HieDG: A Hierarchical Discrete Geometry-Guided Framework for Multi-Animal Tracking

Chenxun Deng[1,2], Zhongde Zhang[3], Ye Yuan[1,4], Chengyang Zhang[5], Yifan Zhang[6], Bohao Chen[1,4], Hongying Yan[7], Hang Zhou[8], Hua Han[1,9], and Xi Chen[1*]

[1] State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] School of Technology, Beijing Forestry University
[4] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[5] College of Computer Science, Sichuan University
[6] C²DL, Institute of Automation, Chinese Academy of Sciences.
[7] School of Automation, Chongqing University
[8] Faculty of Life and Health Sciences, Shenzhen University of Advanced Technology
[9] School of Future Technology, University of Chinese Academy of Sciences
xi.chen@ia.ac.cn

Abstract. Multi-animal tracking (MAT) is critical for wildlife monitoring and behavioral analysis, yet remains challenging due to uniform appearance, high density, and irregular motion. Existing methods typically follow heuristic- or query-based paradigms: the former relies on handcrafted geometric associations without end-to-end optimization, whereas the latter enables joint optimization but relies heavily on appearance embeddings. In such conditions, continuous geometric embeddings can be unstable, as small coordinate perturbations may disproportionately alter cross-frame attention weights, degrading identity association performance. To address this limitation, we propose **HieDG**, a **Hie**rarchical **D**iscrete **G**eometry-guided tracking framework that reformulates geometric dynamics as structured discrete representations within a query-based tracker. Instead of directly using raw geometric signals, HieDG employs a two-stage residual codebook to discretize position, scale, and velocity cues, transforming unstable continuous geometry into structured, stable discrete tokens. These tokens are aligned with visual embeddings and integrated into the tracking queries to enhance identity consistency. Extensive experiments on animal-specific benchmarks (AnimalTrack, BFT, and BuckTales) demonstrate state-of-the-art association performance with significant improvements in HOTA, AssA, and IDF1. Additional evaluations on generic multi-object tracking benchmarks, including DanceTrack and SportsMOT, show competitive performance, indicating the broader applicability of discretized geometric modeling beyond animal-specific scenarios.

Keywords: multi-object tracking · discrete geometric embedding · residual quantization · query-based tracking

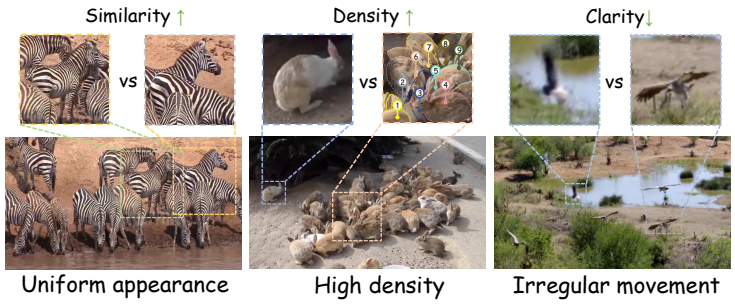
1 Introduction

Multi-animal tracking (MAT), a representative yet challenging multi-object tracking problem, plays a pivotal role in biology, ecology, and wildlife research, supporting applications such as animal monitoring [16,20], population estimation [3], and behavior analysis [30]. MAT involves accurately locating all animals within a video stream while consistently maintaining their identities throughout the sequence. Despite remarkable advances in human and vehicle tracking [8, 9, 21, 44], MAT remains technically demanding, particularly under *low visual discriminability regimes*, where appearance cues provide limited identity separation and geometric dynamics exhibit stronger variability.

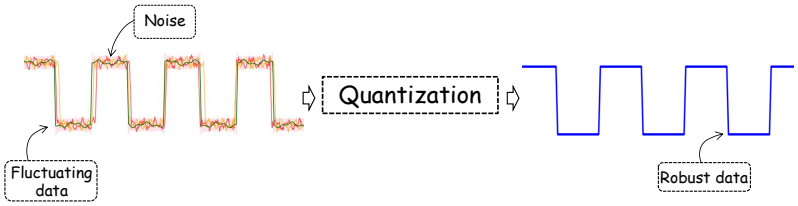
Driven by advances in object detection, the bottleneck in multi-object tracking has primarily shifted from accurate localization to robust and consistent identity association [14,36]. To address this, two major paradigms have emerged: heuristic-based association and Transformer-driven query-based tracking. Traditional heuristic-based paradigms typically adopt a detector–tracker pipeline, relying on meticulously crafted rules (e.g., Kalman filter [39] and Hungarian algorithm [26]) to model geometric cues for association [32, 45]. Despite notable progress, these methods rely on fixed association rules that are inherently incompatible with gradient-based end-to-end optimization. Consequently, they struggle to adapt to evolving visual and geometric cues, leading to suboptimal performance in complex tracking scenarios [41]. In contrast, query-based methods formulate tracking as query propagation across frames, enabling end-to-end identity modeling via cross-frame attention [11, 25, 43, 47]. Although effective on human and vehicle benchmarks, these methods are heavily appearance-driven. In scenarios with weak visual discriminability, such reliance renders identity association inherently fragile.

Multi-animal tracking presents precisely such a regime (Fig. 1a). First, animals of the same species often exhibit a highly uniform appearance due to social grouping behavior, resulting in minimal inter-instance visual distinction. Second, animals frequently form high-density groups, where close spatial proximity and frequent overlap significantly increase identity ambiguity. Third, irregular and non-planar movement patterns, especially in aerial or aquatic environments, introduce strong geometric variability across frames. Overall, these factors place MAT in a low-discriminability, high-variability regime, in which both appearance and continuous geometric cues become less reliable for stable identity modeling.

This motivates incorporating geometric cues into end-to-end query-based tracking. However, geometric modeling under attention-based architectures is non-trivial. As observed in [44], camera motion, object jitter, and environmental noise introduce subtle yet persistent fluctuations in geometric signals. Although small in raw analog space, such perturbations may induce disproportionate shifts in cross-frame attention weight distributions, leading to unstable identity matching. Empirically, as we further demonstrate in Sec. 4.3, directly incorporating continuous geometric representations can even degrade association performance relative to appearance-only baselines. This phenomenon suggests that continuous geometric embeddings may destabilize attention-based identity modeling under noisy and appearance-ambiguous conditions.



(a) Challenges of animal tracking



(b) Illustration of data quantization

Fig. 1: (a) Typical conditions in multi-animal tracking, including uniform appearance, high density, and irregular motion, which reduce appearance discriminability. (b) Quantization transforms fluctuating continuous signals into stable discrete states, suppressing perturbation-induced variability.

Inspired by quantization principles in signal processing—where noisy analog signals are mapped to stable discrete states (Fig. 1b)—we revisit geometric modeling from a representation perspective. Rather than treating geometry purely as a continuous regression signal, we introduce a discretized state-space representation for geometric dynamics. By mapping fluctuating geometric cues to a finite set of learned codewords, we constrain perturbation-induced embedding variance while preserving structural information essential for stable identity association. This discrete formulation mitigates perturbation amplification and yields more stable attention responses.

Based on this insight, we propose **HieDG**, a **Hie**rarchical **D**iscrete **G**eometry-guided framework for multi-animal tracking. HieDG employs a coarse-to-fine residual quantization strategy to discretely encode position, size, and velocity cues through independent two-level codebooks. This hierarchical design captures both coarse geometric structure and fine-grained variations, transforming unstable continuous signals into robust discrete embeddings aligned with the visual feature space. The resulting geometric tokens are fused with tracking queries, enhancing identity consistency under challenging conditions. Our contributions are summarized as follows:

– We propose HieDG, a geometry-guided multi-animal tracking framework that unifies heuristic geometric reasoning with end-to-end query-based identity modeling via discrete geometric representations.

- We design a hierarchical residual quantization strategy that encodes position, scale, and velocity cues into structured discrete tokens, reformulating geometric modeling as a robust representation-learning problem within tracking.
- Extensive experiments on AnimalTrack, BFT, and BuckTales demonstrate state-of-the-art performance with significant improvements in HOTA, AssA, and IDF1. Additional evaluations on generic MOT benchmarks show competitive performance, indicating the general applicability of discretized geometric modeling beyond animal-specific scenarios.

2 Related Work

Tracking-by-Detection has become the dominant paradigm in modern MOT, where object detection and data association are decoupled. With the rapid progress in object detection, recent efforts have increasingly focused on the association stage, aiming to construct coherent trajectories across frames. Many prior works adopt heuristic post-processing strategies to associate current detections with historical trajectories [32]. Early approaches, such as SORT [4], rely on geometry by computing an IoU matrix between frame-wise detections and applying the Hungarian algorithm for optimal matching. ByteTrack [45] further incorporates low-score detections to recover temporarily occluded targets with dropped confidence, significantly improving identity preservation in occlusion-heavy scenes. OC-SORT [6] emphasizes motion continuity by modeling trajectory consistency and smoothing motion direction, showing notable advantages in long-term occlusions and crossing scenarios. While heuristic-based paradigms perform well under certain conditions through carefully crafted geometric modeling, their fixed association rules often require manual tuning and struggle to generalize across diverse or complex scenarios. Nevertheless, their explicit use of geometric consistency provides a useful inductive bias for identity association, motivating us to integrate geometry into end-to-end trainable frameworks.

Tracking-by-Query has emerged as a powerful paradigm in modern MOT due to its end-to-end trainability and its capability for global optimization. By propagating object queries across frames and performing cross-frame attention, these methods unify detection and association, enabling joint learning of localization and identity assignment via gradient descent. TransTrack [34] first introduces Transformers into MOT by propagating tracking queries through temporal attention. TrackFormer [25] further unifies detection and association within a shared decoder to improve alignment. MOTR [43, 47] models identity propagation by dynamically reusing historical queries, allowing flexible handling of object births and terminations. In contrast, MOTIP [11] formulates MOT as an identity prediction problem trained via classification loss, avoiding explicit matching operations.

Although these approaches achieve strong performance on standard human and vehicle benchmarks, their identity modeling remains largely appearance-driven. Under conditions commonly observed in multi-animal tracking, such as

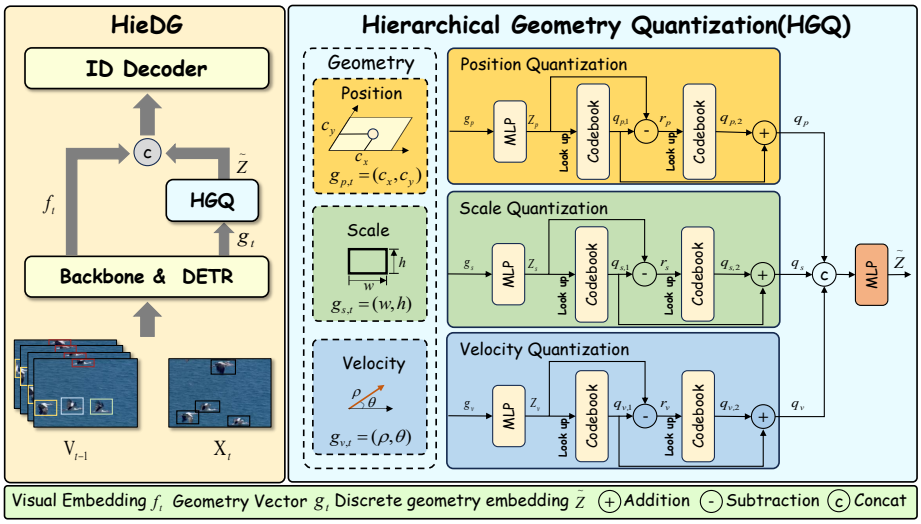


Fig. 2: Overview of the HieDG framework. Appearance and geometric features are extracted via Deformable DETR [50]. The geometric vectors are first projected through MLPs and independently quantized by two-stage residual codebooks for position, size, and velocity. The resulting discrete geometric embeddings are concatenated and aligned to match the visual space. The combined representation is finally passed to the ID decoder for trajectory association.

uniform appearance, high density, and irregular motion, appearance embeddings alone may provide limited separability, making association more challenging. Inspired by the geometric reasoning in classical heuristic pipelines, we incorporate explicit geometric cues (e.g., position, scale, and velocity) into the query-based framework to enhance robustness under visually ambiguous and dynamically complex scenarios.

2.1 Motivation and Overview

3 Method

In this section, we introduce our proposed MAT framework, HieDG, which integrates robust discrete geometric cues into the query-based paradigm. We begin with the motivation and overview in Sec. 2.1. Then, we detail the geometric vector construction and the hierarchical geometry quantization in Sec. 3.1 and Sec. 3.2, respectively. The geometry-aware animal tracking is described in Sec. 3.3. Finally, we outline the training and inference procedures in Sec. 3.4 and Sec. 3.5.

Motivation. In multi-animal tracking scenarios, uniform appearance, high density, and irregular motion patterns pose significant challenges to both heuristic-based and query-based tracking paradigms. Heuristic approaches, while effective

at geometric modeling, lack the capacity for end-to-end global optimization, limiting their adaptability across diverse animal scenes. In contrast, query-based methods benefit from joint optimization but heavily rely on visual features, which are often non-discriminative in animal settings. This motivates a natural and intuitive question: can we combine the strengths of both paradigms by leveraging the global optimization capability of query-based frameworks to model geometric cues directly from video sequences, thereby enhancing identity association?

However, within query-based tracking frameworks, association is typically performed via cross-attention between encoded features. In real-world animal monitoring videos, camera shake, object jitter, and visual noise introduce fluctuations in geometric signals. Although subtle at the raw signal level, these variations can be amplified during embedding, leading the attention mechanism to misinterpret them as genuine geometric discrepancies—ultimately compromising association reliability. In signal processing and circuit analysis, quantization is often employed to convert perturbed analog signals into discrete and deterministic states, enhancing both stability and interpretability. Inspired by this, we seek to improve the discriminability of fluctuating geometric embeddings by discretizing them into structured discrete codeword representations.

Overview. The overall architecture of HieDG is illustrated in Fig. 2. We extract object-level geometric cues (position, scale, and velocity) from Deformable DETR [50] outputs and project them into latent vectors. Each geometric cue is independently quantized using a two-stage residual codebook in a coarse-to-fine manner, producing structured discrete tokens. These tokens constrain perturbation-induced embedding variance while preserving essential geometric information. The resulting discrete geometric embeddings are aligned with the visual feature space and fused with appearance features within the query-based tracking framework to guide robust identity association.

3.1 Geometric Vector Construction

We adopt Deformable DETR [50] as the end-to-end object detector to process each video sequence $V = \{\mathbf{X}_t\}_{t=1}^{T}$, where $\mathbf{X}_t \in \mathbb{R}^{H\times W\times 3}$ denotes the RGB image at frame t. A CNN backbone [15] and a Transformer encoder [35] are employed to extract and enhance image features $f_t \in \mathbb{R}^{h\times w\times c}$. Subsequently, the DETR decoder transforms detection queries into object-level predictions, including bounding box $bbox = (c^i_{x,t}, c^i_{y,t}, w^i_t, h^i_t)$ and class index. Deformable DETR directly provides decoded embeddings as object-level features f^i_t, eliminating the need for post-processing steps such as RoI cropping or feature alignment, obtaining dimension consistent representations across frames.

To accommodate video sequences with varying resolutions and aspect ratios, we normalize the bounding box parameters as $(c^i_{x,t}/W, c^i_{y,t}/H, w^i_t/W, h^i_t/H)$, where W and H denote the frame width and height, respectively. During training, the velocity of object i is computed by differencing the normalized center coordinates across consecutive frames, yielding Cartesian velocity components $(v^i_{x,t}, v^i_{y,t})$. As polar coordinates explicitly decouple two identity-relevant factors:

magnitude and direction [10], we convert the velocity representation from Cartesian $(v^i_{x,t}, v^i_{y,t})$ to polar coordinates(ρ^i_t, θ^i_t). This formulation decouples motion magnitude and direction, reducing entanglement between velocity components and facilitating more stable representation. As a result, the entire geometric state vector $g^i_t = (g^i_{p,t}, g^i_{s,t}, g^i_{v,t})$ for object i at frame t is thus defined as:

$$g^i_t = \left(\frac{c^i_{x,t}}{W}, \frac{c^i_{y,t}}{H}, \frac{w^i_t}{W}, \frac{h^i_t}{H}, \rho^i_t, \theta^i_t \right), \tag{1}$$

where g^i_t represents the geometric vector of object i at frame t, including its position, size, and velocity.

3.2 Hierarchical Geometry Quantization

Given that the constructed geometry vectors exhibit heterogeneous scales and units across dimensions, we perform residual quantization separately on position $g^i_{p,t}$, scale $g^i_{s,t}$, and velocity $g^i_{v,t}$. Each component is projected into a latent space using a multi-layer perceptron, resulting in embeddings ${z_{g\in(p,s,v)}}^i_t \in \mathbb{R}^D$.

To better capture the fluctuating geometric cues, we design a hierarchical quantization strategy that discretizes the MLP-projected continuous embeddings. Specifically, for position, size, and velocity components, we apply a two-stage residual codebook to perform quantization.

$$E^{(\ell)}_{g\in(p,s,v)} = \{e^{(\ell)}_{g,1}, \ldots, e^{(\ell)}_{g,K}\} \subset \mathbb{R}^D, \quad \ell = 1, 2, \tag{2}$$

where D denotes the embedding dimension and K represents the number of codewords per stage. We set $D = 64$ and $K = 64$ to balance representational capacity and computational efficiency. A detailed analysis of the hyperparameter design is provided in Sec. 4.4.

$$\begin{cases} {k_1}^i_{g,t} = \underset{k_1\in\{1,\ldots,K\}}{\arg\min} \left\| z^i_{g,t} - e^{(1)}_{g,k_1} \right\|_2, & q_{g,1} = e^{(1)}_{g,k_1}, \\ r^i_{g,t} = z^i_{g,t} - q_{g,1}, & \\ {k_2}^i_{g,t} = \underset{k_2\in\{1,\ldots,K\}}{\arg\min} \left\| r^i_{g,t} - e^{(2)}_{g,k_2} \right\|_2, & q_{g,2} = e^{(2)}_{g,k_2}. \end{cases} \tag{3}$$

The residual formulation allows the second codebook to capture fine-grained geometric variations that persist after coarse quantization. The final geometric embedding $\tilde{z}^i_{g,t}$ is obtained by summing the coarse and fine codewords, producing a more detailed and robust representation [17]. The discrete embeddings $\tilde{z}^i_{p,t}$, $\tilde{z}^i_{s,t}$, and $\tilde{z}^i_{v,t}$ are then concatenated into a 3D-dimensional vector and linearly projected to $\tilde{z}^i_t \in \mathbb{R}^d$ to ensure compatibility with the visual features.

3.3 Geometry-Aware Animal Tracking

In addition to the object-level visual embedding f_t^i extracted in Sec. 3.1, we follow [11] and incorporate identity information by projecting a one-hot encoded ID index into an embedding l^m for trajectory m. We further integrate the discretized geometric embedding $\tilde{z}_t^i$ to jointly model appearance, identity, and geometric states. All embeddings f_t^i, l^m, and $\tilde{z}_t^i$ are projected to dimension d. At frame t, the historical representation of trajectory m over time window $[t-T, t-1]$ is defined as:

$$F_{t-T:t-1}^m = \left\{ \left(f_{t-T}^m, \tilde{z}_{t-T}^m, l^m \right), \ldots, \left(f_{t-1}^m, \tilde{z}_{t-1}^m, l^m \right) \right\}. \tag{4}$$

Each detected object in the current frame is represented as $o_t^i = (f_t^i, \tilde{z}_t^i, l_{\text{init}}^i)$. A Transformer decoder performs cross-attention between the current object query o_t^i and historical trajectory features $F_{t-T:t-1}^m$. The concatenation of visual, geometric, and identity embeddings preserves modality-specific information before attention interaction. By incorporating discretized geometric tokens, the attention mechanism becomes less sensitive to small coordinate perturbations, leading to more stable identity matching across frames. The attention output is combined with o_t^i and passed through a classification head for identity prediction.

3.4 Training

Loss Function. Following prior works [25, 34, 46], we formulate trajectory association as an identity classification task that includes an additional unknown category. Beyond the standard detection loss $\mathcal{L}_{\text{det}}$ from the DETR, we incorporate an ID classification loss $\mathcal{L}_{\text{id}}$ and a vector quantization loss $\mathcal{L}_{\text{vq}}$ for training process. These components are integrated into a unified objective via a weighted sum, where each coefficient λ_i modulates the contribution of the corresponding loss term $\mathcal{L}_i$. The overall training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{det}} \mathcal{L}_{\text{det}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{vq}} \mathcal{L}_{\text{vq}}, \tag{5}$$

where $\mathcal{L}_{\text{det}}$ consists of a classification loss, a Generalized IoU loss, and an L1 regression loss. For the ID classification loss $\mathcal{L}_{\text{id}}$, we adopt the standard cross-entropy formulation. The vector quantization loss $\mathcal{L}_{\text{vq}}$ is defined as $\|\text{sg}[z] - e\|_2^2 + \beta \|z - \text{sg}[e]\|_2^2$, where $\text{sg}[\cdot]$ denotes the stop-gradient operator. The first term updates the codewords e, while the second term encourages the encoder output z to commit to its assigned codeword [37, 42]. To enable gradient backpropagation through the non-differentiable nearest-neighbor operation, we adopt the straight-through estimator (STE) [28].

3.5 Inference

During inference, object detections in the current frame lack explicit temporal correspondences with the previous frame, which prohibits direct velocity estimation via frame-to-frame differencing. For each detection with normalized center

$\mathbf{c}_t^i = (c_{x,t}^i, c_{y,t}^i)$, we estimate its velocity using a lightweight KNN-based regression over historical trajectory states. Specifically, we retrieve the k nearest historical states $\mathcal{N}_k$ in spatial coordinate space and compute the estimated velocity as $\hat{\mathbf{v}}_t^i = \sum_{j \in \mathcal{N}_k} w_{ij} \mathbf{v}_j$, where the normalized weights are defined as

$$w_{ij} = \frac{\exp\left(-\|\mathbf{c}_t^i - \mathbf{c}_j\|_2^2 / \sigma_s^2\right) \cdot \exp(\cos(\Delta\phi_{ij})/\sigma_d)}{\sum_{l \in \mathcal{N}_k} \exp\left(-\|\mathbf{c}_t^i - \mathbf{c}_l\|_2^2 / \sigma_s^2\right) \cdot \exp(\cos(\Delta\phi_{il})/\sigma_d)}. \tag{6}$$

Here $\mathbf{v}_j$ denotes historical velocity vectors and $\Delta\phi_{ij}$ denotes the angular difference between the motion direction of the current detection and the historical state j. The spatial kernel encourages proximity consistency, while the directional term favors motion-aligned trajectories.

4 Experiments

4.1 Datasets and Metrics

Datasets. To comprehensively evaluate multi-animal tracking, we select a set of challenging and representative benchmarks spanning diverse animal study scenarios. Animaltrack [44] is a multi-class animal tracking dataset featuring complex inter-individual interactions, consisting of 58 sequences collected from wild animals across 10 terrestrial and aquatic categories. BFT [48] is a dataset of 22 bird species with diverse morphological characteristics and irregular motion patterns, composed of 106 videos. BuckTales [27] is a UAV-based animal tracking dataset featuring wild antelopes with highly similar visual appearances. Due to incomplete annotations in the full release, we evaluate on the subset with complete identity labels (12 videos, 680 trajectories). These benchmarks cover a broad spectrum of representative challenges in multi-animal tracking. This diverse setting enables a comprehensive evaluation of HieDG in terms of both performance and robustness across complex animal tracking scenarios.

To further assess the generalization of our study beyond animal-specific scenarios, we conduct experiments on two widely adopted multi-person tracking benchmarks, DanceTrack [33] and SportsMOT [7]. Both datasets exhibit extreme intra-class similarity and complex motion patterns, posing substantial challenges for identity association.

Metrics. We adopt Higher Order Tracking Accuracy (HOTA) [22] to jointly evaluate detection (DetA) and association (AssA) performance. MOTA [2] and IDF1 [31] are also reported to cover complementary tracking aspects. Since our method primarily targets identity association, we place particular attention on HOTA, AssA, and IDF1, while reporting MOTA for completeness.

4.2 Implementation Details

Given its simplicity and generality, we adopt MOTIP [11] as the baseline for this study. Specifically, we employ a ResNet-based [15] Deformable DETR [50] as the detector in the proposed HieDG framework. To maintain consistency with prior work, we initialize the model using COCO [19] pretrained weights. Furthermore, we set the dimension of each codebook and each geometric component to $D = 64$,

Table 1: Performance comparison on the AnimalTrack [44] and BFT [48] test sets. * denotes the use of continuous geometric cues. The best and second-best performance are highlighted in **bold** and underline, respectively.

Methods	AnimalTrack [44]					BFT [48]				
	HOTA	DetA	AssA	MOTA	IDF1	HOTA	DetA	AssA	MOTA	IDF1
Heuristic-based										
JDE [38]	28.7	40.6	20.3	28.3	33.2	30.7	40.9	23.4	35.4	37.4
CSTrack [18]	–	–	–	–	–	33.2	47.0	23.7	46.7	34.5
FairMOT [46]	30.6	–	–	29.0	38.8	40.2	53.3	28.2	56.0	41.8
TADAM [13]	32.5	–	–	36.5	37.2	–	–	–	–	–
DeepSORT [24]	36.4	41.2	32.1	41.0	35.2	–	–	–	–	–
SORT [4]	42.5	48.8	37.0	54.8	49.1	61.2	60.6	62.3	75.5	77.2
IOUTrack [5]	41.6	–	–	55.7	45.7	–	–	–	–	–
Tracktor++ [1]	44.2	–	–	55.2	51.0	–	–	–	–	–
QDtrack [29]	47.0	–	–	55.7	56.3	–	–	–	–	–
TransCenter [40]	–	–	–	–	–	60.0	66.0	61.1	74.1	72.4
CenterTrack [49]	–	–	–	–	–	56.2	58.5	54.0	60.2	61.0
ByteTrack [45]	48.1	46.9	49.3	56.6	55.5	62.5	61.2	64.1	77.2	82.3
OC-SORT [6]	–	–	–	–	–	66.8	65.4	68.7	77.1	79.3
Query-based										
TrackFormer [25]	31.0	–	–	20.4	36.5	63.3	66.0	61.1	74.1	72.4
TransTrack [34]	45.4	–	–	48.3	53.4	65.1	68.5	61.8	74.9	73.1
MOTR [43]	49.5	51.3	47.7	57.2	54.1	64.2	64.4	64.7	74.1	75.2
MeMOTR [12]	52.2	54.3	50.1	62.1	58.6	–	–	–	–	–
MOTIP [11]	54.1	54.2	55.0	61.9	61.7	69.2	**71.1**	67.3	**78.6**	80.1
CO-MOT [41]	55.3	**55.1**	55.5	**62.8**	62.1	69.0	70.5	67.7	78.3	79.0
Ours										
HieDG*	54.6	55.0	54.2	62.4	61.8	69.5	70.4	68.8	77.2	79.8
HieDG	**56.2**	54.9	**58.4**	62.5	**64.4**	**71.3**	70.6	**72.2**	77.4	**82.5**

while the hidden dimension of the quantized geometry embedding $\tilde{\mathbf{z}}$ is aligned with the Deformable DETR backbone, i.e., $C = 256$.

Following prior work [11, 12, 47], we apply standard image-level augmentations, including random resizing, horizontal flipping, and color jittering, for an apples-to-apples comparison. To improve robustness to estimation noise at inference, we apply a 0.3 dropout to velocity embeddings during training, which regularizes the model against potential motion uncertainty. All experiments are implemented in PyTorch and conducted on NVIDIA Tesla V100 GPUs. We use 4 GPUs for training on the AnimalTrack, DanceTrack, and SportsMOT datasets, and 2 GPUs for BFT and BuckTales.

We adopt the AdamW optimizer with a weight decay of 5×10^{-4} and an initial learning rate of 10^{-4}. A linear warmup is applied for the first epoch to stabilize learning rate adjustment. The loss weight coefficients are set to $\lambda_{\text{det}} = 1.0$, $\lambda_{\text{id}} = 2.0$, and $\lambda_{\text{vq}} = 0.1$.

4.3 Comparisons with State-of-the-art Methods

We evaluate the proposed HieDG in comparison with representative heuristic-based and query-based trackers on the AnimalTrack [44], BFT [48], BuckTales [27], DanceTrack [33], and SportsMOT [7] benchmarks. The quantitative results are illustrated in Tab. 1, Tab. 2, and Tab. 3. Since association mechanisms are inherently coupled with detection backbones, we report results following each method's official evaluation protocols and configurations.

AnimalTrack and BFT. We evaluate HieDG on standard multi-animal tracking datasets in AnimalTrack and BFT across terrestrial, underwater, and aerial environments, where individuals frequently exhibit uniform appearance, high density, and irregular movements. Heuristic-based methods (e.g., SORT [4], DeepSORT [24], and OC-SORT [6]) rely on handcrafted motion and association rules. Although OC-SORT performs better at handling abrupt nonlinear motion in BFT for bird tracking, such strategies remain limited under complex multi-animal dynamics. Query-based approaches demonstrate greater adaptability, yet HieDG consistently outperforms existing methods and achieves state-of-the-art association performance, yielding clear gains in HOTA, AssA, and IDF1. A detailed comparison reveals a consistent trend: directly incorporating continuous geometric representation (HieDG*) degrades association accuracy relative to the baseline, whereas discretizing geometric embedding consistently yields performance gains. This contrast suggests that continuous geometry can be unreliable in crowded and appearance-ambiguous scenarios, where it may introduce substantial noise. By contrast, discretization converts raw geometry into stable structural tokens, thereby enhancing association robustness. These results suggest that discretized geometric modeling provides improved robustness in complex multi-animal tracking scenarios.

Table 2: Performance comparison on the BuckTales [27] test set. * denotes the use of continuous geometric cues. Best and second-best results are highlighted in **bold** and underline, respectively.

Methods	HOTA	DetA	AssA	MOTA	IDF1
Heuristic-based					
SORT [4]	30.1	40.7	24.1	42.9	47.5
DeepSORT [24]	9.62	25.6	3.8	16.3	56.5
ByteTrack [45]	<u>49.8</u>	**47.1**	<u>52.6</u>	**49.6**	<u>64.0</u>
OC-SORT [6]	42.4	45.1	39.8	<u>47.3</u>	57.9
Query-based					
TrackFormer [25]	41.7	<u>45.8</u>	37.9	41.4	57.2
MOTR [43]	42.5	44.0	41.1	41.8	58.6
MOTIP [11]	45.6	42.2	50.4	43.9	61.8
CO-MOT [41]	47.2	43.6	51.3	45.9	62.3
Ours					
HieDG*	41.3	43.2	39.4	44.9	56.6
HieDG	**52.4**	42.9	**64.0**	44.6	**69.7**

BuckTales. UAV-based remote sensing, widely used in wildlife monitoring, introduces additional challenges for multi-animal tracking. Aerial viewpoints yield small, blurry targets, while frequent camera shake and abrupt viewpoint shifts destabilize both appearance and spatial relationships, making identity association significantly more difficult. As shown in Tab. 2, ByteTrack achieves competitive performance in this UAV setting, suggesting that motion-driven heuristics can remain effective when appearance cues collapse. Nevertheless, HieDG achieves the best overall association performance. The improvements in HOTA, AssA, and IDF1 indicate stronger identity consistency under severe geometric perturbations. These findings are consistent with the

Table 3: Performance comparison on the DanceTrack [33] and SportsMOT [7] test sets. * denotes the use of continuous geometric cues. The best and second-best performance are highlighted in **bold** and underline, respectively.

Methods	DanceTrack [33]					SportsMOT [7]				
	HOTA	DetA	AssA	MOTA	IDF1	HOTA	DetA	AssA	MOTA	IDF1
Heuristic-based										
FairMOT [46]	39.7	66.7	23.8	82.2	40.8	49.3	70.2	34.7	86.4	53.5
QDtrack [29]	54.2	80.1	36.8	87.7	50.4	60.4	77.5	47.2	90.1	62.3
ByteTrack [45]	47.7	71.0	32.1	89.6	53.9	62.1	76.5	50.5	**93.4**	69.1
OC-SORT [6]	55.1	80.3	38.3	**92.0**	54.6	68.1	**84.8**	54.8	**93.4**	68.0
Query-based										
TrackFormer [25]	–	–	–	–	–	63.3	66.0	61.1	74.1	72.4
MeMOTR [12]	63.4	77.0	52.3	85.4	65.5	68.8	82.0	57.8	90.2	69.9
TransTrack [34]	45.5	75.9	27.5	88.4	45.2	68.9	82.7	57.5	92.6	71.5
CO-MOT [41]	65.3	80.1	53.5	89.3	66.5	–	–	–	–	–
MOTIP [11]	69.6	80.4	60.4	90.6	74.7	72.6	83.5	63.2	92.4	77.1
Ours										
HieDG*	68.5	**80.8**	60.0	91.4	73.1	71.4	82.7	59.6	92.5	71.7
HieDG	**70.3**	80.7	**61.8**	90.9	**75.4**	**72.8**	83.4	**63.9**	92.1	**77.5**

observations on AnimalTrack and BFT, further demonstrating the robustness of hierarchical geometric modeling in extreme aerial tracking scenarios.

DanceTrack and SportsMOT. To evaluate generalization in generic low-discriminability multi-person tracking scenarios, we report results on DanceTrack and SportsMOT in Tab. 3. HieDG achieves superior association performance across both benchmarks. The consistent gains in AssA and IDF1 demonstrate improved identity preservation in crowded and appearance-ambiguous scenes. These results indicate that the proposed geometry-guided modeling strategy generalizes beyond animal-specific tracking and remains effective in broader multi-object tracking settings.

4.4 Ablations

We conduct ablation studies on the BFT [48] dataset, which exhibits dense interactions, strong appearance similarity, and irregular motion patterns. These characteristics make BFT a suitable testbed for validating the effectiveness of each component in our framework under low-discriminability conditions. For clarity, we denote HieDG* as the variant using continuous geometric embeddings without quantization.

Geometric Component. Heuristic-based methods typically leverage geometric cues, such as position and scale, from detectors to guide tracking. To capture the fast yet coherent motion patterns in animal scenarios, we model velocity based on inter-frame position differences. As shown in Tab. 4, each geometric cue independently improves association performance, while combining all three

Table 4: Impact of different geometric components. The best performance is shown in **bold**, and the gray background is the choice for our final experiment.

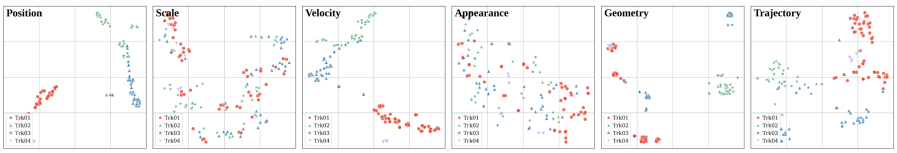
Pos.	Scale	Vel.	HOTA	DetA	AssA	MOTA	IDF1
			69.2	71.1	67.3	78.6	80.1
✓			70.6	70.5	71.0	76.9	81.6
	✓		70.1	71.3	68.9	78.9	81.3
		✓	69.7	**71.8**	68.0	**79.1**	80.9
✓	✓	✓	**71.3**	70.6	**72.2**	77.4	**82.5**

Table 5: Impact of hierarchical geometric quantization strategies. The best performance is shown in **bold**, and the gray background is the choice for our final experiment.

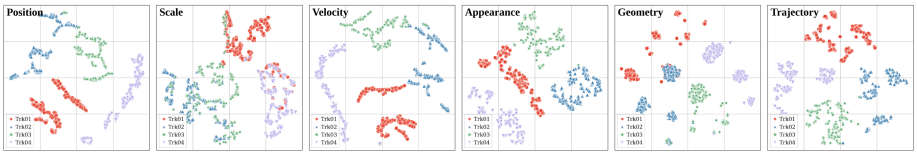
Geo. Emb.	HOTA	AssA	IDF1	Para. (K)	FLOPs (K)
Cont. Enc.	69.5	68.8	79.8	+**199.2**	+198.1
FixVQ	69.8	69.1	80.2	+49.9	+99.0
1-Stage VQ	70.5	68.9	81.3	+99.1	+197.4
2-Stage VQ	71.3	**72.2**	82.5	+148.3	+295.6
3-Stage VQ	**71.5**	**72.2**	**82.7**	+197.4	+**394.0**

yields the highest performance gain. These results demonstrate that complementary geometric cues provide additive benefits for identity association, and that jointly modeling position, scale, and velocity yields the most stable cross-frame matching.

We visualize the t-SNE [23] of trajectory embeddings for all frames in sequence Fa5001 and An1005, as shown in Fig. 3, where each color and shape denotes a unique ID. The first three columns illustrate embeddings derived from individual geometric components (position, scale, and velocity), whereas the last three columns show appearance features, discretized geometry embeddings, and the fused trajectory representation. While appearance features show moderate separability, the quantized geometric embeddings exhibit tighter intra-ID clustering compared to continuous appearance features. The fused embeddings further improve inter-ID separation, suggesting that discretized geometry stabilizes identity representation across frames.



(a) t-SNE [23] visualization of all frames in Fa5001 (BFT).



(b) t-SNE [23] visualization of all frames in An1005 (BFT).

Fig. 3: t-SNE [23] visualization of track embeddings across all frames in the BFT dataset. Different IDs are marked by distinct colors and shapes.

Table 6: Impact of different codebook embedding dimensions. The best performance is shown in **bold**, and the gray background is the choice for our final experiment.

Code. Dim.	HOTA	AssA	IDF1	Para. (K)	FLOPs (K)
16	70.2	70.3	80.7	+74.6	+148.2
32	70.8	71.7	81.6	+99.1	+197.4
64	71.3	**72.2**	82.5	+148.3	+295.6
128	**71.4**	**72.2**	**82.7**	**+246.6**	**+492.3**

Table 7: Impact of Hungarian matching and polar coordinate representation. The best performance is shown in **bold**, and the gray background is the choice for our final experiment.

Hung.	Polar vel.	HOTA	DetA	AssA	MOTA	IDF1
		69.8	**71.3**	68.3	**78.8**	80.7
✓		69.9	71.1	68.7	78.5	81.0
	✓	**71.3**	70.6	**72.2**	77.4	**82.5**
✓	✓	71.1	70.7	71.5	77.9	81.9

Geometric Vector Quantization. Discretizing geometric cues into stable representations significantly enhances tracking performance. To evaluate our quantization design, we compare it with fixed-bin quantization and explore different depths of quantization. As shown in Tab. 5, fixed-bin quantization yields only limited gains, while hierarchical residual quantization consistently improves association performance. Although extending from two to three stages slightly increases HOTA and IDF1, the improvement is marginal relative to the additional 49.1K parameters and 98.4K FLOPs. This disproportionate increase in computational cost does not justify the minor accuracy gain. Therefore, we adopt the two-stage design as a favorable trade-off between performance and efficiency.

To assess the impact of codebook size, we set equal dimensions for the position, scale, and velocity codebooks to ensure a fair comparison. As shown in Tab. 6, increasing the codebook size to 64 substantially improves association performance with moderate computational overhead. However, further scaling to 128 offers marginal gains while introducing significant additional cost. These results suggest that our two-stage residual quantization sufficiently captures fluctuating geometry in animal scenarios, rendering larger codebooks less beneficial.

Hungarian Algorithm and Velocity Representation. Inspired by [11], although the Hungarian algorithm is widely used for global ID assignment, our association strategy (detailed in Sec. 3.3) is formulated independently of this matching mechanism. To enhance robustness to low-speed jitter and rapid directional changes common in animal scenarios, we model velocity in polar coordinates. The ablation results in Tab. 7 show that incorporating Hungarian matching brings negligible improvement, indicating that the ID classification-based formulation already provides effective global assignment within the learned framework. In contrast, modeling velocity in polar coordinates consistently improves AssA and IDF1, suggesting that decoupling magnitude and direction provides a more stable motion representation under irregular animal dynamics.

5 Conclusion

We propose HieDG, a hierarchical discrete geometry-guided framework for multi-animal tracking. By reformulating continuous geometric cues into structured dis-

crete embeddings through residual quantization, HieDG stabilizes cross-frame identity association within attention-based tracking architectures. Extensive experiments demonstrate consistent improvements across diverse multi-animal benchmarks, with competitive generalization to generic tracking datasets. Overall, these findings suggest that discretized geometric representations can improve the stability of attention-based identity modeling under low-discriminability and noisy motion conditions.

6 Acknowledgement

This work was supported by the National Key Research and Development Program of China under Grant 2023YFC3208303; the Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project under Grants 2022ZD0211900 and 2022ZD0211902; the National Natural Science Foundation of China (NSFC) under Grant 82471245 from Hang Zhou, Grant 12301651 from Weifu Li, and Grant 62273347 from Yifan Zhang; the Shenzhen Natural Science Foundation under Grant JCYJ20241202130204005; the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010844; and the Characteristic Innovation Project of Guangdong Higher Education Institutions (Natural Science) under Grant 2024KTSCX027.

References

1. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (October 2019)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**(1), 246309 (2008)
3. Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H.E., Van Oosterhout, C.: Genetic load: genomic estimates and applications in non-model animals. Nature Reviews Genetics **23**(8), 492–503 (2022)
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: Proceedings of the International Conference on Image Processing. pp. 3464–3468 (2016)
5. Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 1–6 (2017). https://doi.org/10.1109/AVSS.2017.8078516
6. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: 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 (June 2023)
7. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9921–9931 (October 2023)

8. Deng, C., Ji, L., Zhang, C., Yan, H., Zhang, Z., Wang, L., Zhang, J.: Pgmm: Prior guided multi-expert model for long-tailed classification. Pattern Recognition **179**, 113669 (2026). https://doi.org/https://doi.org/10.1016/j.patcog.2026.113669, https://www.sciencedirect.com/science/article/pii/S0031320326006345
9. Deng, C., Li, D., Ji, L., Zhang, C., Li, B., Yan, H., Zheng, J., Wang, L., Zhang, J.: Chatdiff: A chatgpt-based diffusion model for long-tailed classification. Neural Networks **181**, 106794 (2025). https://doi.org/https://doi.org/10.1016/j.neunet.2024.106794, https://www.sciencedirect.com/science/article/pii/S0893608024007184
10. Fiquet, P.É.H., Simoncelli, E.P.: A polar prediction model for learning to represent visual transformations. In: Advances in Neural Information Processing Systems (2023), https://openreview.net/forum?id=hyPUZX03Ks
11. Gao, R., Qi, J., Wang, L.: Multiple object tracking as id prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27883–27893 (June 2025)
12. Gao, R., Wang, L.: Memotr: Long-term memory-augmented transformer for multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9901–9910 (October 2023)
13. Guo, S., Wang, J., Wang, X., Tao, D.: Online multiple object tracking with cross-task synergy. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 8132–8141 (2021)
14. Han, G., Lim, S.N.: Few-shot object detection with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28608–28618 (June 2024)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
16. Jetz, W., Tertitski, G., Kays, R., Mueller, U., Wikelski, M., Åkesson, S., Anisimov, Y., Antonov, A., Arnold, W., Bairlein, F., et al.: Biological earth observation with animal sensors. Trends in Ecology & Evolution **37**(4), 293–298 (2022)
17. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11523–11532 (June 2022)
18. Liang, C., Zhang, Z., Zhou, X., Li, B., Zhu, S., Hu, W.: Rethinking the competition between detection and reid in multiobject tracking. IEEE Transactions on Image Processing **31**, 3182–3196 (2022)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755. Springer (2014)
20. Liu, D., Li, W., Ma, C., Zheng, W., Yao, Y., Tso, C.F., Zhong, P., Chen, X., Song, J.H., Choi, W., et al.: A common hub for sleep and motor control in the substantia nigra. Science **367**(6476), 440–445 (2020)
21. Liu, Y., Li, W., Liu, X., Li, Z., Yue, J.: Deep learning in multiple animal tracking: A survey. Computers and Electronics in Agriculture **224**, 109161 (2024)
22. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International Journal of Computer Vision **129**(2), 548–578 (2021)
23. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (nov 2008)

24. Maggiolino, G., Ahmad, A., Cao, J., Kitani, K.: Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In: Proceedings of the International Conference on Image Processing. pp. 3025–3029 (2023)
25. Meinhardt, T., Kirillov, A., Leal-Taixé, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8844–8854 (June 2022)
26. Mills-Tettey, G.A., Stentz, A., Dias, M.B.: The dynamic hungarian algorithm for the assignment problem with changing costs. Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27 **7** (2007)
27. Naik, H., Yang, J., Das, D., Crofoot, M.C., Rathore, A., Sridhar, V.H.: Bucktales: A multi-uav dataset for multi-object tracking and re-identification of wild antelopes. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 81992–82009. Curran Associates, Inc. (2024)
28. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
29. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 164–173 (June 2021)
30. Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., Papadoyannis, E.S., Normand, E., Deutsch, D.S., Wang, Z.Y., et al.: Sleap: A deep learning system for multi-animal pose tracking. Nature Methods **19**(4), 486–495 (2022)
31. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European Conference on Computer Vision. pp. 17–35. Springer (2016)
32. Shim, K., Ko, K., Yang, Y., Kim, C.: Focusing on tracks for online multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11687–11696 (June 2025)
33. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20993–21002 (June 2022)
34. Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
36. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: Yolov10: Real-time end-to-end object detection. In: Advances in Neural Information Processing Systems. vol. 37, pp. 107984–108011. Curran Associates, Inc. (2024)
37. Wang, J., Jiang, Y., Yuan, Z., PENG, B., Wu, Z., Jiang, Y.G.: Omnitokenizer: A joint image-video tokenizer for visual generation. In: Advances in Neural Information Processing Systems (2024)
38. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Proceedings of the European Conference on Computer Vision. pp. 107–122. Springer (2020)

39. Welch, G., Bishop, G., et al.: An introduction to the kalman filter (1995)
40. Xu, Y., Ban, Y., Delorme, G., Gan, C., Rus, D., Alameda-Pineda, X.: Transcenter: Transformers with dense representations for multiple-object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7820–7835 (2023)
41. Yan, F., Luo, W., Zhong, Y., Gan, Y., Ma, L.: CO-MOT: Boosting end-to-end transformer-based multi-object tracking via coopetition label assignment and shadow sets. In: International Conference on Learning Representations (ICLR) (2025)
42. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved VQGAN. In: Proceedings of the International Conference on Learning Representations (2022)
43. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: Proceedings of the European Conference on Computer Vision. pp. 659–675. Springer (2022)
44. Zhang, L., Gao, J., Xiao, Z., Fan, H.: Animaltrack: A benchmark for multi-animal tracking in the wild. International Journal of Computer Vision **131**(2), 496–513 (2023)
45. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: Proceedings of the European Conference on Computer Vision. pp. 1–21. Springer (2022)
46. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021)
47. Zhang, Y., Wang, T., Zhang, X.: Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 22056–22065 (2023)
48. Zheng, G., Lin, S., Zuo, H., Fu, C., Pan, J.: Nettrack: Tracking highly dynamic objects with a net. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19145–19155 (June 2024)
49. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: Proceedings of the European Conference on Computer Vision. pp. 474–490. Springer (2020)
50. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable {detr}: Deformable transformers for end-to-end object detection. In: Proceedings of the International Conference on Learning Representations (2021)