Diffusion Models in Video Object Tracking

Bartłomiej Olechno

Agenda

- What are diffusion models?
- What is object tracking? (taxonomy)
- Challenges in video tracking
- Metrics of diffusion tracking
- Advantages of diffusion models
- Overview of recent methods
- Results and analysis
- Limitations and future directions

What are Diffusion Models?

- Generative models that learn to reverse a noise-adding process.
- Two-phase approach:
  - **Forward process**: progressively adds noise to data.
  - **Reverse process**: refines step-by-step to generate samples.
- Used for image, video, and trajectory generation.

What is Object Tracking?

- **Object tracking** involves locating and following the position of a target across video frames.
- Types:
  - **Single Object Tracking (SOT)** – one target, known in the first frame.
  - **Multiple Object Tracking (MOT)** – multiple targets, often with detection.
- Common approaches:
  - Tracking-by-detection **TBD**: detection + association between frames.
  - End-to-end learning **JDT**: directly predict object trajectories.

Baseline: Tracking-by-Detection (TBD) Pipeline

Tracking-by-Detection (TBD) is a two-stage tracking framework:

1. **Object Detection:**
   - A detector (e.g., YOLO, DETR) identifies objects in each frame independently.
2. **Data Association:**
   - Detected objects are linked across frames to form trajectories.
   - Common techniques:
     - IoU-based matching
     - Motion models (Kalman filter)
     - Appearance models (Re-ID embeddings)
     - Hungarian algorithm for assignment

Advantages: Modular, scalable, compatible with pretrained detectors.
Limitations: Sensitive to detector errors and occlusions; no end-to-end learning.

Challenges in Video Tracking

- Changes in lighting, occlusions, camera motion.
- Nonlinear and unpredictable object motion.
- Need for spatiotemporal context modeling.

Why Use Diffusion Models?

- Naturally captured uncertainty in object motion.
- Efficient for modeling of trajectory distributions.
- Enable generation of alternative future scenarios.
- Foundation models understand spatiotemporal context.

Overview of Methods

- **DiffusionTrack** – one shot JTD approach. [6]
- **DiffMOT** – models motion in latent space. [2]

DiffusionTrack

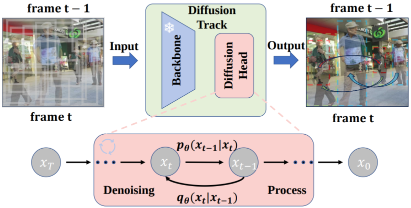

Figure: DiffusionTrack formulates object association as a denoising diffusion process from paired noise boxes to paired object boxes within two adjacent frames t 1 and t. The diffusion head receives the two-frame image information extracted by the frozen backbone and then iteratively denoises the paired noise boxes to obtain the final paired object boxes.

DiffMOT

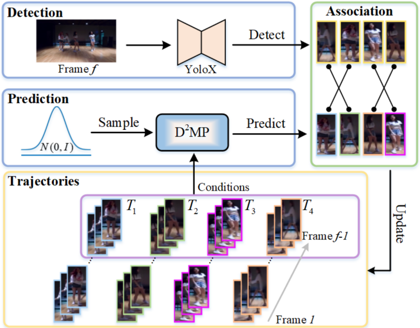

Figure: DIffMOT integrates diffusion process into TBD approach. Standard approaches uses linear motion models like the Kalman Filter. The authors developed Decoupled Diffusion-based Motion Predictor, which increases performance of the TBD approach.

Tracking Evaluation Metrics

1. **IDF1 (ID F1 Score)** [1]
   - Evaluates how well the tracker preserves object identities over time.
   - Harmonic mean of ID precision and recall:

$$\text{IDF1} = \frac{2 \cdot \text{ID Precision} \cdot \text{ID Recall}}{\text{ID Precision} + \text{ID Recall}}$$

   - Sensitive to identity switches.

2. **MOTA (Multiple Object Tracking Accuracy)** [1]
   - Captures overall tracking quality, combining different error types.
   - $$\text{MOTA} = 1 - \frac{\text{FN} + \text{FP} + \text{IDSW}}{\text{GT}}$$
   - Penalizes false negatives, false positives, and identity switches.

HOTA (Higher Order Tracking Accuracy)

What is HOTA? [3]

- A modern tracking metric that jointly evaluates association accuracy and detection accuracy

$$\text{DetA} = \frac{TP}{TP + 0.5 \cdot (FP + FN)}$$

$$\text{AssA} = \frac{TP_{assoc}}{TP_{assoc} + 0.5 \cdot (FP_{assoc} + FN_{assoc})}$$

$$\text{HOTA} = \sqrt{\text{DetA} \times \text{AssA}}$$

Why HOTA over MOTA?

- MOTA mixes false positives, false negatives, and ID switches but does not reflect identity quality.
- HOTA evaluates **entire trajectories**, not just frame-level matches.

Tracking Example: Why HOTA Outperforms MOTA

Frame	GT A	GT B	Tracker 1	Tracker 2
1	A	B	A	B
2	A	B	A	B
3	A	B	B (!)	B
4–10	A	B	B	B

- **Tracker 1:** One ID switch in frame 3, but then tracks A as B for the rest of the sequence.
- **Tracker 2:** Maintains correct identities throughout.
- **MOTA:** Both may have similar score (only 1 IDSW).
- **HOTA (AssA):** Tracker 1 heavily penalized for losing identity across 7 frames.

Experimental Results

Method	HOTA (MOT17)
GTR (no diffusion SOTA JTD)	59.1
DiffusionTrack	60.8
DiffMOT	64.5
BoT-SORT (popular no diffusion TBD)	64.6 (without ReID)

Table: Comparison of HOTA scores for different tracking methods on MOT17 [4].

Experimental Results

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
DiffusionTrack	52.4	47.5	33.5	89.5	82.2
BoT-SORT (w/ CMC)	56.9	57.7	–	91.7	–
BoT-SORT (w/o CMC)	58.8	58.9	–	93.4	–
Deep OC-SORT	61.3	61.5	45.8	92.3	82.2
DiffMOT	**62.3**	**63.0**	**47.2**	**92.8**	**82.5**

Table: Evaluation on the DanceTrack, source [5]

Pros and Cons

Advantages:

- Better uncertainty modeling.
- Performs well in complex tracking scenarios.
- Enables generative trajectory sampling.

Limitations:

- High computational cost.
- Slower inference times.

Future Research Directions

- Spatiotemporal diffusion modeling.
- Deeper research on generalization across various datasets.
- Research on fusion of diffusion models with SOTA algorithms.

Conclusion

- Diffusion models are a promising new approach to object tracking.
- They improve the handling of uncertainty and motion complexity.
- The performance is still worse than other models.

Questions?

Thank you for your attention!

Any questions?

Bibliografia I

[1] Keni Bernardin and Rainer Stiefelhagen. "Evaluating multiple object tracking performance: the CLEAR MOT metrics". In: *EURASIP Journal on Image and Video Processing*. 1. Springer, 2008, pp. 1–10.

[2] Messaoud Chaabane, Richard J Radke, and Xitong Li. "DiffMOT: Diffusion Models for Robust Multiple Object Tracking". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

[3] Jonathon Luiten et al. "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking". In: *International Journal of Computer Vision (IJCV)* 130 (2022), pp. 1103–1121.

[4] Anton Milan et al. "MOT16: A Benchmark for Multi-Object Tracking". In: *arXiv preprint arXiv:1603.00831*. 2016.

Bibliografia II

[5] Peize Sun et al. "DanceTrack: Multi-Person Tracking in Uniform Appearance and Diverse Motion". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[6] Yuchao Yang et al. "DiffusionTrack: Diffusion Model for Multiple Object Tracking". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023.