DuARE: Automatic Road Extraction with Aerial Images and Trajectory Data at Baidu Maps

Jianzhong Yang*
Xiaoqing Ye*
yangjianzhong@baidu.com
yexiaoqing@baidu.com
Baidu Inc.
Haidian District, Beijing, China

Bin Wu
Yanlei Gu
Ziyu Wang
wubin16@baidu.com
Baidu Inc.
Haidian District, Beijing, China

Deguo Xia
Jizhou Huang†
xiadeguo@baidu.com
huangjizhou01@baidu.com
Baidu Inc.
Haidian District, Beijing, China

ABSTRACT

The task of road extraction has aroused remarkable attention due to its critical role in facilitating urban development and up-to-date map maintenance, which has widespread applications such as navigation and autonomous driving. Existing solutions either rely on a single data source for road graph extraction or simply fuse the multimodal information in a sub-optimal way. In this paper, we present an automatic road extraction solution named DuARE, which is designed to exploit the multimodal knowledge for underlying road extraction in a fully automatic manner. Specifically, we collect a large-scale real-world dataset for paired aerial image and trajectory data, covering over 33,000 km^2 in more than 80 cities. First, road extraction is performed on the abundant spatial-temporal trajectory data adaptively based on the density distribution. Then, a coarse-to-fine road graph learner from aerial images is proposed to take advantage of the local and global context. Finally, our cross-check-based fusion approach keeps the optimal state of each modality while revisiting the original trajectory map with the guidance of aerial predictions to further improve the performance. Extensive experiments conducted on large-scale real-world datasets demonstrate the superiority and effectiveness of DuARE. In addition, DuARE has been deployed in production at Baidu Maps since June 2021 and keeps updating the road network by 100,000 km per month. This confirms that DuARE is a practical and industrial-grade solution for large-scale cost-effective road extraction from multimodal data.

CCS CONCEPTS

• **Information systems** → **Data mining**.

KEYWORDS

spatial-temporal, transportation, road extraction, Baidu Maps

*Both authors contributed equally to this paper.
†Corresponding author: Jizhou Huang.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539029

ACM Reference Format:
Jianzhong Yang, Xiaoqing Ye, Bin Wu, Yanlei Gu, Ziyu Wang, Deguo Xia, and Jizhou Huang. 2022. DuARE: Automatic Road Extraction with Aerial Images and Trajectory Data at Baidu Maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539029

1 INTRODUCTION

Road network is a fundamental backbone for a wide range of intelligent transportation applications such as navigation, route planning, and autonomous driving. Baidu Maps has covered a total road length of over 11 million km nationwide. However, due to successive construction and upgrade of roads, it is important to construct and maintain the road network. One indispensable task in this context is automatic road extraction, which aims at extracting the road network from abundant data and has been capturing extensive interests due to its significance to widespread downstream applications. For example, the partial absence of road maps can lead to navigation detours and travel efficiency suffering, even accidents. On the contrary, a more complete road map can offer significant benefits to the task of Travel Time Estimation [19].

The task of automatic road map extraction can be formulated as constructing and updating the road network given abundant road data. Nevertheless, large-scale road map extraction is often costly and labor-intensive. Conventional manual approaches such as field surveys and requiring local mappers to collect OpenStreetMap (OSM) [21] are impractical for millions of roads. Instead, recent works appeal to multiple sources for extracting the underlying road networks, mainly by the Global Positioning System (GPS) trajectories and the aerial images. Thanks to the ubiquitous GPS sensors throughout various mobile devices such as mobile phones and moving vehicles, we are able to collect a vast amount of GPS trajectories as long as there is an underlying road.

However, map extraction from GPS trajectories is non-trivial due to two reasons. **First**, the collected GPS points can have different sampling rates and are always noisy. Attempts have been made to alleviate the noise by data-processing techniques such as point clustering [17] and Kernel Density Estimation (KDE) [8]. **Second**, there are fewer GPS trajectories in unfrequented and remote areas whereas in building-intensive regions, the GPS trajectory is prone to causing deviation due to signal shielding. Alternatively, recent years have witnessed the prosperity of deep CNN to recover road networks from the aerial images [6, 7, 13, 24, 31, 41, 43]. Note that

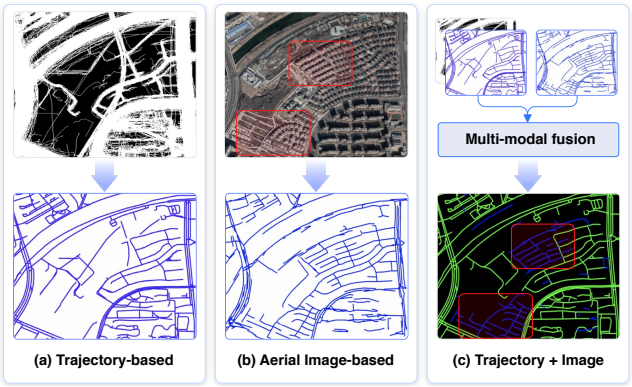

Figure 1: Examples of road extraction using different data. Figure (a)/(b) shows that road extraction with trajectory data/aerial images can achieve promising precision/recall, respectively. By leveraging both of them in (c), we can update the road network with more detailed roads recovered.

aerial here includes images collected by satellites or airplanes and for simplicity, we do not discriminate them particularly.

There are two main challenges that limit the aerial-based real-world application. **First**, the deep neural network relies on heavy labor for annotation and is prone to be overfitting to common scenarios. In other words, the model is non-trivial to generalize well to unseen terrains or to distinguish roads with other structures, such as railway, building tops, and waterways. **Second**, the underlying roads can be easily occluded by trees and tall buildings or be partially shadowed in aerial images. Bad weather and inappropriate angle shot further deteriorate the problem.

As aforementioned, both the GPS trajectories and aerial images have limitations and superiority for road network extraction since the two modalities capture different types of intrinsic knowledge. Intuitively, if digging deep into incorporating the two modalities, it is conceivable that we can obtain more complete and reliable road maps. There are limited works that study fusion strategies, with most of them rendering the GPS trajectory maps to the aerial-based neural networks as another input channel [34] (early-fusion) or appeal to deep neural networks for mutual fusion [27, 39] (deep-fusion). Nevertheless, due to the variance of the two modalities, a unified network might search for a trade-off between the two modality-based features and attain a sub-optimal balance. Alternatively, we pay attention on the late fusion strategy to keep the optimum of each modality independently.

To fully explore the complementary association of the multi-modal sources of data without any manual labor, we propose an automatic industrial-grade solution termed DuARE for road extraction and have applied it in production at Baidu Maps. Our framework consists of three stages. **First**, given the vast spatial-temporal trajectory data, we adopt the density-adaptive processing policies for different-density regions since a single uniformed operation can cause errors in regions with different densities. **Second**, given aerial images as input, as road networks are intuitively represented as a graph of road segments and intersections, we specially design a coarse-to-fine framework by leveraging both the local neighboring context from the coarse stage and the global context from the fine graph-based stage for road extraction. **Third**, given the observation that aerial branch predicts several potential roads that are ignored in trajectory-based road maps, we propose a cross-check-based fusion strategy by revisiting the original spatial-temporal trajectory map with the guidance of aerial predictions. Remote roads that are predicted by the aerial branch and validated by more than two trajectories are recovered.

As shown in Figure 1, although the aerial images and the GPS trajectories have their shortcomings, our fusion strategy exploits the complementarity of the two modalities to a great extent and automatically updates the road map by extracting minor roads that are previously ignored. Our key contributions to both the research and industrial communities are as follows:

- **Potential impact:** We propose a novel DuARE solution for large-scale automatic multimodal road extraction by leveraging the extensive spatial-temporal trajectory data as well as the aerial images. In addition, the proposed framework has been successfully deployed at Baidu Maps and keeps automatically updating the road networks by 100,000 km per month.
- **Novelty:** The solution of fusing aerial image and trajectory is designed to explore the two modalities for complementation. The novelty lies in each stage, from the density-adaptive trajectory-based road extraction, the coarse-to-fine aerial prediction branch, to the cross-check-based fusion policy, making the road extraction task fully automatic and cost-efficient.
- **Scalability:** Our approach can be widespreadly applied to both urban and rural areas and be simply integrated into existing concurrent trajectory-based or aerial image-based map generation approaches.
- **Technical quality:** Extensive experiments conducted on large-scale real-world datasets collected from Baidu Maps demonstrate the superiority of DuARE against the strong baselines.

2 PRELIMINARIES

In this section, we define the two modalities of data and then set up the road extraction problem. Thanks to the large amount users of Baidu Maps, we can collect abundant GPS trajectories every day in various regions. Each GPS point is denoted as Tra=$\{long, lat, t\}$, where $long$ and lat are short for longitude, latitude, respectively, and t is the timestamp. Due to the unbalanced frequencies of different regions, our framework automatically accumulates trajectories to the trajectory map $\mathcal{T}$ gathered from different time spans according to the trajectory density within every 10 km $\times$ 10 km region. As illustrated in Figure 2, the upper diagram reports the numbers of samples with different densities in Log-form, which indicates that with the trajectories getting denser, the regions decrease. The bottom diagram shows how many months of trajectory data are accumulated for various densities. The density is computed as the statistical total length of the trajectories within per region. For example, in frequently-visited regions with more than 1600 km of trajectory length, we select the accumulated trajectory map within two months (the highest density level). However, for untraversed regions, multiple months of data are acquired.

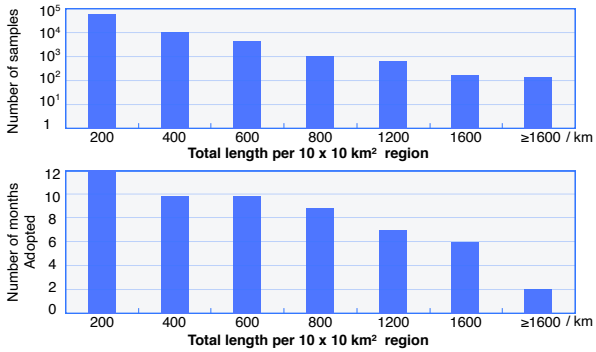

Figure 2: The relationship of total mileage within per 10 × 10km region and the number of months adopted for trajectories accumulation.

The Large-scale Aerial Ground Truth Generation. The task of road extraction from aerial images is defined as follows: given an aerial image $\mathcal{I}$ collected from satellites or airplanes as input, the network is supposed to predict the underlying road graphs $\mathcal{G}^{aer}$. Rather than appealing to heavy labeling manpower, we solve the large-scale annotation problem by the Baidu Map Database in an automatic manner. In specific, given an aerial image I with $W \times H$ resolution, with the spatial resolution 50cm × 50cm for each pixel, it covers $W/2$ meters by $H/2$ meters region with a certain geographic coordinate range. First, we extract the road graph $\mathcal{G}$ from Baidu Map Database according to the latitude and longitude range and then project them onto the corresponding aerial image to get a binary $W \times H$ road map, denoted as $\mathcal{GT}$. The $\mathcal{GT}$ map is then served as the ground truth for follow-up aerial-based road graph learning, where 1 means road areas and 0 means non-road region.

3 APPROACH

To fully exploit the complementary features of the spatial-temporal trajectory data and the aerial images, we propose a novel multi-modal solution for road network extraction termed DuARE. The overview is demonstrated in Figure 3, consisting of three stages. (I) the spatial-temporal trajectory data-based road extraction, in which we extract reliable underlying roads from noisy data in a density-aware manner. (II) the coarse-to-fine road network learner from aerial images. (III) the cross-check-based fusion policy from the two predictions. The visualization of the fused map shows that after the fusion stage, the discontinuous or over-killed roads can be largely recovered.

3.1 Trajectory-based Road Extraction

Given numerous GPS trajectories collected from various sensors at different times and locations, our trajectory-based approach utilize the trajectory density to perform adaptive extraction strategies. The density-aware pipeline is presented in Figure 4, with the adaptive policies embodied in three aspects. First, the trajectories are adaptively accumulated with different time spans according to the statistic density grades, as shown in Figure 2. For example, frequently-visited areas need fewer months of data to gather into the trajectory map $\mathcal{T}$. Second, adaptive selection of various filters is conducted for image denoising. Morphological filter, Gaussian filter, and the combination of them are adaptively selected based on the density distribution. In general, we adopt morphological filtering to suppress noises but for high-density regions when morphological filters fail to deal with the tangled edges, we adopt a Gaussian filter with different variances to filter the noises efficiently while keeping the edges smooth. Then the grade imaging is performed based on the density distribution of the filtered trajectory map.

Third, we propose adaptive thinning policies to extract road networks from the filtered trajectory map. We adopt progressive convolution-based erasing operation for thinning in most regions, *e.g.* 3×3, 9×9 kernels. However, it fails to distinguish the main road from the nearby relief road since the trajectories overlap with each other because of deviation. Given a road, we statistic the accumulated GPS trajectories along the width and length dimension of the road for attaining density distribution. Then a density-specific threshold is applied to extract the high-density lines. In this way, the two neighboring roads can display two different peaks and thus can be both recovered.

3.2 Aerial Image-based Road Extraction

In general, there are mainly two branches of utilizing the deep neural networks for road extraction, the segmentation-based, and the graph-based approaches. The former relies more on local context and is prone to be overfitting to easy samples whereas the latter considers more global context information by constructing a road network. Inspired by Sat2Graph [22] and CenterNet [16], we make the best possible use of local and global context and propose a coarse-to-fine road network learner approach by first detecting the coarse vertices and then learning local feature embedding extracted from the early-level neighboring regions around the coarse vertices for supplementing the fine stage prediction.

Given an aerial image I as input, the model aims to predict the road network $\mathcal{G}^{aer} = \{V, E\}$, with V the set of vertices and E the edges between vertices. After obtaining the ground truth road map $\mathcal{GT}$ from the database as well as its topological connectivity, we first encode $\mathcal{GT}$ into an undirected graph by identifying the intersection points as well as sampling the uniformly spaced points with the distance threshold d as vertices. As suggested in [22], we set d=20 meters to distinguish stacked roads while keeping the sparsity for stable training. For graph-encoding, we follow a similar operation to predict a 19-dimension tensor in each pixel with the first element indicating the probability of existing a vertex in each pixel and the following tuples denoting the edgeness toward the i-th 60-degree sector and the relative position of the i-th edge. More details is explained in [22]. Furthermore, we add an addition tensor for indicating the roadness in each pixel, *i.e.*, the probability of the point lying on the road no matter whether it is a vertex or not.

Network design. It is worth noting that directly regressing the vertexness (probability of being a vertex) at each position is non-trivial for the network without any prior knowledge. Possibly if the first element, *i.e.*, the vertexness is erroneously predicted as a background point, the network has no second chance to recover it. To solve the problem, we propose the two-stage pipeline in a coarse-to-fine manner. The framework is elaborated in Figure 5. In specific, our coarse-to-fine pipeline consists of two heads, both of which share the same encoder. Many auto-decoder architectures can be

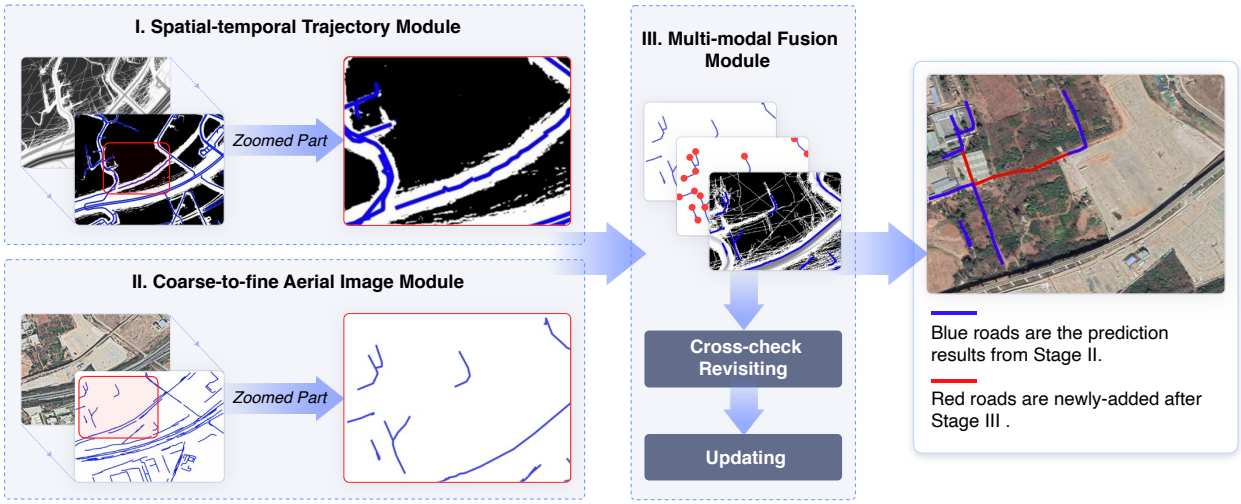

Figure 3: The overview of our DuARE solution for multimodal road extraction. Stage I: road extraction from spatial-temporal trajectory data, with detailed roads failing to be extracted. Stage II: The coarse-to-fine road network learner from aerial images. Roads within the red boxed region are incomplete. Stage III: The multimodal cross-check-based fusion module. The visualization of the fused map shows that after the fusion stage, the discontinuous or over-killed roads can be largely recovered.

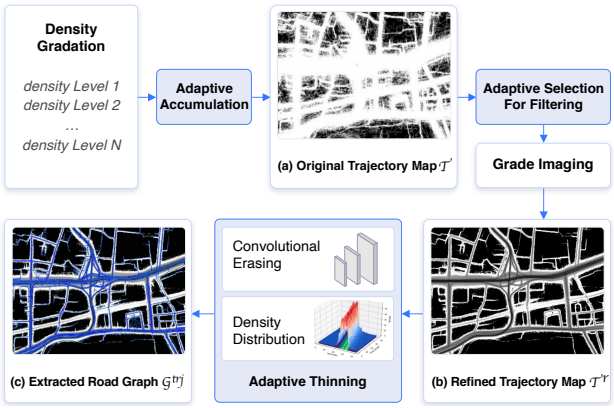

Figure 4: The pipeline for reliable road extraction from the collected abundant noisy trajectories.

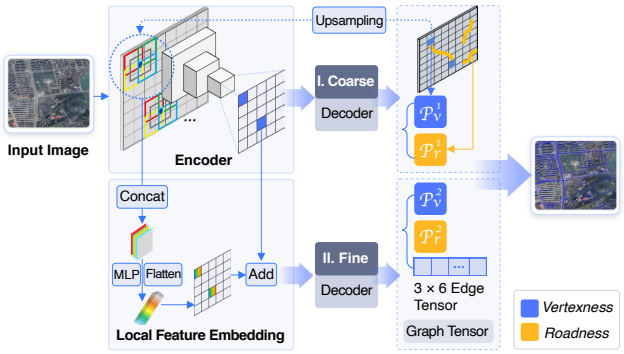

Figure 5: The architecture of the coarse-to-fine road network learner from aerial images. The coarse stage is first trained to detect the coarse locations of vertices. Then neighboring features are extracted for local feature embedding to be fused with the deep-layer features containing the global context. The Fine stage finally predicts the vertexness and roadness of the aerial image given the fused features.

utilized for feature learning [25, 33, 41]. We adopt the DLA [40] as the backbone, with the resolutions of the feature maps becoming smaller along with the layers deeper, from $W \times H$ to $W/32 \times H/32$.

Coarse-stage design. First, we specially design a coarse stage with a semantic segmentation head predicting the roadness p_r of each pixel (the probability of being a road point) and a vertex head predicting the vertexness p_v of each pixel (the probability of being a vertex). Road segmentation knowledge offers supplementary semantic guidance for localization of the vertices.

Local feature embedding. Next, for a certain predicted vertex $V(u, v)$ from the coarse head (*i.e.* the vertexness exceeds a predefined threshold) which is predicted on the deep-level feature map of size $W/32 \times H/32$, we first map it onto the shallow feature map of size $W \times H$ in the encoder by upsampling to obtain the anchor point $V_s(u', v')$. Taking the anchor point as the center (c), bottom-left (bl), bottom-right (br), top-left (tl) and top-right (tr) corners respectively, we extract the corresponding region-of-interest (RoI) features with a fixed window size S and concatenate them to form the stacked local features:

$$f_{local} = [f_c, f_{bl}, f_{br}, f_{tl}, f_{tr}] \in \mathbb{R}^{5\times C\times S\times S} \quad (1)$$

where f_c denotes the RoI region centered with the point V_s and so forth, C is the channel of the feature map. Rather than simply extracting the local feature map f_c, we take full advantage of the topology of road graph to combine the four neighboring feature maps cornered with the coarse vertex. The concatenated features

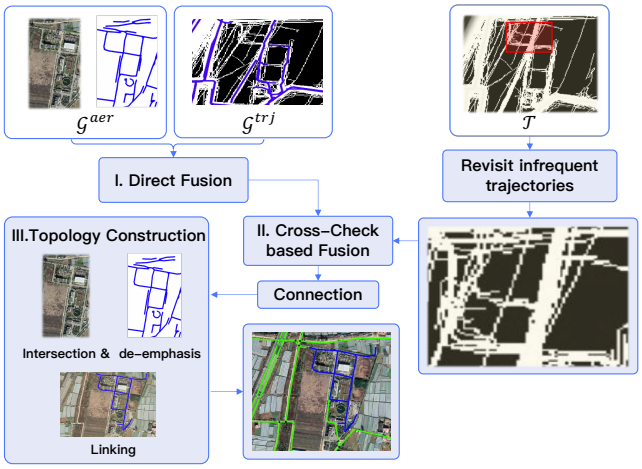

Figure 6: The multimodal fusion strategy by cross-checking.

are fed into a lightweight Multi-layer Perceptron (MLP) block for local feature embedding to obtain $f_{embed} = MLP\{f_{local}\}$. Assume the deep features at the output of the DLA encoder are denoted as f_{deep} with the resolution of $W/32 \times H/32$. We add the f_{deep} and the local feature embedding f_{embed} for the input of fine decoder (the bottom branch of Figure 5).

Fine-stage design. The input for the fine-stage decoder is the addition of local features extracted from the neighboring regions around the coarse vertices and the deep features involving more global context by the encoder. As for the output of the fine stage, apart from the original 19-dim graph tensor explained in Sat2Graph [22], we additionally predict the roadness attribute as the coarse stage does, to leverage the semantic segmentation guidance.

Training Process. The coarse-to-fine pipeline is trained end-to-end in three stages. For the first stage, we only train the encoder and the coarse decoder under the supervision of road segmentation and the ground truth location of vertices. The cross-entropy loss is applied to vertexness and segmentation loss.

$$\mathcal{L}_{stage1} = \mathcal{L}_{CE}(p_v^1, \hat{p_v^1}) + \mathcal{L}_{CE}(p_r^1, \hat{p_r^1}) \tag{2}$$

where $\hat{p_v}$ and $\hat{p_r}$ are the ground truth vertexness and roadness.

Second, the parameters of coarse decoder are fixed to finetune the encoder as well as train the fine decoder for graph tensor prediction. L2 loss is adopted to supervise the edge vector.

$$\mathcal{L}_{stage2} = \mathcal{L}_{CE}(p_v^2, \hat{p_v^2}) + \mathcal{L}_{CE}(p_r^2, \hat{p_r^2}) + \Gamma \cdot \mathcal{L}_{edge} \tag{3}$$

where Γ is 1 for positive vertices and 0 for background pixels. $\mathcal{L}_{edge}$ is the edge loss, including the cross-entropy loss for edgeness towards the i-th 60 degree sector and the L2 loss of the relative offset of the i-th edge. The details of edge loss can be found in [22].

Third, we unlock the coase and fine decoders and finetune the model end-to-end. The overall loss is:

$$\mathcal{L}_{stage3} = \omega_1 \mathcal{L}_{stage1} + \omega_2 \mathcal{L}_{stage2} \tag{4}$$

where ω_1 and ω_2 are the weights for balancing the two decoders. The recovery of road networks from the predicted graph tensors exactly follows the work [22] and is omitted here.

3.3 Road Map Fusion

Algorithm 1 Framework of the trajectory-aerial fusion pipeline

Input: trajectory-based road network $\mathcal{G}^{trj}$, aerial-based road network $\mathcal{G}^{aer}$, original trajectory map $\mathcal{T}$
Output: Fused road network G^{fuse}

Stage **I**: First fusion; assume the fused road network is the union set of $\mathcal{G}^{trj}$ and $\mathcal{G}^{aer}$: $\mathcal{G}_I = \mathcal{G}^{trj} \cup \mathcal{G}^{aer}$.
for all (r_i, r_j), $r_i \in \mathcal{G}^{trj}$, $r_j \in \mathcal{G}^{aer}$ **do**
  if $\measuredangle(r_i, r_j) < 30^\circ$ and **SIM** $(r_i, r_j) > 0.7$ **then**
    delete the shorter road within $\{r_i, r_j\}$ from $\mathcal{G}_I$.
  end if
end for
Stage **II**: Cross-check fusion by revisiting $\mathcal{T}$.
for $tr \in \mathcal{T}$ **do**
  set dict[tr] = []
  for $r_i \in \mathcal{G}_I$ **do**
    compute the **SIM**=(r_i, tr) and add r_i to dict[tr] if **SIM**>0.7.
  end for
end for
Compute the LCSs for every $(tr_i, tr_j) \in \mathcal{T}$ and connect them to obtain the new road set $\mathcal{G}_{II}$.
Stage **III**: Automatic topology construction.
Find the intersection nodes within $\mathcal{G}_{II}$ and cut the road into segments if the angle between two adjoining segments smaller than 120°. The cut segments form a new road set $\mathcal{G}_{III}$.
for all $r_i \in \mathcal{G}_{III}$ **do**
  obtain the endpoint nodes of r_i and add to the set $\{\mathcal{N}\}$.
end for
Cluster the nodes in $\{\mathcal{N}\}$ into multiple subsets $\{\mathcal{N}_1\}$, $\{\mathcal{N}_2\}$, ..., $\{\mathcal{N}_n\}$ with a radius of 5 meters.
for each subset $\mathcal{N}_k \subset \mathcal{N}$ **do**
  Find the captain node cap_k according to Equation 5.
end for
Adjust and link the nodes within each $\{\mathcal{N}_k\}$ *w.r.t.* cap_k into a road subgraph $\mathcal{G}_k$.
Multiple subgraphs $\mathcal{G}_k$,$\mathcal{G}_{k+1}$, ... are combined into $\mathcal{G}^{fuse}$.
return $\mathcal{G}^{fuse}$

One of our critical designs is the aerial-trajectory fusion module based on revisiting infrequent trajectories under the guidance of aerial predictions. Previous works usually stack the GPS trajectory map to the aerial image as input of the neural networks [34] (*i.e.*, early fusion), or adopt two encoders for separate feature encoding and fuse the features with MLPs [27, 39] (*i.e.*, deep fusion). However, two challenges are existing. First, the network is prone to overfitting simple samples where the aerial images and the dense trajectories overlap. This is also admitted in [34], in which multiple data augmentation techniques are leveraged to alleviate the overfitting problem. Second, the underlying topology within the trajectory map can be overlooked if simply learned by a unified neural network. Since the aerial image is expected to extract road networks, especially in regions with sparse trajectories whereas the trajectory-based approach usually recovers highly reliable roads, we prefer to keep high recall for the aerial branch and keep high accuracy for the trajectory branch. Due to the variance of the two

modalities, if enforced within a unified framework, the network might search for a trade-off between the two features and attain a sub-optimal balance. In consequence, we alternatively propose the cross-check-based late-fusion strategy by first seeking the optimum of each modality independently and then fusing the two results by further revisiting the over-killed trajectories with the guidance of aerial predictions.

Our full fusion pipeline depicted in Figure 6 includes three main stages: (I) Direct fusion by combining the road networks $\mathcal{G}^{trj}$ and $\mathcal{G}^{aer}$ generated by trajectories and aerial images, respectively. We delete the shorter one if two roads overlap with each other (See Algorithm 1 Stage I) with the similarity larger than a predefined threshold of 0.7 and then we can obtain the road networks $\mathcal{G}_I$. (II) Cross-check fusion by revisiting the sparse trajectories under the guidance of $\mathcal{G}^{aer}$. In specific, we recover the road lines from the original trajectory map $\mathcal{T}$ if two conditions are satisfied: there is a road detected by the aerial image and there are more than two lines along the same road by computing the Longest-Common-Subsequence (LCSs) of two trajectories. Those who are over-killed in the extracted road network $\mathcal{G}^{trj}$ due to low trajectory density are now recovered with the guidance of aerial image prediction. (III) Automatic topology construction. First, the roads are cut into segments by intersections and we obtain the endpoints of the segments as nodes. Then the nodes are clustered into multiple subsets within a radius of 5 meters. Further, for each subset $\mathcal{N} = \{n_1, n_2, ..., n_N\}$, we find the captain node cap, which has the minimum summed distance to all the other nodes within the subset:

$$cap = \underset{k}{arg\min} \sum_{i=0}^{N} \|n_i - n_k\|_2 \tag{5}$$

where N is the number of the subset $\mathcal{N}$ and we traverse each node within $\mathcal{N}$ to obtain the captain node, $\|*\|_2$ denotes Euclidean distance. For each node within the subset, we adjust its location to connect with the captain node smoothly and form it into a sub-graph. Multiple sub-graphs are finally linked into the road networks $\mathcal{G}^{fuse}$. Detailed elaboration is depicted in Algorithm 1.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Dataset. We collect a large-scale dataset containing paired aerial images and spatial-temporal trajectory data from Baidu Maps. The dataset covers more than 80 cities in China, including Beijing, Guangzhou, and so on, with a total coverage area of 33,000 square kilometers, which is much larger than the City-Scale Dataset [22], SpaceNet Roads Dataset [36], and the dataset used in [34]. The trajectory data derived from the road network database are adaptively accumulated according to the statistic density. The generated trajectory map has the resolution of 5000 × 5000, with each pixel covering 2m × 2m. The corresponding aerial images are collected crowdsourced with 2048 × 2048 resolution, each covering 1km × 1km. There are 33,368 aerial images in total and are randomly split into training, validation, and testing set with the ratio of 7: 1: 2.

4.1.2 Comparison Methods. We conduct extensive comparison experiments with methods selected from three categories.

(1) Trajectory-based road extraction approaches. We select KDE [8], vanilla cluster-based approach as well as the trajectory-branch in DeepDualMapper[39] learned by U-Net-like structure [33] for comparing with our density-adaptive road extraction from the trajectories.

(2) Aerial image-based approaches. This branch mainly relies on deep neural networks for feature learning. To evaluate the effectiveness of our coarse-to-fine aerial image-based road extraction method, we select the segmentation-based approach Seg-Unet [33], D-linkNet [43] and the graph-based Sat2Graph [22] for comparison.

(3) Fusion-based approach from multimodal data.

To compare our cross-check-based fusion strategy with concurrent fusing strategies, we re-implement $early$, $deep$ and $late$ fusion for comparison. $Early$ fusion is implemented by concatenating the trajectory map as an additional channel to the RGB aerial image to form a 4-channel input, inspired by [34]. Differently, $deep$ fusion is conducted by first adopting a siamese Unet-like network for independent feature learning, with an aerial image and a trajectory map taken as input, respectively. Then the two features are fused to yield the final prediction. For a fair comparison, the Unet-structure with Resnet34 backbone is adopted for fused-based feature learning. In addition, we conduct the vanilla $late$ fusion by directly computing the overlapping roads between the extracted roads from trajectories and aerial images, respectively.

4.1.3 Evaluation Metrics and Hyberparameters. We adopt the relaxed precision, recall, and F1 scores to evaluate the performance of road extraction following [18]. For end-to-end evaluation, we adopt ρ meters as the distance threshold for detecting true positives (TP), to keep consistent with the aerial-based evaluation (ρ meter-threshold corresponds to $2*\rho$ pixels in aerial images). A predicted road is defined as TP only if it is predicted as a road, and the neighboring ρ meters around the centerline of the predicted road exist a ground truth road point. In all experiments, we train the networks by the Adam optimizer, with the learning rate 0.0001 and batch size as 2 to train 100k steps. S in Equation 1 is set as 32. ω_1 and ω_2 in Equation 4 are set to 0.3 and 0.7.

4.2 Overall Evaluation

4.2.1 Performance of different approaches. We conduct experiments on the collected large-scale real-world dataset from Baidu Maps to evaluate the performance of different approaches. The results are shown in Table 1. Traditional trajectory-based approaches such as KDE and clustering methods present inferior performance. U-Net$_{traj}$ derived from DeepDualMapper[39] improves the precision and recall with the power of convolutional neural networks. By contrast, thanks to our density-aware trajectory-based method by leveraging accumulated trajectory data and adaptive applying various operations for regions with various densities, the performance is improved to a great extent.

For aerial image-based approaches, graph-based Sat2Graph performs better than segmentation-based methods Unet and D-linknet. It has been validated in [22] that the improvement comes from the graph-tensor rather than a stronger backbone. Our coarse-to-fine aerial method performs best among all the aerial image-based methods as we make use of the coarse stage for vertices localization to extract neighboring local context for fusing with the global context.

Table 1: The performance of trajectory-based, aerial image-based and multimodal fusion based approaches on the collected large-scale trajectory-image paired dataset.

Modality	Method	Precision	Recall	F1
Trajectory	KDE	0.631	0.455	0.528
	Cluster-based	0.733	0.355	0.478
	U-Net$_{traj}$	0.762	0.698	0.729
	Our Trajectory branch	0.776	0.742	0.759
Aerial Image	Seg-Unet	0.597	0.672	0.632
	D-linknet	0.623	0.711	0.664
	Sat2Graph	0.741	0.827	0.782
	Our Aerial branch	0.748	0.849	0.795
Multi-modality	Early fusion	0.752	0.841	0.794
	Deep fusion	0.780	0.859	0.818
	Vanilla fusion	0.892	0.595	0.714
	Our DuARE	0.782	**0.881**	**0.829**

All in all, the multimodal approaches attain better performance than the single-modality methods, which demonstrates that different modalities of data have the potential to complement each other. Simply fusing the two modalities into a stacked input by early fusion might achieve little improvement and deep fusion better learns the feature representation of each modality by two different encoders, at the cost of more computation head. However, due to the variance lying in the two modalities, a unified network might search for a trade-off and achieve a sub-optimal balance. Alternatively, we pay attention to the late fusion strategy to keep the optimum of each modality independently. Note that adopting vanilla late fusion by directly computing the union of the outputs from two modalities achieves high precision since only the common roads that have been detected by both the trajectory and the aerial image will be preserved, resulting in a low recall. However, the road extraction pipeline prefers higher recall rather than higher precision since we expect to dig out the undiscovered roads. On the contrary, the proposed cross-check-based fusion strategy largely improves the recall while keeping high precision by revisiting the original trajectories under the guidance of aerial image prediction. The overall pipeline of our DuARE solution achieves superior performance over all the other approaches cost-effectively.

4.2.2 Qualitative visualization of different fusion policies. We show the visualization results of various fusion strategies in Figure 7 given the same input, *i.e.*, the two predictions from the trajectory map and the aerial image. Early fusion attains discontinuous roads and deep fusion over-segments the road networks. In contrast to the vanilla late fusion which directly intersects the two predictions from different modalities, our cross-check-based fusion obtains continuous roads and also recovers minor roads. This is because our fusion strategy keeps the optimum of each modality while revisiting the original trajectory map with the guidance of aerial prediction so as to recall the over-killed roads, which largely keeps the continuity and integrity.

4.2.3 Precision-Recall curves of different aerial image-based approaches. Figure 8 presents the precision-recall curves of different aerial image-based approaches. The graph-based approach Sat2Graph achieves better precision-recall over the segmentation-based approaches like Seg-Unet and D-linknet. Our aerial branch

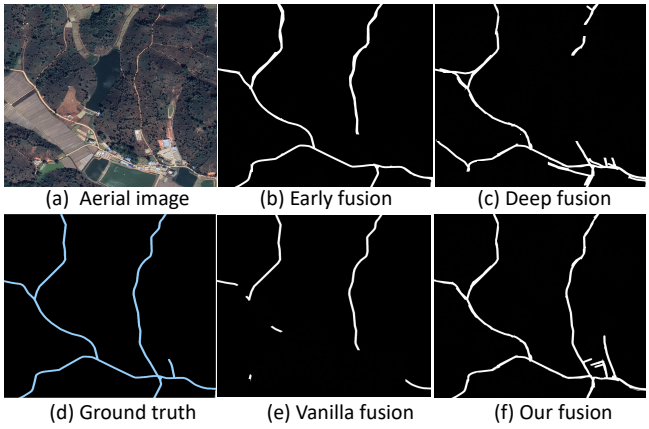

Figure 7: Qualitative visualization of different fusion policies.

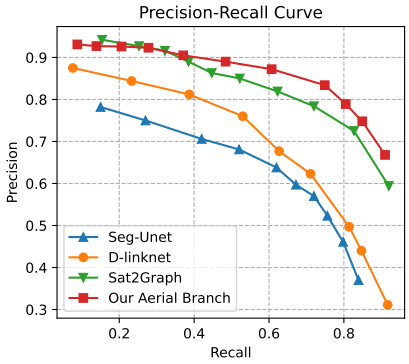

Figure 8: PR curves of different aerial image-based methods.

achieves superior performance over the other approaches at most positions, though a slightly lower precision is observed at an extremely low recall rate compared with Sat2Graph. We believe the advantage lies in the coarse-to-fine local and global context fusion.

4.3 Ablation Studies

4.3.1 Effect of the core components in aerial image-based approach. From the perspective of the aerial image-based approach, we conduct extensive ablation studies to evaluate the key designs of our aerial image-based method and present the results in Table 2. Group I is the baseline without the proposed coarse stage. *Vertex* and *seg* denote the vertexness and the roadness (segmentation) head at the specially designed coarse stage. The Fusion column denotes whether adopting concatenation or addition operation for fusing the local and global context. Since the two stages both predict the vertex information, we also conduct experiments on the adopted stage for final prediction.

From Group I and II, it is proved that simply adding the vertex detection head can not improve the performance since no semantic knowledge is leveraged. Comparing Group II and III, adopting both the discrete vertex regression and the continuous road segmentation brings a 5.1% improvement. Group III and V explore the specific fusion strategy in incorporating the local features extracted from

Table 2: Effects of core components in aerial image branch.

Group	Coarse Stage		Fusion		Adopted Stage		Precision	Recall	F1
	vertex	seg	concat	add	Coarse	Fine			
I						✓	0.741	0.827	0.782
II	✓			✓		✓	0.694	0.801	0.744
III	✓	✓		✓		✓	**0.748**	**0.849**	**0.795**
IV	✓	✓		✓	✓		0.609	0.695	0.649
V	✓	✓	✓			✓	0.746	0.845	0.792

Table 3: Effect of selecting various regions for local feature embedding. The second column denotes how many neighboring regions are extracted within the encoder of Figure 5.

Neighboring regions	local region number	Precision	Recall	F1
-	0	0.741	0.827	0.782
center-based	1	0.739	0.838	0.785
corner-based	4	0.740	0.842	0.788
center+corners	5	**0.748**	**0.849**	**0.795**

the low-level feature maps guided by coarse prediction and the deep-level features. The dims of the two features before fusion are $[H, W, C]$, and the fused features after concatenation and addition operation result to be $[H, W, 2C]$ and $[H, W, C]$, respectively. We find that add-based fusion performs slightly better than concatenation (0.3% gain), even with fewer parameters. Comparing Group III and IV, we find that the prediction from the coarse stage is inferior to the fine stage, which benefits from incorporating local and global context from both stages.

4.3.2 Effect of different neighboring regions for local features embedding. As is shown in Figure 5, we extract the neighboring local features given the coarse vertex prediction and then combine them with the deep-level features. To explore how the neighboring regions affect the performance, we conduct ablation studies on various neighboring regions in Table 3. The center-based strategy extracts the local region with a fixed window size centered at the coarse vertex and improves the recall by 1.1%. Furthermore, if adopting the coarse vertex as the top-right, top-left, bottom-right, and bottom-left corners and extracting corresponding neighboring regions, the recall observes a 1.5% gain compared to baseline. By adopting both center-based and corner-based neighboring regions, we obtain a noticeable improvement (+1.3% F1). We believe that incorporating the two kinds of neighboring regions helps to better learn the edgeness in the road graph.

5 RELATED WORK

5.1 Trajectory-based Road Extraction

Fueled by the ubiquitous GPS tracking data, there have been extensive attempts in trajectory-based road extraction methods. We refer the readers to [3] for a detailed literature review and comparison. The point clustering-based methods cluster discrete trajectory points to form continuous road segments by clustering strategies such as k-means algorithm [17] and the 2D Gaussian fitting [20]. Multiple neighboring distance measurements considering the location, heading and proximity can be applied, such as Voronoi diagrams, Delaunay triangulations, and Vietoris-Rips complex [1, 2, 9, 29, 38]. Kernel Density Estimation (KDE) methods are also utilized to transform the points to a density-based image [8, 14], which is robust against noise. The intersection-linking approaches first detect the intersecting vertices and then link them together to construct the road network. Multiple point characteristics can be leveraged for detecting the intersections, such as direction, speed, density and geometric features [26, 30, 44].

5.2 Aerial Images-based Road Extraction

With the advent of convolutional neural networks (CNNs) in recent years, there has been tremendous progress in semantic segmentation [5, 10, 42]. In consequence, we only briefly review the approaches leveraging deep neural networks rather than traditional methods [24, 37]. There are mainly two branches in the aerial image-based road extraction task: the semantic segmentation-based and the graph-based approaches. Segmentation-based methods formulate the task as a road segmentation problem from aerial images to predict the *roadness* probability and even the fine-grained categories of each pixel [4, 11, 15, 31, 41, 43]. For example, road-specific contextual information and road structured networks are utilized to preserve the continuity of road segments [12, 35]. Such methods rely on semantic knowledge and thus the performance can deteriorate in textureless regions or the tree-arched roads. Besides, they require post-processing strategies such as morphological thinning and line following to recover the topology from segments.

Alternatively, graph-based approaches learn the road graph directly from aerial images. Markov random field is parameterized to direct infer topologically correct roads [32]. RoadTracer [6] generates the road network graph iteratively, followed by recurrent neural networks for considering more context knowledge [13]. The recently presented Sat2Graph directly predicts the locations of vertices and edges of the road graph networks and achieves outperforming results [22]. RoadTagger [23] further utilizes both the CNN and Gated Graph Neural Networks to infer the road network.

5.3 Multimodal fusion-based Road Extraction

Single modality has its limitations and advantages, thus it is feasible to incorporate multimodal data for road extraction. There is limited work fusing both GPS trajectory and aerial images. The first idea is to render the trajectory map as an additional input channel stacked to the aerial image [34]. Alternatively, deep fusion is exploited in DeepDualMapper [28, 39], by first feeding the two modalities into two networks for feature extraction separately, and then fusing the learned features at multi-scale layers. However, due to the variance in the two modalities, a unified network might search for a trade-off and obtain a sub-optimal balance.

6 CONCLUSION

In this paper, we present an industrial solution: DuARE, for automatic road extraction at Baidu Maps from multimodal data: the abundant spatial-temporal trajectory data as well as the aerial images. We dig deep into the features of each modality and present novel designs including three modules. (1) Density-adaptive trajectory-based road extraction which utilizes the density distribution for indicating specific operations towards different regions. (2) Coarse-to-fine road extraction from aerial images, so as to involve local and global context. (3) The road map fusion via cross-checking through

the trajectory map under the guidance of aerial predictions. Extensive experiments conducted on the collected large-scale real-world multimodal dataset from Baidu Maps validate the superiority of the proposed solution. DuARE has been deployed in production at Baidu Maps since June 2021 and keeps updating the road networks by 100,000 km per month.

REFERENCES

[1] Mridul Aanjaneya, Frederic Chazal, Daniel Chen, Marc Glisse, Leonidas J Guibas, et al. 2011. Metric graph reconstruction from noisy data. In *Proceedings of the twenty-seventh annual symposium on Computational geometry*. 37–46.

[2] Gabriel Agamennoni, Juan I Nieto, and Eduardo M Nebot. 2010. Robust inference of principal road paths for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems* 12, 1 (2010), 298–308.

[3] Mahmuda Ahmed, Sophia Karagiorgou, Dieter Pfoser, and Carola Wenk. 2015. A comparison and evaluation of map construction algorithms using vehicle tracking data. *GeoInformatica* 19, 3 (2015), 601–632.

[4] Rasha Alshehhi, Prashanth Reddy Marpu, Wei Lee Woon, and Mauro Dalla Mura. 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing* 130 (2017), 139–149.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.

[6] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. 2018. Roadtracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4720–4728.

[7] Anil Batra, Suriya Singh, Guan Pang, Saikat Basu, CV Jawahar, and Manohar Paluri. 2019. Improved road connectivity by joint learning of orientation and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 10385–10393.

[8] James Biagioni and Jakob Eriksson. 2012. Map inference in the face of noise and disparity. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. 79–88.

[9] Daniel Chen, Leonidas J Guibas, John Hershberger, and Jian Sun. 2010. Road network reconstruction for organizing paths. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 1309–1320.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, et al. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[11] Guangliang Cheng, Ying Wang, Shibiao Xu, Hongzhen Wang, Shiming Xiang, and Chunhong Pan. 2017. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing* 55, 6 (2017), 3322–3337.

[12] Guangliang Cheng, Chongruo Wu, Qingqing Huang, Yu Meng, Jianping Shi, Jiansheng Chen, and Dongmei Yan. 2019. Recognizing road from satellite images by structured neural network. *Neurocomputing* 356 (2019), 131–141.

[13] Hang Chu, Daiqing Li, David Acuna, Amlan Kar, Maria Shugrina, Xinkai Wei, Ming-Yu Liu, Antonio Torralba, and Sanja Fidler. 2019. Neural turtle graphics for modeling city road layouts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4522–4530.

[14] Jonathan J Davies, Alastair R Beresford, et al. 2006. Scalable, distributed, real-time map generation. *IEEE Pervasive Computing* 5, 4 (2006), 47–54.

[15] Ilke Demir, Forest Hughes, Aman Raj, Kleovoulos Tsourides, Divyaa Ravichandran, Suryanarayana Murthy, Kaunil Dhruv, et al. 2017. Robocodes: Towards generative street addresses from satellite imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition Workshops*. 1–10.

[16] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6569–6578.

[17] Stefan Edelkamp and Stefan Schrödl. 2003. Route planning and map inference with global positioning traces. In *Computer science in perspective*. 128–151.

[18] Marc Ehrig and Jérôme Euzenat. 2005. Relaxed precision and recall for ontology matching. In *Proc. K-Cap 2005 workshop on Integrating ontology*. No commercial editor., 25–32.

[19] Xiaomin Fang, Jizhou Huang, Fan Wang, Lingke Zeng, Haijin Liang, and Haifeng Wang. 2020. ConSTGAT: Contextual Spatial-Temporal Graph Attention Network for Travel Time Estimation at Baidu Maps. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2697–2705.

[20] Tao Guo, Kazuaki Iwamura, and Masashi Koga. 2007. Towards high accuracy road maps generation from massive GPS traces data. In *2007 IEEE international geoscience and remote sensing symposium*. IEEE, 667–670.

[21] Mordechai Haklay and Patrick Weber. 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive computing* 7, 4 (2008), 12–18.

[22] Songtao He, Favyen Bastani, Satvat Jagwani, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Mohamed M Elshrif, Samuel Madden, and Mohammad Amin Sadeghi. 2020. Sat2Graph: road graph extraction through graph-tensor encoding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 51–67.

[23] Songtao He, Favyen Bastani, Satvat Jagwani, Edward Park, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Samuel Madden, and Mohammad Amin Sadeghi. 2020. RoadTagger: Robust road attribute inference with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10965–10972.

[24] Stefan Hinz and Albert Baumgartner. 2003. Automatic extraction of urban road networks from multi-view aerial imagery. *ISPRS journal of photogrammetry and remote sensing* 58, 1-2 (2003), 83–98.

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[26] Sophia Karagiorgou and Dieter Pfoser. 2012. On vehicle tracking data-based road network generation. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. 89–98.

[27] Yali Li, Longgang Xiang, Caili Zhang, Fengwei Jiao, and Chenhao Wu. 2021. A Guided Deep Learning Approach for Joint Road Extraction and Intersection Detection From RS Images and Taxi Trajectories. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), 8008–8018.

[28] Lingbo Liu, Zewei Yang, Guanbin Li, Kuo Wang, Tianshui Chen, et al. 2021. Aerial Images Meet Crowdsourced Trajectories: A New Approach to Robust Road Extraction. *arXiv preprint* (2021).

[29] Xuemei Liu, James Biagioni, Jakob Eriksson, Yin Wang, George Forman, and Yanmin Zhu. 2012. Mining large-scale, sparse GPS traces for map inference: comparison of approaches. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 669–677.

[30] Radu Mariescu-Istodor and Pasi Fränti. 2018. CellNet: Inferring road networks from GPS trajectories. *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 4, 3 (2018), 1–22.

[31] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. 2017. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3438–3446.

[32] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. 2015. Enhancing road maps by parsing aerial images around the world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1689–1697.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[34] Tao Sun, Zonglin Di, Pengyu Che, Chun Liu, et al. 2019. Leveraging crowdsourced gps data for road extraction from aerial imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7509–7518.

[35] Chao Tao, Ji Qi, Yansheng Li, Hao Wang, and Haifeng Li. 2019. Spatial information inference net: Road extraction using road-specific contextual information. *ISPRS Journal of Photogrammetry and Remote Sensing* 158 (2019), 155–166.

[36] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. 2018. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232* (2018).

[37] Jan Dirk Wegner, Javier Alexander Montoya-Zegarra, and Konrad Schindler. 2015. Road networks as collections of minimum cost paths. *ISPRS Journal of Photogrammetry and Remote Sensing* 108 (2015), 128–137.

[38] Stewart Worrall and Eduardo Nebot. 2007. Automated process for generating digitised maps through GPS data compression. In *Australasian Conference on Robotics and Automation*, Vol. 6. Brisbane: ACRA.

[39] Hao Wu, Hanyuan Zhang, Xinyu Zhang, Weiwei Sun, Baihua Zheng, and Yuning Jiang. 2020. DeepDualMapper: A gated fusion network for automatic map extraction using aerial images and trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1037–1045.

[40] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. 2018. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2403–2412.

[41] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. 2018. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters* 15, 5 (2018), 749–753.

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2881–2890.

[43] Lichen Zhou, Chuang Zhang, and Ming Wu. 2018. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE conference on computer vision and pattern recognition Workshops*. 182–186.

[44] S Zourlidou and Monika Sester. 2016. Intersection detection based on qualitative spatial reasoning on stopping point clusters. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives 41 (2016)* 41 (2016), 269–276.

A APPENDIX

A.1 Quantitative results on the public dataset

In addition to the presented performance evaluated on the collected large-scale dataset, we also compare the performance on the public aerial image dataset SpaceNet Roads Dataset[36] - for further comparison.

SpaceNet Roads Dataset is a large corpus of labeled satellite imagery on Amazon Web Services (AWS) and is initially designed for competitions including automated building footprint extraction and road network extraction. Note that the ground truth of the testing data in the dataset is not public. Following the same splitting with Sat2Graph[22], we split the 2549 tiles (non-empty) of the original training dataset into training (80%), testing (15%) and validating (5%) sets for comparison. On the public dataset, we select two branches of approaches for comparison.

(1) Segmentation-based approaches:

- Seg-UNet[33], which adopts U-Net-like CNN structure for learning the segmentation given aerial image as input.
- Seg-DRM[31] adopts a larger CNN backbone to predict the initial road segmentation, and then reasons about missing connections in the extracted road topology as a shortest path problem. As stated in [22], the original soft-IoU loss is replaced with cross-entropy loss for regressing the topology, since the latter achieves better performance.
- Seg-Orientation[7] develops a multi-branch convolutional framework to utilize the mutual information between orientation learning and road segmentation tasks.

(2) Graph-based approaches:

- RoadTracer [6] uses an iterative search process guided by a automatically construct accurate road network maps from a CNN-based decision function to derive the road network graph directly from the network outputs, which discards the complicated post-process procedures.
- Sat2Graph [22] acts as the baseline of our aerial-based road extraction branch.

As is shown in Table 4, RoadTracer seems to achieve a lower F1-score compared to Seg-DRM and Seg-Orientation, since it mainly focuses on the coarse road connectivity and results in a lower recall. Compared with Sat2Graph, our coarse-to-fine aerial image-based road extraction method achieves similar precision and a higher recall, leading to an overall superior performance.

Table 4: The performance of road extraction approaches from aerial images on the public SpaceNet Roads Dataset.

Branch	Method	Precision	Recall	F1
Seg-based	Seg-Unet	0.690	0.663	0.676
	Seg-DRM	0.828	0.726	0.773
	Seg-Orientation	0.816	0.714	0.761
Graph-based	RoadTracer	0.786	0.625	0.696
	Sat2Graph	**0.859**	0.766	0.810
	Our Aerial branch	**0.859**	**0.778**	**0.816**

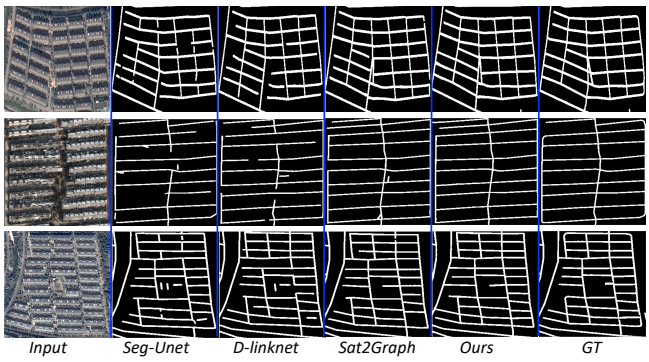

Figure 9: The qualitative results of different aerial image-based approaches on the dataset collected by Baidu Maps.

A.2 Ablation study of the trajectory branch

We conduct ablation studies to evaluate the effect of three adaptive parts in Figure 4 and present the results in Table 5. From Group I and VI, the adaptive accumulation policy obtains +0.78% F1 gain compared with adopting fixed months of trajectory data (6 months). Given Group II, III and VI, we find that adaptive filtering policy performs better than a single strategy like morphological or Gaussian filter. Analogously, the third adaptive thinning policy also surpasses single convolutional-based erasing strategy and density-distribution-based thinning by 0.25% and 0.2% in F1 score.

A.3 Qualitative results

We present more qualitative examples to demonstrate the effectiveness of our approach.

A.3.1 Qualitative visualization of different aerial image-based approaches. We present the visualization results of aerial image-based methods in Figure 9. The segmentation-based approaches such as Seg-Unet, D-linknet fail to recover the details, and the graph-based method Sat2Graph contains more false positives. By contrast, our coarse-to-fine approach demonstrates more reasonable results due to the local and global context fusion aided by the coarse stage.

A.3.2 Qualitative visualization of different fusion strategies. As is depicted in Figure 10, we give more examples comparing different fusion policies. In contrast to the vanilla late fusion which directly intersects the two predictions from different modalities, our cross-check-based fusion obtains continuous roads and also recovers minor roads.

A.3.3 Qualitative visualization of the DuARE solution. As is shown in Figure 11, given the aerial images as input, we depict the extracted road by the proposed DuARE solution colored in blue. As can be seen, the extracted lines are visually well-aligned with the roads.

Given the aerial images as input, we depict the extracted road by the proposed DuARE solution colored in blue. As can be seen, the extracted lines are visually well-aligned with the roads.

Table 5: Effects of adaptive parts in the trajectory branch of the proposed solution.

Group	Adaptive Accum.		Adaptive Filter			Adaptive Thinning			Precision	Recall	F1
	fixed	adaptive	morp.	Gaussian	adaptive	conv-erase	density	adaptive			
I	✓				✓			✓	0.695	0.668	0.681
II		✓	✓					✓	0.719	0.738	0.728
III		✓		✓				✓	0.747	0.703	0.724
IV		✓			✓	✓			0.758	0.693	0.724
V		✓			✓		✓		0.743	0.735	0.739
VI		✓			✓			✓	**0.776**	**0.742**	**0.759**

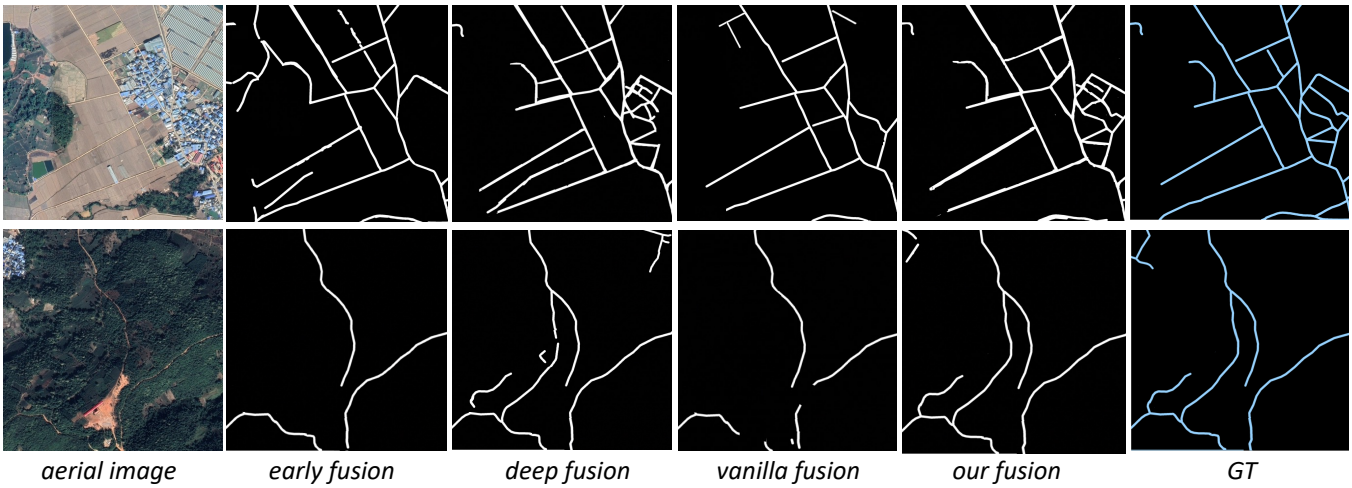

Figure 10: Qualitative visualization of different fusion policies.

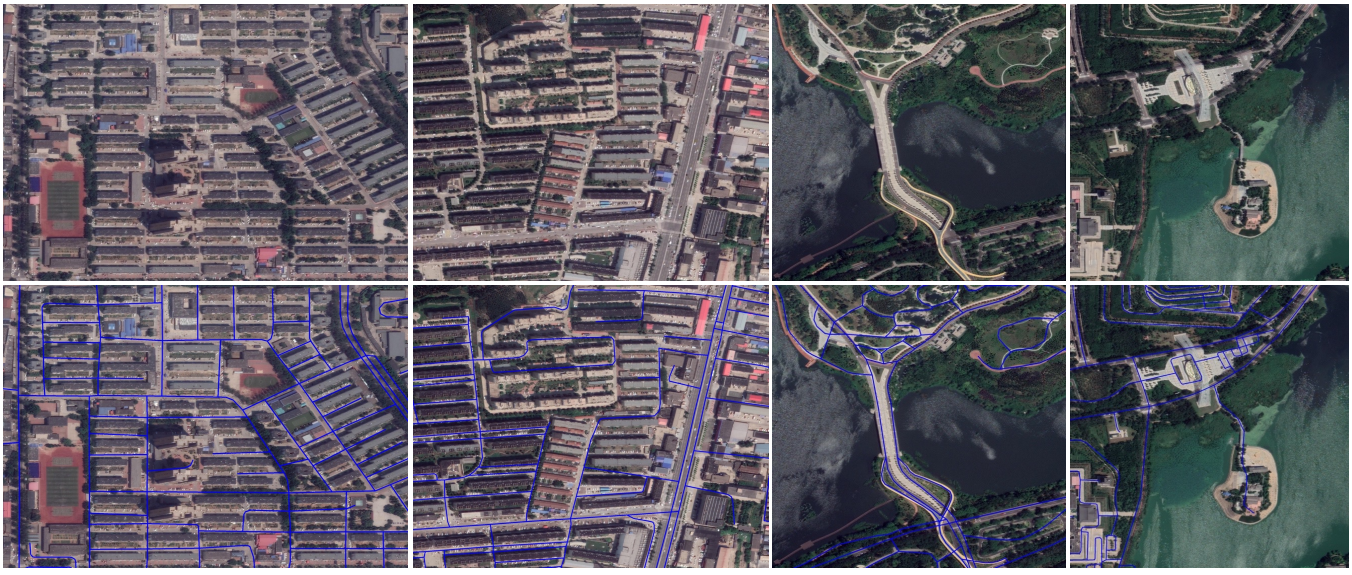

Figure 11: Qualitative visualization of our DuARE road extraction method. Top: original image, bottom: the extracted road painted in blue line on the image. (Best viewed when zoomed in.)