WORLD LAUREATES FORUM
世界顶尖科学家论坛
2026
国际工程智能大会
2026 International Conference on AI for Engineering
World Laureates Foundation
上海世界顶尖科学家发展基金会
工程智能研究院
INSTITUTE OF AI FOR ENGINEERING

Semi-supervised Nighttime Pedestrian Tracking via Dual Teachers and VLM-guided Semantic Enhancement

Dingzhong Cai[1], Wending Yan[1], Chunrong Hua[1], Yan Huang[1], Dawei Dong[1], Imdad Ullah Khan[1], Bing Yan[1]

[1]School of Mechanical Engineering, Southwest Jiaotong University, Chengdu, China

01 ABSTRACT

Nighttime object tracking supports autonomous driving and surveillance. We contribute STST with a dual-teacher structure, VLM-guided semantic pseudo-label refinement and a novel multi-scale student network, and establish NMOT2024. STST achieves 57.9 HOTA on real-night sequences, exceeding the strongest retrained baseline by 10.2 percentage points.

Research question Nighttime object tracking faces several challenges, including scarce nighttime annotations, weak feature representations, and significant object scale variations.

02 METHOD

NMOT2024 benchmark

We establish NMOT2024 with 20 unlabeled training sequences and 16 annotated test sequences (11 real, 5 synthetic nighttime). Road-view scenes include low visibility, occlusion, motion blur and scale variation. Labeled MOT17 supplies additional supervised training data.

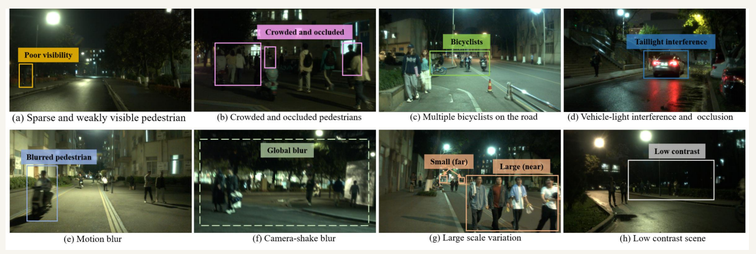


Fig. 1. NMOT2024 benchmark (brightened).

Training pipeline

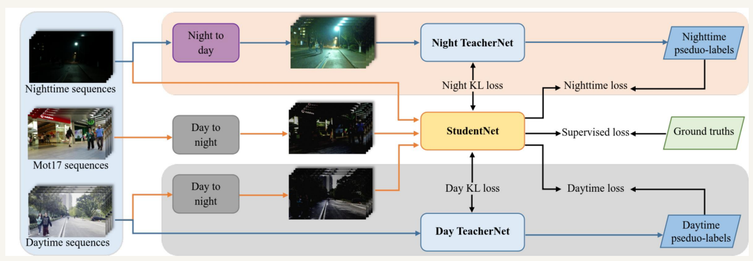


Fig. 2. STST training pipeline.

Unlabeled nighttime/daytime and labeled MOT17 sequences train the student via Night TeacherNet, Day TeacherNet, and supervised loss. Night/day KL losses align teacher–student feature distributions.

Night TeacherNet

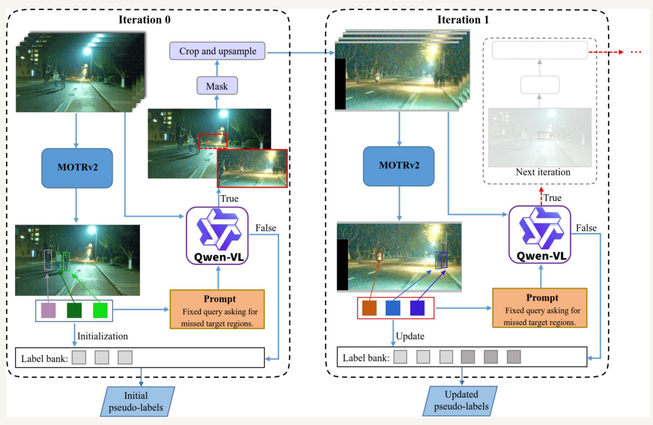


Fig. 3. Iterative pseudo-label refinement.

Night TeacherNet incorporates V-IPLM strategy, using Qwen VLM semantic guidance and iterative teacher re-tracking to refine pseudo-labels.

Student: Multi-scale tracking

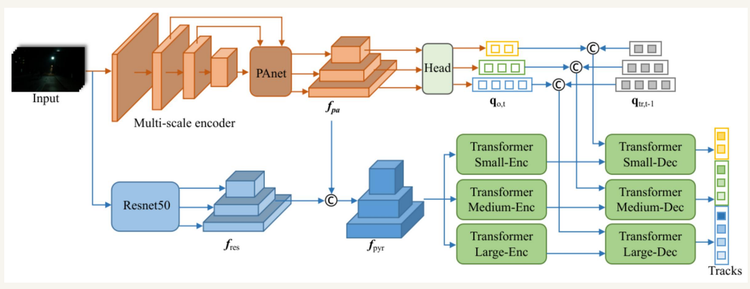


Fig. 4. Student network.

The student jointly detects and associates objects. Dual backbones strengthen pyramid features under low visibility; scale-specific encoder-decoder pairs handle size variation. Only the student network is activated during inference.

03 RESULTS

Baselines (*) are retrained on synthetic-night MOT17 and annotated Real-Train to reduce domain shift; STST uses no night training annotations.

Quantitative Results

Table 1. NMOT2024-Real-Test; higher is better.

Method	HOTA ↑	DetA ↑	AssA ↑	MOTA ↑	IDF1 ↑
ByteTrack*	47.7	43.3	52.7	49.7	57.3
TrackFormer*	34.5	41.0	29.7	39.3	37.3
MOTR*	32.0	27.5	37.2	32.4	37.7
MOTRv2*	41.9	43.0	41.3	48.0	49.7
MOTIP*	42.4	46.6	38.8	50.5	46.6
STST (ours)	**57.9**	**51.4**	**65.3**	**55.4**	**64.8**

Our method leads all retrained baselines across the five tracking metrics.

Qualitative Results



Fig. 5. Ours method maintains dark-clothed and small-pedestrian tracks across frames #96 and #106.

Pseudo-label refinement

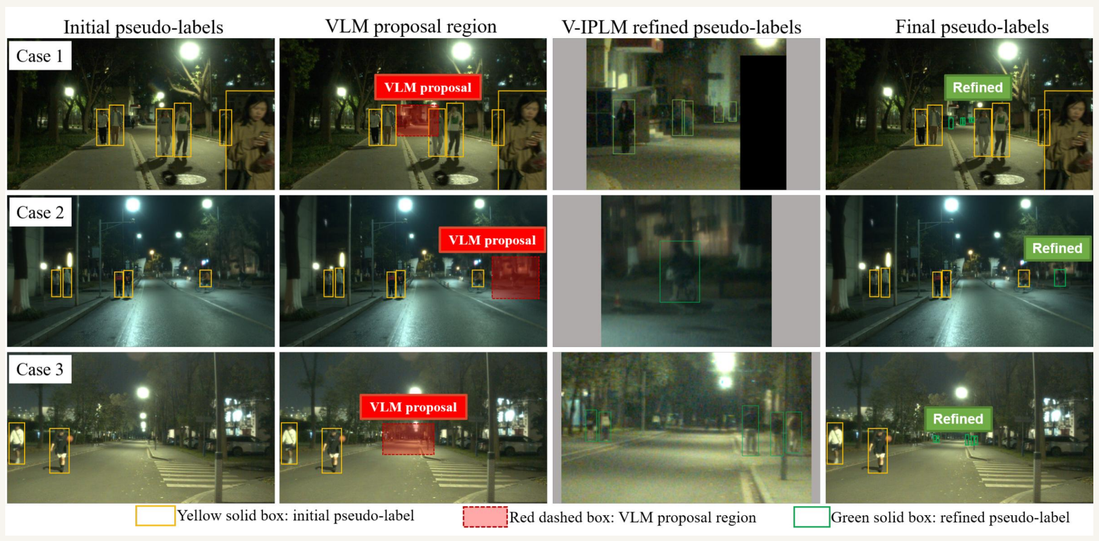


Fig. 6. V-IPLM recovers missed pedestrians and improves pseudo-label coverage (brightened for visibility).

04 DISCUSSION



Fig. 7. Bbox-level masking may remove distant small pedestrians, causing missed pseudo-labels. Future work will explore finer masking and instance-aware separation.

05 CONCLUSIONS

Our method provides dual-teacher supervision, VLM-guided semantic pseudo-label refinement, and multi-scale student tracking. These contributions support nighttime perception in autonomous driving, robotics, and surveillance.

CODE / PROJECT
https://github.com/cdzzzz/STSTtracker

ACKNOWLEDGEMENTS
NSFC 51875482; Sichuan NSF 2022NSFSC0416