic* branch captures identity cues consistent across viewpoints. This separation (i) integrates seamlessly with off-the-shelf trackers such as DeepSORT Wojke et al. (2017), (ii) eliminates calibration and labeling overhead, and (iii) opens new directions for cross-source representation learning (e.g., audio–video or LiDAR–camera).

B Datasets

MMP-MvMHAT. Adapted from MMPTRACK mmp, MMP-MvMHAT features 4–6 overlapping indoor cameras and 28 individuals. It provides 8,000 fully annotated frames across four training scenes and 4,000 frames for validation, with no calibration data. This setup, focused on crowded and occluded environments, poses a challenging multi-view tracking task.

MvMHAT. MvMHAT Feng et al. (2024) is a large-scale dataset containing 26 video groups (98 sequences) sourced from Campus Xu et al. (2016), EPFL Fleuret et al. (2008), and newly collected footage. Each group includes 3–4 synchronized camera views, totaling over 90,000 annotated frames. Split into training and testing (13 groups each) with a 2:1 ratio, MvMHAT covers diverse scenarios and camera angles—often with 90° viewpoint differences—to facilitate robust multi-view tracking evaluation.

C Evaluation Metrics

Over-time Tracking. We adopt Multiple Object Tracking Accuracy (MOTA) Bernardin & Stiefelhagen (2008) to assess single-view tracking performance in terms of false positives, missed detections, and identity switches. Given the emphasis on robust identity association over time, we further use ID Precision (IDP), ID Recall (IDR), and ID F1 (IDF1) Ristani et al. (2016a), as well as

High Order Tracking Accuracy (HOTA) Luiten et al. (2021) for a balanced evaluation of detection, association, and localization.

Cross-view Tracking. For multi-camera scenarios, we use Association ID Precision (AIDP), Association ID Recall (AIDR), and Association ID F1 (AIDF1) Han et al. (2020; 2021), which average pairwise matching accuracy across different cameras. We also include Multi-view Multi-Human Association Accuracy (MHAA), which penalizes identity-consistency errors in multi-camera contexts with frequent occlusions and appearance shifts.

Overall. To provide a holistic MCMOT assessment, we calculate the MCMOT F1 score (*F*) and accuracy score (*A*) by taking the average of F1 and accuracy across both over-time and cross-view tracking Feng et al. (2024):

$$F = \text{Mean(IDF1, AIDF1)} \quad , \quad A = \text{Mean(MOTA, MHAA)}.$$

D Implementation Details

Pretraining Phase. After detecting the bounding box for each person, the region of interest (ROI) is cropped based on the bounding box coordinates and resized to a fixed resolution of 224×224 pixels through upsampling or downsampling. These resized ROIs are divided into non-overlapping patches of size 16×16.

The *single-view encoder* is implemented as a vanilla Vision Transformer (ViT) base model with 12 transformer blocks and 12 attention heads, using an embedding dimension of 768. During inference, the output of the single-view encoder is split into view-agnostic and view-specific features, each with a dimension of 384.

The *single-view decoder*, *cross-view encoder*, and *cross-view decoder* are shallow Vision Transformer models, each comprising 8 attention blocks and 16 attention heads with an embedding dimension of 512. For knowledge distillation, the *teacher model* is a ViT large model pretrained on ImageNet, with an output feature dimension of 1024. To align dimensions during distillation, the 512-dimensional output of the single-view decoder is projected to 1024 dimensions using a linear layer.

Pretraining is conducted over 400 epochs with a base learning rate of 1.5×10^{-4} and a 40-epoch warmup phase. The weight decay is set to 0.05. Reconstruction loss is computed using mean squared error (MSE) loss applied to normalized pixel values rather than raw pixel values. The maximum number of detections is set to be 10 for both datasets.

Inference Phase. During inference, a detector identifies bounding boxes for all persons, and the ROIs are cropped based on the bounding box coordinates. Features for each detection are generated by feeding all patches (without masking) to the single-view encoder. The per-patch features are then aggregated using max pooling to produce a single feature vector (matching the encoder dimension) representing each detection.

For *over-time tracking*, we utilize DeepSort, with the generated features serving as the primary matching criterion, complemented by a Kalman filter as a secondary criterion. For *cross-view tracking*, aggressive matching is employed, computing pairwise similarity between detections from different views to establish associations.

E Visualization

Figure 3 illustrates how CALIBFREE captures identity-preserving features across different viewpoints, showing the original input, masked patches, and reconstructed outputs. Substantial portions of each subject are masked to simulate partial observations, yet CALIBFREE reliably restores these regions by leveraging both view-agnostic and view-specific features. Crucial details like posture, clothing texture, and overall silhouette remain largely intact, supporting consistent identity tracking across camera views. Even when significant information is obscured, the model reconstructs occluded areas accurately based on the available visible patches, all without requiring camera calibration. This visualization underscores CALIBFREE's robustness and practical utility in multi-camera scenarios with frequent occlusions and substantial viewpoint variations.

F Computation and Runtime

We report the hardware setup and training time for each dataset:

Dataset	# GPUs	GPU Model	Mem/GPU	Batch	Epochs	Wall-clock
MMP-MvMHAT	4	NVIDIA H100	80 GB	5	400	13.4 h
MvMHAT	4	NVIDIA A100	80 GB	5	400	38.3 h

At inference, using four cameras with 1080P input resolution, a ViT-B encoder (as single-view encoder) for each view, and DeepSORT as the tracker, the average runtime per timestep is 154 ms (± 32 ms) on one NVIDIA A100 80GB GPU.

G Use of Large Language Models (LLMs).

We used LLMs only as a writing assistant for improving clarity and conciseness of our text (e.g., rewriting and rephrasing). LLMs were not involved in research ideation, design, experimentation, or analysis. The authors take full responsibility for all content.