DM^{3}T: Harmonizing Modalities via Diffusion for Multi-Object Tracking

Supplementary Material

6. Qualitative Visualzation

We provide qualitative comparisons against the strong baseline PFTrack in Fig. 5 and Fig. 6. In the LashHer-020 sequence Fig. 5, PFTrack loses and incorrectly re-identifies the scooter after occlusion. Our method, however, maintains the correct identity throughout the entire sequence. This visually confirms the superior **identity preservation** (our high IDF1 score) enabled by the discriminative, harmonized features generated by our C-MDF module, which are robust to temporary occlusion.

In the Photo-0310-42 sequence Fig. 6, which features camera jitter and object adhesion, our tracker remains stable, while PFTrack exhibits frequent identity jumps. This further demonstrates that our iterative refinement process successfully resolves cross-modal conflicts and noisy inputs, producing a stable and temporally coherent feature representation that is less prone to error.

7. Network Architecture Details

To ensure reproducibility and enable fair comparison with existing methods, we provide comprehensive architectural specifications for our proposed approach. We adopt the widely-used DLA-34 backbone [28] due to its demonstrated effectiveness across MOT benchmarks, which allows for an unbiased evaluation of our diffusion-based contributions independent of any architectural enhancements. It is worth noting that our framework maintains flexibility in backbone selection, and can readily accommodate alternative architectures as needed.

7.1. Backbone: DLA-34

Unlike traditional sequential networks, DLA-34 employs recursive and hierarchical connections, enabling it to retain both fine-grained spatial information and high-level semantic cues. This design is particularly beneficial for object detection and tracking tasks, where multi-scale representations are critical. Our implementation of DLA-34 consists of six hierarchical stages. The early levels (Level 0 and Level 1) comprise conventional convolutional layers, while the deeper levels (Levels 2–5) adopt a tree-like structure implemented using the *Tree* class. These stages utilize the Hierarchical Deep Aggregation (HDA) mechanism, which incrementally combines features from different depths to construct a robust feature hierarchy. The overall configuration is summarized in Table 7. Each tree module is composed of *Basic Block* units, which follow the classic residual design with two 3×3 convolutions. This enables both stable training and efficient gradient flow, making the backbone well-suited for real-time detection and tracking pipelines.

7.2. Neck: Iterative Deep Aggregation Upsampling

To recover spatial resolution and fuse multi-scale features from the backbone, we employ a feature aggregation neck composed of *DLAUp* and Iterative Deep Aggregation Upsample (*IDAUp*) modules. These modules, inspired by the original DLA design, progressively combine low-resolution, high-semantic features with high-resolution, spatially detailed ones.

The fusion process begins from the deepest backbone output (Level 5) and iteratively merges features up to Level 2. Each *IDAUp* module aligns the channel dimensions of its input features, upsamples the lower-resolution maps using transposed convolutions, and merges them via element-wise addition. The architectural details are provided in Table 8.

To enhance geometric adaptability, all projection and aggregation operations within the *IDAUp* modules are implemented using Deformable Convolutions (DCNv2) [34]. This allows the model to dynamically adjust sampling locations based on object shapes and motion patterns, which is particularly beneficial in multi-object tracking scenarios. The final output of the neck is a unified feature map with 1/4 resolution of the input image, serving as input to the detection and tracking heads.

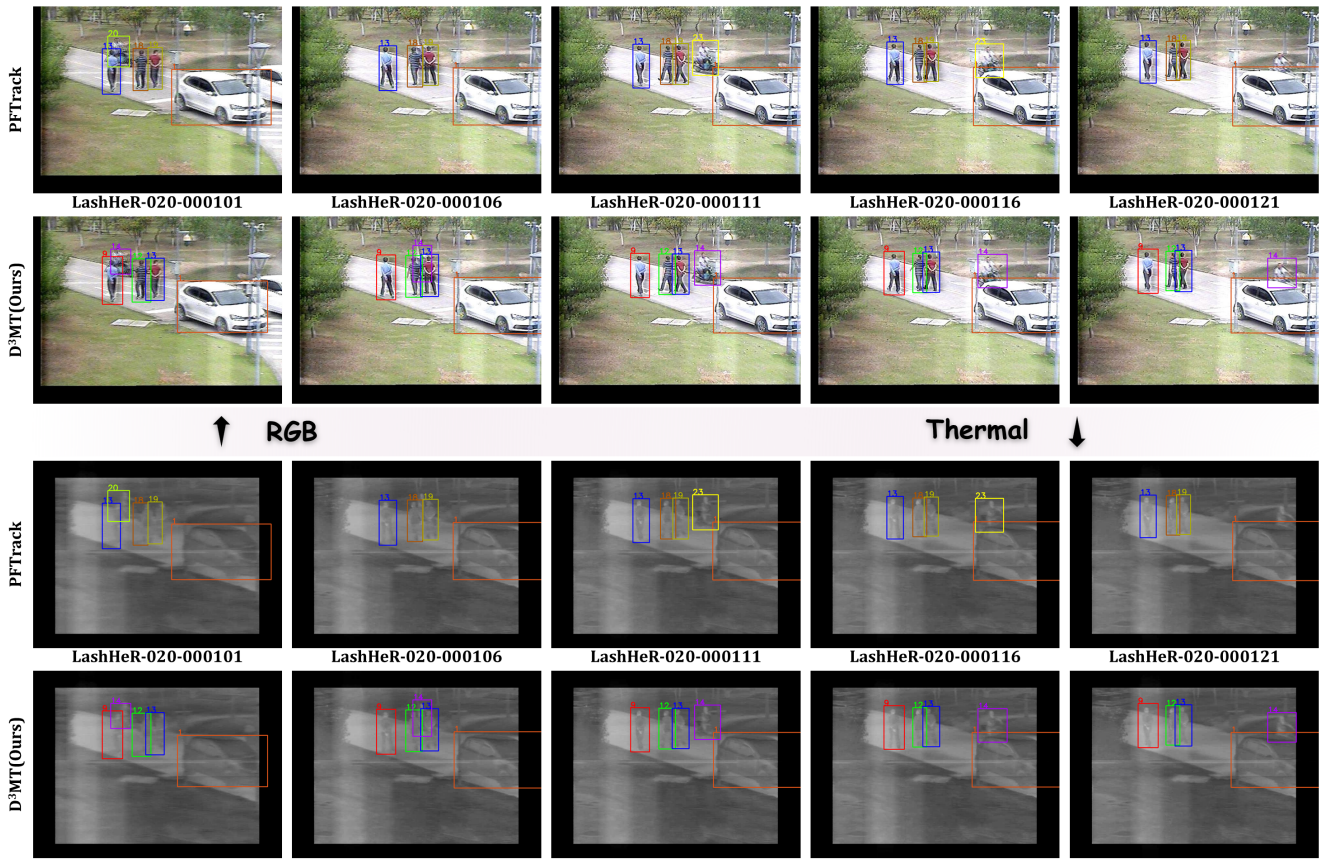

Figure 5. Tracking performance on LashHer-020 sequence. Our method maintains robust multi-object tracking despite significant occlusion. Different object identities are distinguished by bounding box colors and ID numbers, best viewed in color and zoomed. Same as below.

Table 7. Structure of the DLA-34 backbone used in our model. Levels 2–5 are constructed with *Tree* modules that employ *Basic Block* as the internal unit.

Level	**Module**	**Blocks/Convs**	**Output Channels**	**Output Stride**
base_layer	Conv7x7	1	16	1
Level 0	Conv3x3	1	16	1
Level 1	Conv3x3	1	32	2
Level 2	Tree	1	64	4
Level 3	Tree	2	128	8
Level 4	Tree	2	256	16
Level 5	Tree	1	512	32

Table 8. Structure of the feature aggregation neck. Each *IDAUp* module merges features from two levels. All projection and aggregation operations utilize Deformable Convolutions (DCNv2). L_i denotes the feature map from Level i of the backbone. 'd' indicates the number of channels.

Fusion Stage	**Inputs**	**Upsample Operation**	**Output**
ida_0	L_5 (512-d), L_4 (256-d)	Project L_5, Deconv $\times 2$, Add	Fused L_4 (256-d)
ida_1	Fused L_4 (256-d), L_3 (128-d)	Project, Deconv $\times 2$, Add	Fused L_3 (128-d)
ida_2	Fused L_3 (128-d), L_2 (64-d)	Project, Deconv $\times 2$, Add	Fused L_2 (64-d)
Final Output	Fused L_2 (64-d, stride 4)		

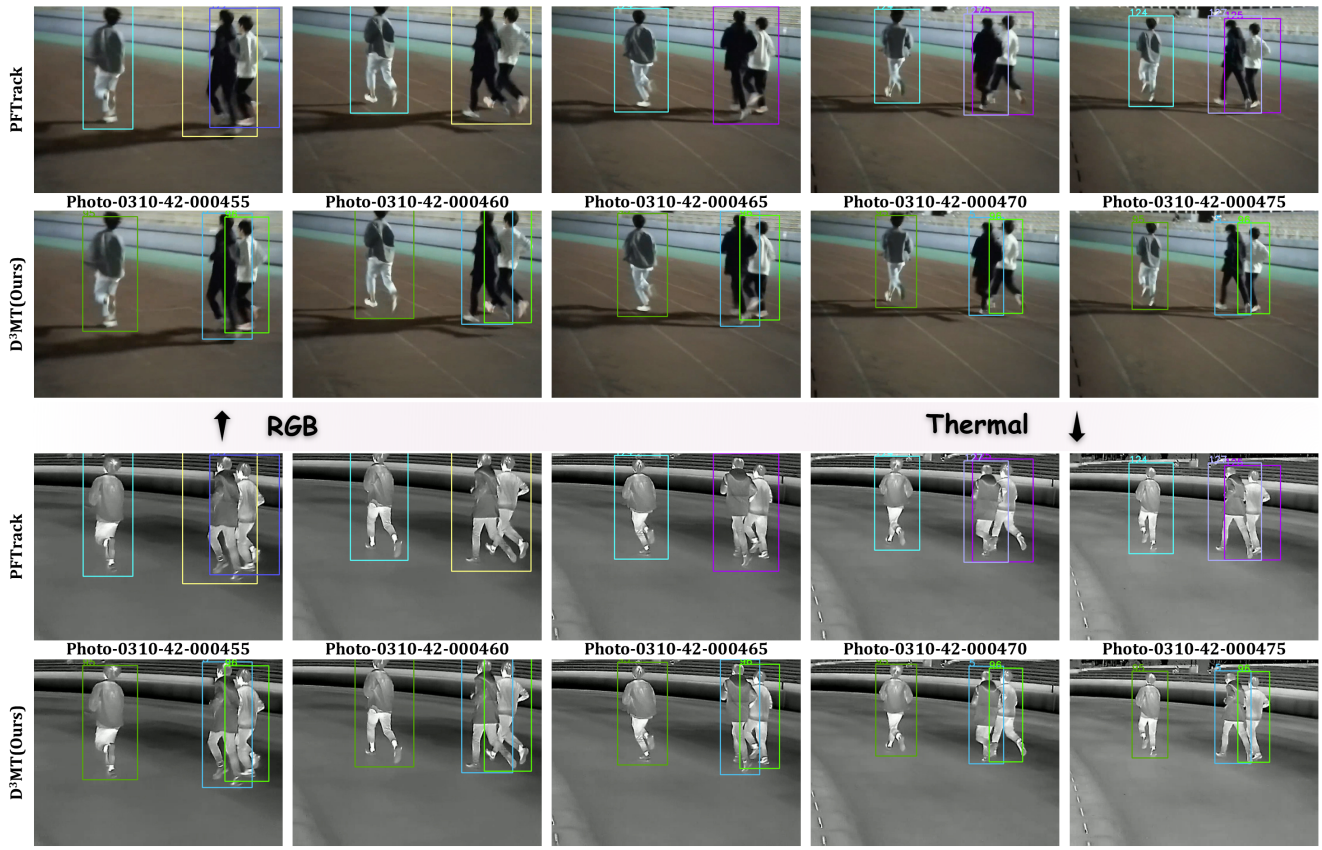

Figure 6. Comparative tracking performance on Photo-0310-42 sequence. Our method (bottom) remains stable despite camera jitter and object adhesion.