MEVAR: MOBILITY-ENHANCED VEHICLE TRAJECTORY RECONSTRUCTION FROM CAMERA SENSING NETWORKS

Jingtian Ma[1,2], *Jingyuan Wang*[1,2*]

[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Engineering Research Center of Advanced Computer Application Technology, Ministry of Education, Beihang University, Beijing, China

ABSTRACT

Vehicle trajectory reconstruction is crucial for understanding mobility patterns and supporting downstream applications. Existing approaches for camera-sensing data mainly cluster images by visual similarity, with limited spatio-temporal constraints, making them highly sensitive to image quality. We propose *MEVAR*, a mobility-enhanced framework that integrates an instance-level visual module with a road-embedded spatio-temporal point process to model the joint probability of trajectory points. This relaxes reliance on appearance features and enforces trajectory-level consistency. Furthermore, we introduce a self-supervised iterative optimization scheme that alternately refines the mobility model and clustering results without labeled data. Experiments on two large-scale highway datasets demonstrate that MEVAR achieves superior accuracy and generalization compared with strong baselines.

***Index Terms*—** Vehicle trajectory reconstruction, camera sensing network, spatio-temporal modeling, self-supervised learning

1. INTRODUCTION

Vehicle trajectory reconstruction is fundamental for understanding traffic patterns and supporting applications such as flow estimation [1, 2], signal control [3], and route planning [4, 5]. Traditional GPS-based approaches suffer from privacy concerns and limited coverage, whereas urban camera networks continuously capture vehicles across intersections, offering a complementary opportunity to recover complete trajectories. We therefore study trajectory reconstruction in a *camera sensing network*, aiming to cluster images from distributed cameras into coherent vehicle trajectories.

Unlike GPS data, camera data is network-wide but lacks vehicle IDs, making association challenging. Re-identification (ReID) methods [6, 7, 8, 9] rely on visual similarity, but performance degrades under illumination, viewpoint, or occlusion variations [10, 11]. Moreover, visually similar vehicles (e.g., same model) may be misclassified as identical. To mitigate this, spatio-temporal cues can provide trajectory-level constraints. Several prior studies combine visual and spatio-temporal cues [8, 12, 13], but most impose only local constraints (e.g., pairwise time windows or adjacent-camera links), leaving long-range consistency across the road network underexplored. As illustrated in Fig. 1, two visually similar vehicles A and B travel along the blue and green routes, respectively. If only appearance features and local spatio-temporal constraints are considered, they may be erroneously linked into the gray dashed route, which contradicts road topology and mobility patterns.

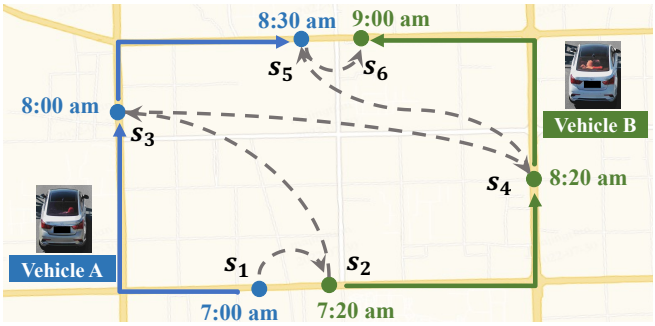


Fig. 1. Two visually similar vehicles moving in different directions; local cues may confuse them, while long-range consistency disambiguates trajectories.

To tackle these challenges, we propose *MEVAR (Mobility-Enhanced Vehicle trAjectory Reconstruction)*, a model that jointly leverages visual and spatio-temporal information while enforcing trajectory-level consistency across the road network. MEVAR contains two complementary modules: (i) an instance-level visual module that extracts robust appearance features to measure local similarity, and (ii) a trajectory-level mobility module based on a road-embedded spatio-temporal point process (RSTPP) that evaluates the joint probability of candidate trajectories, capturing long-range mobility constraints beyond image quality. Built upon these modules, a self-supervised iterative optimization framework

*Jingyuan Wang is the corresponding author. This work was partially supported by the National Natural Science Foundation of China (No. 72171013, 72222022, 72242101), and the Fundamental Research Funds for the Central Universities (JKF-2025017226182).

alternately refines clustering results and the mobility model, enabling mutual improvement without labeled data. This design makes MEVAR more resilient to visual ambiguities and better aligned with real-world traffic dynamics than existing ReID- or heuristic-based methods.

The main contributions of this work are threefold: 1) We design a neuralized road-embedded spatio-temporal point process that enforces trajectory-level consistency across the road network. 2) We develop a self-supervised iterative framework that jointly refines clustering and mobility modeling without requiring labeled data. 3) We validate MEVAR on two large-scale highway datasets, where it consistently outperforms state-of-the-art baselines.

2. PRELIMINARIES

We summarize the notations and define the problem setting.

Definition 1 (Road Network) *A road network is a directed graph $\mathcal{G} = \langle \mathcal{S}, \boldsymbol{A} \rangle$, where $\mathcal{S}$ is the set of n_s intersections and $\boldsymbol{A} \in \{0,1\}^{n_s \times n_s}$ is the adjacency matrix, with $\boldsymbol{A}[i,j] = 1$ if a directed link exists from s_i to s_j.*

Definition 2 (Traffic Camera Record) *A record is $r = \langle c, s, p, t, i \rangle$, denoting camera c (mapped to intersection s by OSMnx [14]) captures vehicle p at time t with image i.*

Definition 3 (Vehicle Trajectory) *A trajectory of vehicle p is $\mathcal{T}_p = [(s_0, t_0), \ldots, (s_n, t_n)]$, where (s_i, t_i) indicates p passing intersection s_i at time t_i.*

Definition 4 (Vehicle Trajectory Reconstruction) *Given traffic camera records $\mathcal{R}$ on road network $\mathcal{G}$, the goal is to recover vehicle trajectories $\mathcal{T}$:*

$$\mathcal{T} \leftarrow \mathcal{F}(\mathcal{R}, \mathcal{G}), \tag{1}$$

where $\mathcal{F}$ maps images to trajectories. Equivalently, this is an image clustering problem, with each cluster representing one vehicle and its trajectory obtained by chronological sorting.

3. METHOD

As in Fig. 2, the pipeline alternates between (i) an *instance-level* visual module that provides appearance similarity and (ii) a *trajectory-level* mobility module that scores the spatio-temporal consistency of candidate trajectories. Clustering fuses the two scores; the resulting high-confidence trajectories supervise the mobility module, which in turn improves the next clustering round. No manual labels are required.

3.1. Instance-level Visual Module

Video streams from roadside cameras are sampled at fixed intervals and resized to 256×256. Standard augmentations (random flip/erase, p=0.5) are applied to improve robustness to illumination and occlusion; all privacy-sensitive content is anonymized. A frozen, pre-trained ResNet-50 is adopted as backbone due to its strong generalization ability and efficiency, extracting d-dimensional embeddings that capture

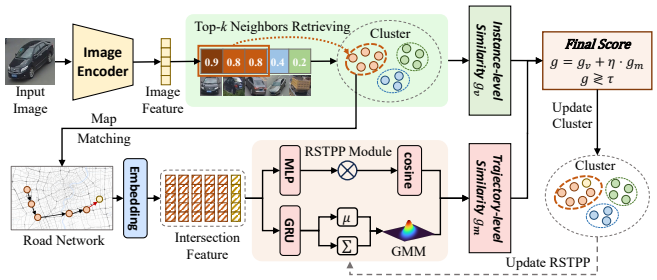


Fig. 2. Self-supervised framework that integrates an instance-level visual module for appearance similarity with a trajectory-level mobility module (RSTPP) for long-range spatio-temporal consistency.

discriminative appearance features. For each image, we retrieve top-k nearest neighbors via cosine similarity using FAISS [15], forming candidate associations for trajectory reconstruction. These instance-level matches provide a visual prior, which is later refined and validated by the mobility module to enforce trajectory-level consistency.

3.2. Trajectory-level Mobility Module

We model trajectory rationality on the road network with a Road-embedded Spatio-Temporal Point Process (RSTPP).

Graph Embedding. We learn intersection embeddings $\boldsymbol{H} \in \mathbb{R}^{n_s \times d}$ by first initializing a trainable embedding matrix for all intersections and then refining them through a multi-layer GAT [16], which aggregates neighborhood information with learnable attention weights. The embeddings are optimized via a graph reconstruction loss:

$$\mathcal{L}_{\text{graph}} = |\boldsymbol{A} - \boldsymbol{H}\boldsymbol{H}^\top|_F^2, \tag{2}$$

where $\boldsymbol{h}_s \in \mathbb{R}^d$ denotes the embedding of intersection s.

RSTPP. The traditional Temporal Point Process (TPP) is a probabilistic generative model describing variable-length time series. The core of TPP is the *intensity function*, which means the expectation of the number of events occurring per unit time. We extend TPP to RSTPP by incorporating the intersection embeddings into the intensity function.

Given a partial trajectory $\{(s_j, t_j)\}_{j<t}$, the conditional intensity of visiting (s,t) consists of two parts: (i) a *static term* that encodes road priors from intersection embeddings, and (ii) a *dynamic term* that models history gain from past trajectory points. Formally,

$$\lambda^*(s,t) = \underbrace{\cos\Big(h_s, \sum_{t_j<t} w_j\, h_{s_j}\Big)}_{\text{static (road prior)}} + \alpha \underbrace{\sum_{t_j<t} \gamma(s, s_j, t, t_j)}_{\text{dynamic (history gain)}}, \tag{3}$$

with temporal weights $w_j = \frac{e^{-(t-t_j)}}{\sum_\ell e^{-(t-t_\ell)}}$ and trade-off $\alpha > 0$.

Dynamic term. We then use a K-component Gaussian mixture over spatial–temporal deltas $\Delta = [\Delta_s, \Delta_t]$ with $\Delta_s =$

$\|r_s - r_{s_j}\|_2$ and $\Delta_t = |t - t_j|$ to model history gain:

$$\gamma(s, s_j, t, t_j) = \sum_{k=1}^{K} \phi_j^{(k)} \, g\Big(\Delta \,\Big|\, \mu_j^{(k)}, \Sigma_j^{(k)}\Big), \tag{4}$$

where $\mu_j^{(k)}$ and $\Sigma_j^{(k)}$ are mean and covariance matrix parameters respectively. $\phi^{(k)}$ is the weight of k-th kernel which satisfies $\sum_{k=1}^{K} \phi_j^{(k)} = 1$. For each Gaussian diffusion kernel g, we define it as

$$g = \frac{1}{2\pi\sqrt{|\Sigma_j^{(k)}|}} \exp\Big(-\frac{(\Delta - \mu_j^{(k)})^\top {\Sigma_j^{(k)}}^{-1} (\Delta - \mu_j^k)}{2}\Big), \tag{5}$$

where $C > 0$ is a hyper-parameter which controls the magnitude. We specify μ_j and Σ_j (omitting superscript k) as

$$\mu_j = [\mu_j(s), \mu_j(t)], \tag{6}$$

$$\Sigma_j = \begin{bmatrix} \sigma_j^2(s) & \rho_j(st)\sigma_j(s)\sigma_j(t) \\ \rho_j(st)\sigma_j(s)\sigma_j(t) & \sigma_j^2(t) \end{bmatrix}, \tag{7}$$

where $\mu_j(s)$ and $\sigma_j(s)$ correspond to spatial mean and variance, $\mu_j(t)$ and $\sigma_j(t)$ to temporal mean and variance, and $\rho_j(st)$ to the spatio-temporal correlation coefficient. Instead of being predefined, these five parameters are adaptively generated via nonlinear mappings of spatio-temporal features.

By substituting these five learned parameters into the Gaussian kernels and then into the conditional intensity in Eq. (3), the log-likelihood of a trajectory X on $(0, T] \times \mathcal{S}$ is

$$\ell(\theta) = \sum_{(s_i, t_i) \in X} \log \lambda_\theta^*(s_i, t_i) \; - \; \int_0^T \sum_{s \in \mathcal{S}} \lambda_\theta^*(s, \tau) \, d\tau, \tag{8}$$

which admits an analytic integral following [17]. For N_t credible trajectories, the mobility loss is defined as the average negative log-likelihood:

$$\mathcal{L}_{\text{mob}} = -\frac{1}{N_t} \sum_{i=1}^{N_t} \ell_i(\theta), \tag{9}$$

where $\ell_i(\theta)$ denotes the log-likelihood of trajectory i. The total objective is $\mathcal{L} = \mathcal{L}_{\text{graph}} + \mathcal{L}_{\text{mob}}$.

3.3. Self-supervised Clustering with Iterative Refinement

Clustering. For each record r, let $\mathcal{C}_r$ be its top-k visual neighbors. Given a candidate $r' \in \mathcal{C}_r$ that belongs to cluster u, we first compute the visual similarity between r and r' as the *instance-level* similarity: $g_v(r, r') = \cos(\boldsymbol{f}_r, \boldsymbol{f}_{r'})$. Then we append r at the end of the chronological sequence of u to form the trajectory X and compute a trajectory-level score $g_m(r, r')$ from RSTPP (proportional to $\ell(\theta)$). The final matching score is

$$g(r, r') = g_v(r, r') + \eta \, g_m(r, r'), \tag{10}$$

Algorithm 1 Self-supervised Iterative Clustering

1: Build $\mathcal{R}$ chronologically; learn $\boldsymbol{H}$ by (2); precompute top-k visual neighbors.
2: **for** iteration $= 1, \ldots, N_i$ **do**
3: **Clustering:** for each record r, compute $g(r, r')$ over $\mathcal{C}_r$ and assign by threshold θ.
4: **RSTPP update:** train with current trajectories by maximizing $\ell(\theta)$.
5: Increase η, decrease θ.
6: **end for**

Table 1. Dataset statistics. Mean/Max: images per trajectory.

Dataset	#Image	#Vehicle	#Camera	Mean	Max
HD-20	197,753	44,718	11	4	36
HD-120	1,225,769	66,436	79	18	135

with iteration-dependent weight η (increasing across rounds). We assign r to the cluster of $r^* = \arg\max_{r'} g(r, r')$ if $g(r, r^*) \geq \tau$, otherwise start a new cluster.

Iteration schedule. Round 1 uses η=0 and a high threshold τ to harvest high-confidence trajectories for training RSTPP. Subsequent rounds gradually increase η and relax τ, alternating clustering, merge, and RSTPP update until convergence. See Algorithm 1 for more details.

4. EXPERIMENTS

4.1. Experimental Setup

Datasets. We use two highway datasets from roadside cameras. After detection and filtering (keep the largest vehicle box per frame), each record stores camera geo-coordinates and timestamps. The large set **HD-120** contains ∼1.23M images; **HD-20** is a short-distance subset with ∼0.20M images. Statistics are in Tab. 1.

Metrics. For image i, let $P(i)$ be its predicted cluster, $Q(i)$ its ground-truth trajectory, $\mathcal{U}$ the set of clusters. We define

$$\text{Prec} = \tfrac{1}{|\mathcal{I}|} \sum_i \tfrac{|P(i) \cap Q(i)|}{|P(i)|}, \text{Rec} = \tfrac{1}{|\mathcal{I}|} \sum_i \tfrac{|P(i) \cap Q(i)|}{|Q(i)|},$$

$$\text{Expansion} = \tfrac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{u \in \mathcal{U}} \mathbb{I}[|Q(i) \cap u| \neq 0],$$

where lower Expansion means fewer fragmented trajectories.

Baselines. (1) **BNN** [18]: a strong ReID baseline, used as a pure-visual clustering pipeline. (2) **B-HMM** [19]: BNN augmented with an HMM over the road graph to impose spatio-temporal priors. (3) **SPLSTM** [12]: a spatio-temporal LSTM that models local camera transitions with temporal windows, serving as a joint visual–ST baseline. (4) **VeTrac** [20]: a camera-network trajectory reconstruction method; we adopt its MDS+GCN+HDBSCAN pipeline for comparison.

Table 2. Main results (**bold** = best, underline = second-best).

Dataset	Metric	BNN	B-HMM	SPLSTM	VeTrac	**MEVAR**
HD-20	Prec↑	0.885	0.875	0.884	0.743	**0.937**
	Rec↑	0.583	0.578	0.580	0.593	**0.594**
	F1↑	0.703	0.696	0.700	0.660	**0.727**
	Exp↓	1.790	1.806	1.795	1.823	**1.774**
HD-120	Prec↑	0.821	0.827	0.844	0.765	**0.862**
	Rec↑	0.471	0.476	0.491	0.481	**0.524**
	F1↑	0.599	0.604	0.621	0.591	**0.652**
	Exp↓	2.700	2.558	2.506	2.568	**2.373**

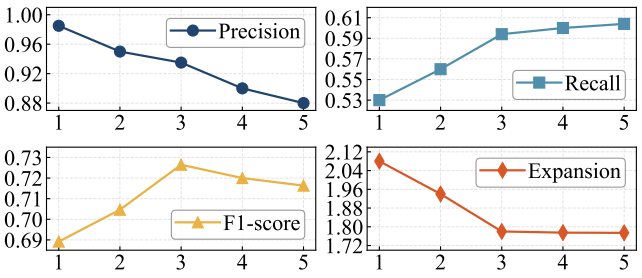


Fig. 3. Parameter study on HD-20 vs. iteration n_i.

4.2. Results

From Table 2, we draw three key conclusions. First, MEVAR consistently achieves the best results on both datasets, improving both accuracy (Prec/Rec/F1) and trajectory coherence (Expansion). Its advantage comes from jointly modeling visual similarity and long-range spatio-temporal consistency under a self-supervised framework, which balances precision and recall while reducing fragmentation. Second, baseline behaviors exhibit dataset-dependent patterns: BNN is strongest on the smaller HD-20, where vehicles have more distinct appearances and visual features suffice; on the larger HD-120, SPLSTM becomes the best baseline, as increased visual ambiguity makes spatio-temporal modeling more important. Finally, VeTrac shows the weakest performance due to heavy reliance on visual similarity without plate cues, and B-HMM provides only marginal gains over BNN, reflecting the limitation of first-order Markov assumptions in modeling complex spatio-temporal dependencies.

4.3. Parameter Study

We schedule iteration-dependent weights as $\theta_i = 1 - 0.05n_i$ and $\eta_i = 0.05(n_i - 1)$ for iteration $n_i \in \{1, \ldots, 5\}$. On HD-20 (Fig. 3), when $n_i \leq 3$, decreasing θ and increasing η trade precision for recall, yielding the best F1 and a simultaneous drop in Expansion at n_i=3. For $n_i > 3$, recall saturates while precision keeps falling, leading to a slight decline in F1, whereas Expansion remains nearly unchanged.

4.4. Ablation Study

We conduct ablations by removing visual (**NV**), mobility (**NM**), static (**NS**), and dynamic (**ND**) components. Results on HD-20 (Fig. 4) indicate that: (a) Without visual cues, trajectories collapse into overly large clusters due to highway homogeneity, leading to high recall but poor precision. (b) Removing the mobility module reduces both precision and recall, highlighting its complementary role to visual features. (c) The static component mainly enhances precision, while the dynamic component primarily contributes to recall.

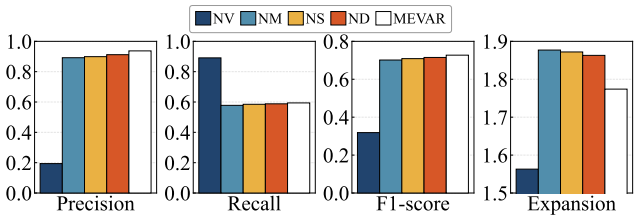
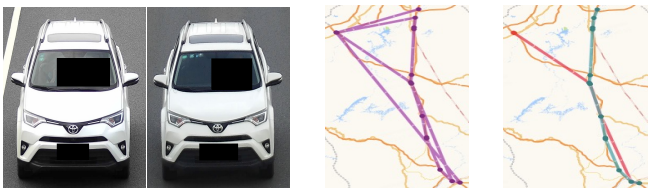


Fig. 4. Ablations on HD-20.



(a) Similar vehicles (b) SPLSTM (c) MEVAR

Fig. 5. Mobility disambiguates visually similar vehicles.

4.5. Qualitative Cases

We illustrate the benefit of the mobility module with two vehicles of nearly identical appearance (Fig. 5(a)). Using only visual similarity and local spatio-temporal cues, SPLSTM incorrectly merges them into one trajectory (Fig. 5(b)), yielding an implausible path that violates road topology. With our road-embedded spatio-temporal modeling, MEVAR correctly separates them into distinct trajectories (Fig. 5(c)). This case highlights that appearance features and local constraints alone are insufficient in complex traffic scenes, while mobility modeling effectively resolves such ambiguities and improves reconstruction robustness. Moreover, it demonstrates the necessity of enforcing long-range trajectory consistency across the camera network, rather than relying solely on local cues.

5. CONCLUSION

We proposed **MEVAR**, a mobility-enhanced framework for vehicle trajectory reconstruction in camera networks. By integrating an instance-level visual module, a road-embedded spatio-temporal point process, and a self-supervised iterative optimization scheme, MEVAR effectively combines visual and mobility cues without requiring labeled data. Experiments on two highway datasets demonstrate clear advantages over strong baselines. Future directions include developing parallelization strategies to accelerate inference and exploiting the trained RSTPP to generate synthetic trajectories for cost-effective data augmentation.

6. REFERENCES

[1] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., Cham, 2016, pp. 869–884, Springer International Publishing.

[2] Ping-Yang Chen, Jun-Wei Hsieh, Munkhjargal Gochoo, Ming-Ching Chang, Chien-Yao Wang, Yong-Sheng Chen, and Hong-Yuan Mark Liao, "Drone-based vehicle flow estimation and its application to traffic conflict hotspot detection at intersections," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020.

[3] Hua Wei, Guanjie Zheng, Vikash Gayah, and Zhenhui Li, "A survey on traffic signal control methods," *arXiv preprint arXiv:1904.08117*, 2019.

[4] Jingyuan Wang, Ning Wu, Wayne Xin Zhao, Fanzhang Peng, and Xin Lin, "Empowering a* search algorithms with neural networks for personalized route recommendation," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 539–547.

[5] Jian Dai, Bin Yang, Chenjuan Guo, and Zhiming Ding, "Personalized route recommendation using big trajectory data," in *2015 IEEE 31st international conference on data engineering*. IEEE, 2015, pp. 543–554.

[6] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[7] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[8] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *European conference on computer vision*. Springer, 2016, pp. 869–884.

[9] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] Chong Liu, Yuqi Zhang, Hao Luo, Jiasheng Tang, Weihua Chen, Xianzhe Xu, Fan Wang, Hao Li, and Yi-Dong Shen, "City-scale multi-camera vehicle tracking guided by crossroad zones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4129–4137.

[11] Minghu Wu, Yeqiang Qian, Chunxiang Wang, and Ming Yang, "A multi-camera vehicle tracking system based on city-scale vehicle re-id and spatial-temporal information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4077–4086.

[12] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1900–1909.

[13] Hung-Min Hsu, Yizhou Wang, and Jenq-Neng Hwang, "Traffic-aware multi-camera tracking of vehicles based on reid and camera link model," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 964–972.

[14] Geoff Boeing, "Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Computers, Environment and Urban Systems*, vol. 65, pp. 126–139, 2017.

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[16] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al., "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.

[17] Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel, "Neural spatio-temporal point processes," in *International Conference on Learning Representations*, 2021.

[18] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[19] Sean R Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.

[20] Panrong Tong, Mingqian Li, Mo Li, Jianqiang Huang, and Xiansheng Hua, "Large-scale vehicle trajectory reconstruction with camera sensing network," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 188–200.