Game4Loc: A UAV Geo-Localization Benchmark from Game Data (Supplementary)

Yuxiang Ji[1]*, Boyong He[1]*, Zhuoyue Tan[1], Liaoni Wu[1,2]†

[1]Institute of Artifcial Intelligence, Xiamen University
[2]School of Aerospace Engineering, Xiamen University
yuxiangji@stu.xmu.edu.cn, wuliaoni@xmu.edu.cn

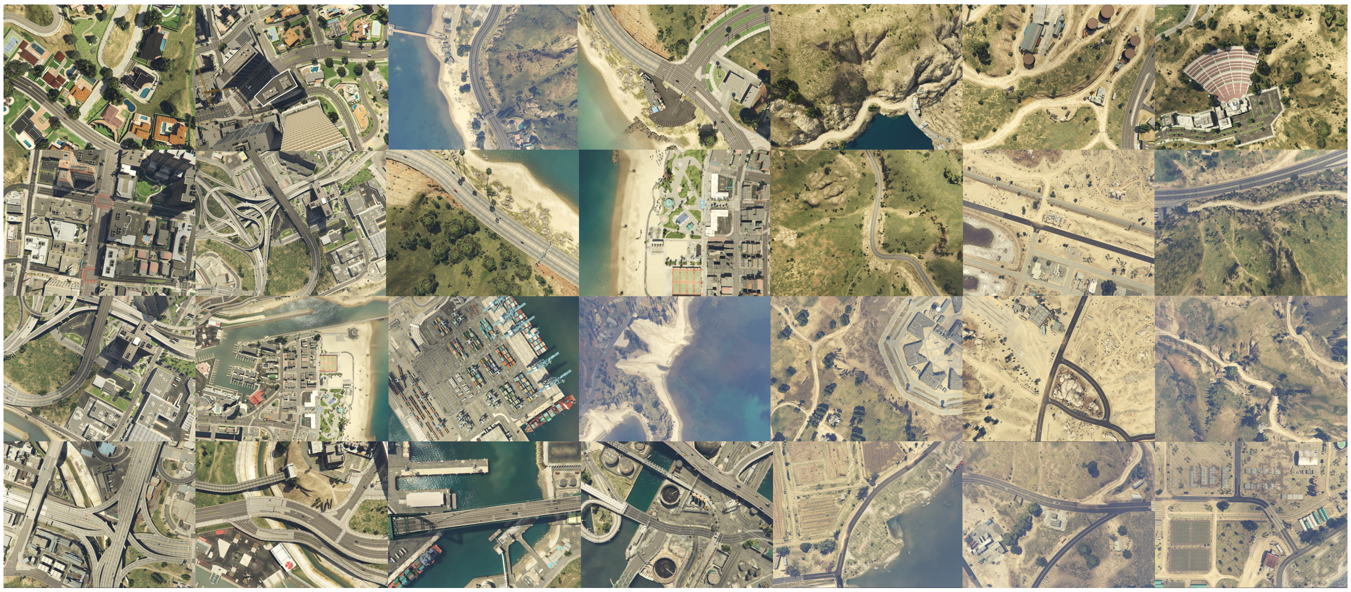


Figure 1: Example drone-view images of GTA-UAV dataset, from multiple scenes, altitudes, and attitudes.

A Introduction

This document provides supplementary materials for the main paper. Fig. 1 shows some drone-view examples of our GTA-UAV dataset from multiple scenes, altitudes, and attitudes. Specifically, Sec. B offers more details about organization and statistics of GTA-UAV. Sec. C provides more experiment setup details about training and testing pipeline. Additional experiments and corresponding analysis are presented in Sec. D. Some visualization is provided in Sec. E. The limitation and discussion are put in Sec. F.

B GTA-UAV Dataset

Dataset Collection

In GTA-UAV dataset, 33,763 drone-view images are collected from the whole game map of commercial video game GTAV. We use an open-source automated framework *Deep-GTA* (Kiefer, Ott, and Zell 2022), to simulate UAV flights at various altitudes and attitudes, then collect drone-view images. The simulated flight scenes cover city, mountain, desert, forest, field, and seaside, with flight altitudes ranging from $80m$ to $650m$ and various flight attitudes. Compared to existing UAV geo-localization datasets (Zheng, Wei, and Yang 2020; Zhu et al. 2023; Dai et al. 2023; Xu et al. 2024), this provides an opportunity to explore more complex and comprehensive localization tasks. For each drone-view data instance, we provide comprehensive meta-data, including:

- GPS information
- Flight altitude
- Flight attitude
- Camera parameters
- Paired satellite-view images

Based on the GPS information and camera attitude, the satellite-view images are paired by calculating the IOU of the approximate ground coverage from two FOV. The related ground coverage could be approximated by Eq. 5.

$$W = 2 \times h \times \tan(\frac{FOV_h}{2}), L = 2 \times h \times \tan(\frac{FOV_v}{2}) \quad (1)$$

$$R_z(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

$$R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \tag{3}$$

$$R_x(\psi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix} \tag{4}$$

$$\text{Corners}_{\text{related}} = R_z(\phi) \cdot R_y(\theta) \cdot R_x(\psi) \cdot \begin{bmatrix} W/2 & -W/2 & -W/2 & W \\ L/2 & L/2 & -L/2 & -L/2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5}$$

Dataset Statistics

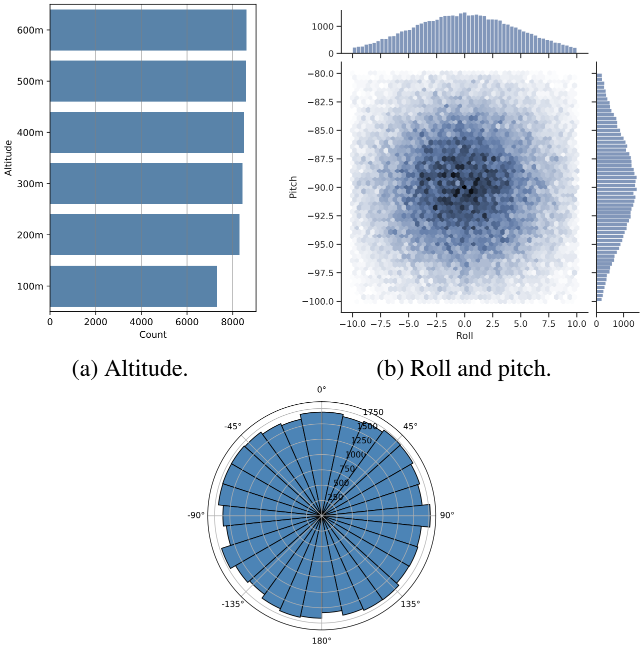
As illustrated in Fig. 2, we collect drone-view images focused on six altitude categories ranging from $80m$ to $650m$. We set the camera at a 90° on the pitch axis, making it perpendicular to the ground. Throughout the simulated flight, the drone-view is normally distributed on the roll and pitch axes with ranges of $[-10°, 10°]$ and $[-100°, -80°]$, respectively. The yaw angle is randomly distributed. For scene categories, we collect scenes including *urban, suburban, mountain, forest, coast, and sea area*. In data collection, the *urban* category occupies a major portion, which is considered for two main reasons: (1) The urban areas in the game map have more details and less homogenization; (2) Most existing datasets focus on urban areas, making it easier to align with them to evaluate generalization capabilities. The data also contains many areas that are difficult to recognize, such as pure sea areas. We did not exclude them currently for mainly two reasons: (1) One reason is that in the cross-area setting, the sea data is actually mainly distributed in the training set, and the reason for the failure in the cross-area setting is more the generalization ability to unseen scenes; (2) Another reason is that this phenomenon actually reflects some bad cases of visual localization of drones in real-world applications (that is, the recognition of some scenes is unrealistic).

The samples from same-area and cross-area settings we mentioned in main paper are shown in Fig. 3. As the same illustration in the main paper, the training and test drone-view images from the same-area setting are sampled from the sharing area, while the training and test drone-view images from the cross-area setting are sampled from seperate areas.

C Experiment Setup Details

Including Dataset

University-1652 University-1652 dataset (Zheng, Wei, and Yang 2020) consists of 37,854 drone-view images and 951 satellite-view images from 701 university buildings, where the task is to match the drone-view to the according satellite-view and vice versa. All images are collected through Google Earth simulation, and the image pairs have a strict one-to-one aligned correspondence.



(a) Altitude.

(b) Roll and pitch.

(c) Yaw.

Figure 2: Data statistics of GTA-UAV.

SUES-200 SUES-200 dataset (Zhu et al. 2023) contains 24,120 drone-view images acquired by the drone at four different heights and only 200 corresponding satellite-view of the same target scene. The data is also collected from Google Earth simulation, using a discrete sampling method, and cannot be extended from retrieval to localization tasks. Due to its limited satellite-view images, the retrieval task on this dataset is relatively simple and lacks practical significance.

DenseUAV In DenseUAV dataset (Dai et al. 2023), 9k drone-view and 18k satellite-view images of 14 university campuses are collected, where drone-view images are captured in the real-world low-alltitude setting. Due to its dense sampling method, localization tasks can be performed on this data, while the pairing process still follows a perfect matching format.

UAV-VisLoc UAV-VisLoc (Xu et al. 2024) includes 6,742 high-altitude drone-view images and 11 satellite maps covering the 11 corresponding areas. The released available dataset and technical report do not tiles satellite maps or pair drone-view images with satellite-view images.

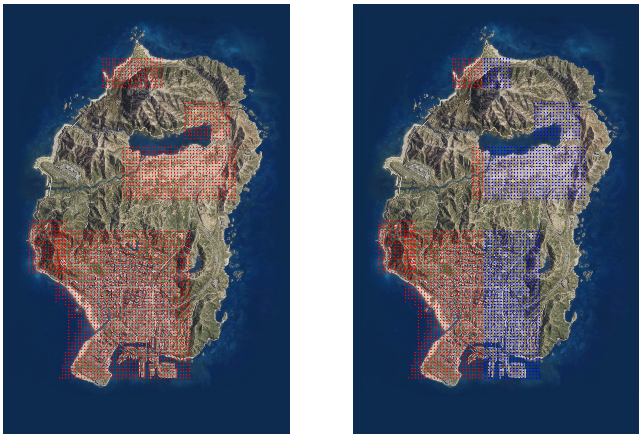
Evaluation Metrics

Spatial Continuity Index (SDM@K) Following DenseUAV (Dai et al. 2023), the SDM@k is an evaluation metric combines the characteristics of Recall@k while also considering the performance of localization, which is defined by Eq. 6:

$$\text{SDM}_k = (K - k + 1)/\exp(s \times d_k), \tag{6}$$

Table 1: Performance on GTA-UAV with label smoothing.

Methods	Cross-Area					Same-Area				
	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
$\mathcal{L}_{\text{InfoNCE}}$, $\epsilon = 0.05$	53.44%	79.55%	64.51%	73.69%	402.42m	77.55%	95.47%	85.40%	87.64%	197.06m
$\mathcal{L}_{\text{InfoNCE}}$, $\epsilon = 0.1$	53.40%	79.79%	64.60%	74.16%	395.55m	75.49%	95.38%	84.36%	87.72%	193.98m
$\mathcal{L}_{\text{InfoNCE}}$, $\epsilon = 0.2$	52.75%	79.71%	64.13%	73.85%	393.86m	74.85%	95.65%	83.87%	87.97%	189.23m
$\mathcal{L}_{\text{weighted-InfoNCE}}$	55.91%	81.07%	66.56%	76.35%	342.05m	84.95%	97.59%	90.15%	88.03%	149.07m



(a) Same Area.

(b) Cross Area.

Figure 3: Data samples coverage in the whole game map. In the same-area setting (left), red dots represent training and test samples, where they share the same area. In the cross-area setting (right), red dots represent training samples and blue dots represent test samples, where they are sampled from cross area.

where $d_i = \sqrt{(x_q - x_i)^2 + (y_q - y_i)^2}$, and $K - k + 1$ is the weight of $k - th$ result. In our experiments setup, all distances are in terms of meters, and the s is set to 0.001.

Implementation Details

In our GTA-UAV dataset, there are 33,763 drone-view images and 14,640 satellite-view images. In the cross-area setting, we divide the entire game map into two mutually exclusive groups, with 15,693 drone-view images used for training and the other 18,070 drone-view images used for testing. In the same-area setting, both the training set and the test set are sampled from the entire game map, with 26,964 drone-view images used for training and the reamining 10,799 drone-view images used for testing. All 14,640 multi-level satellite-view images are used as the retrieval gallery during test.

To evaluate the transferability and generalization capabilities of the GTA-UAV dataset and the proposed method for UAV geo-localization tasks, we conduct trasnfer experiments on a recently released dataset UAV-VisLoc (Xu et al. 2024), as discussed in the main paper. We select 7 regions totaling 4,427 drone-view images according to the area size and image quality as validation data for transferability. Based on the undefined dataset, we divide the 7 satellite maps using an identical tiling methods with GTA-UAV, and pair them by estimating the IOU between the FOV of two views. By testing different pre-trained models on the UAV-VisLoc dataset in both zero-shot and fine-tuned settings, we demonstrate the trasnferability of the GTA-UAV data in UAV geo-localization tasks that resemble real-world scenarios in the main paper.

D More Experiments

Analysis of Weighted InfoNCE

As formulated in Eq. 7, Eq. 8, and Eq. 9, our proposed weighted InfoNCE can be viewd as a form of label smoothing based on weights for standard InfoNCE.

$$\begin{aligned}\mathcal{L}_{\text{InfoNCE}}(F_q, \epsilon, F_R) = \\ &- \epsilon \log \frac{\exp(F_q \cdot F_{r^+}/\tau)}{\sum_i^R \exp(F_q \cdot F_{r_i}/\tau)} \\ &- (1-\epsilon)\frac{1}{|R|}\sum_i^R \log \frac{\exp(F_q \cdot F_{r_i^{+,-}}/\tau)}{\sum_j^R \exp(F_q \cdot F_{r_j}/\tau)} \\ &= \epsilon \mathcal{L}_{\text{InfoNCE}}(F_q, F_R) + (1-\epsilon)\mathcal{L}_{\text{uniform-InfoNCE}}(F_q, F_R)\end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{L}_{\text{weighted-InfoNCE}}(F_q, \alpha_q, F_R) = \\ &- \alpha_q \log \frac{\exp(F_q \cdot F_{r^+}/\tau)}{\sum_i^R \exp(F_q \cdot F_{r_i}/\tau)} \\ &- (1-\alpha_q)\frac{1}{|R|}\sum_i^R \log \frac{\exp(F_q \cdot F_{r_i^{+,-}}/\tau)}{\sum_j^R \exp(F_q \cdot F_{r_j}/\tau)} \\ &= \alpha_q \mathcal{L}_{\text{InfoNCE}}(F_q, F_R) + (1-\alpha_q)\mathcal{L}_{\text{uniform-InfoNCE}}(F_q, F_R)\end{aligned} \tag{8}$$

$$\alpha_q = \sigma(k, \text{IOU}_{qr^+}) = \frac{1}{1 + \exp(-k \times \text{IOU}_{qr^+})} \tag{9}$$

Unlike label smoothing, which uses a fixed hyperparameter ϵ for flexibility, our weighted InfoNCE calculates positive weights α_q based on IOU. In this way, the degree of smoothing could be adaptively controlled through positive weights α_q, leading to more flexible learning targets. As results in Tab. 1, our weighted-InfoNCE could improve the results on both retrieval and localization metrics under different label smoothing settings.

Table 2: Performance at different altitudes.

Flight Altitude			Cross-Area					Same-Area				
100m+200m	300m+400m	500m+600m	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
✓	-	-	52.39%	77.00%	63.08%	76.24%	419.22m	74.63%	93.12%	82.76%	89.79%	233.66m
-	✓	-	59.95%	84.82%	71.04%	74.57%	355.32m	87.81%	99.15%	93.06%	88.45%	115.61m
-	-	✓	55.32%	82.33%	66.18%	70.61%	403.67m	91.43%	99.90%	95.22%	84.53%	160.66m

Table 3: Performance comparison of models at different scales.

Approach	#Params	#Flops	Shared W.	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
Cross-Area								
ViT-Small/16	21.6M	12.4G	✓	49.79%	77.44%	62.25%	72.57%	430.49m
ViT-Medium/16	38.9M	22.0G	✓	52.67%	79.67%	64.13%	73.31%	420.01m
ViT-Base/16	64.3M	37.1G	×	42.09%	71.97%	54.84%	67.34%	547.26m
ViT-Base/16	64.3M	37.1G	✓	55.91%	81.07%	66.56%	76.35%	342.05m
ViT-Large/16	307.2M	174.8G	✓	59.26%	83.42%	69.73%	76.89%	337.29m
Same-Area								
ViT-Small/16	21.6M	12.4G	✓	82.49%	95.74%	88.71%	86.92%	225.51m
ViT-Medium/16	38.9M	22.0G	✓	82.87%	96.49%	89.04%	87.19%	211.12m
ViT-Base/16	64.3M	37.1G	×	85.20%	96.82%	90.30%	87.68%	167.46m
ViT-Base/16	64.3M	37.1G	✓	84.95%	97.59%	90.15%	88.03%	149.07m
ViT-Large/16	307.2M	174.8G	✓	85.01%	97.85%	90.36%	88.58%	134.22m

Impact of Altitudes

Flight altitude is an important variable in UAV geo-localization, as it directly affects the scale and the amount of scenery within the FOV. Considering the image retrieval task, different flight altitudes primarily have the following two effects. (1) The first effect is the scale difference between drone-view and satellite-view images. In the satellite-view database we construct, we use four zoom level tiles as the reference map. These four zoom levels cover an image scale range of approximately 70m to 560m, with a scale difference of a factor of 2 between consecutive levels. This means that within the continuously distributed UAV flight altitudes ranging from 80m to 650m, some images will have a high scale match with the reference map, while others will exhibit scale differences. (2) The second effect is the difference in scene appearance across different scales. When the flight altitude is low, there are fewer objects and scenes within the FOV for matching, increasing the probability of misjudgment. As the flight altitude increases, the resolution of scenery within the FOV decreases, potentially leading to a loss of fine-grained information. Existing UAV geo-localization datasets derived from real captures or Google Map simulations are often limited in altitude, which prevent this critical issue from being addressed. Our GTA-UAV dataset covers a wider range of flight altitudes, introducing flight altitude as a variable in this task. To validate the effect of data at different altitudes on geo-localization tasks, we divide the training set into three groups: $100m + 200m$, $300m+400m$, and $500m+600m$, then evaluate their performance separately. The results in Tab. 2 show that the model's performance varies across data from different altitudes. Notably, in the training with $500m+600m$ data, the larger coverage area of individual images results in more virtual edges between images, leading to fewer available training samples after mutually exclusive sampling. This is unfavorable for mining hard negatives in contrastive learning, thereby limiting the effectiveness of the learning process.

Model Scale Evaluation

As a supplement to the evaluation of different architectures in the main paper, we expand the experiments to the Vision Transformer (ViT) (Dosovitskiy et al. 2021) models with different parameter scales, as detailed in Tab. 3. All these models are with a patch-size 16 × 16. The results show that the performance follows a certain form of scaling law. In addition, the results with shared-weight encoder for cross-view images are better than without it, especially in the cross-area setting. This is because weight sharing can be viewed as a natural regularization term for cross-view models.

Satellite to Drone Retrieval

In addition to the original UAV geo-localization task (drone-view to satellite-view retrieval, D2S), we also extend the task to include satellite-view to drone-view (S2D) retrieval. This means finding the closest drone-view image given a satellite-view image. Tab. 4 shows the results comparing different training methods, where the proposed weighted-InfoNCE with Mutual Exclusive Sampling achieves the best performance on both cross-area and same-area settings. Fig. 4 shows the meter-level localization accuracy based on different thresholds.

Table 4: Performance of Sattellite-view to Drone-view (S2D) retrieval on GTA-UAV comparing different training methods. MES means Mutual Exclusive Sampling.

Methods	Cross-Area					Same-Area				
	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
Positive-only										
Triplet Loss ($\mathcal{L}_{\text{triplet}}$)	76.43%	86.37%	65.54%	84.18%	371.16m	86.41%	96.54%	87.46%	88.86%	135.84m
InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$)	77.87%	88.09%	65.82%	85.60%	367.02m	88.03%	97.23%	89.97%	89.61%	107.47m
InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$, w/. MES)	79.19%	89.50%	67.92%	86.27%	344.99m	89.36%	97.31%	90.65%	90.07%	100.66m
Ours ($\mathcal{L}_{\text{weighted-InfoNCE}}$, w/. MES)	**82.69%**	89.69%	**72.05%**	87.10%	323.36m	**93.97%**	98.97%	**94.61%**	91.24%	68.13m
Positive + Semi-positive										
Triplet Loss ($\mathcal{L}_{\text{triplet}}$)	45.86%	62.25%	37.39%	75.68%	560.06m	77.18%	98.21%	82.23%	**94.19%**	52.71m
InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$)	59.12%	76.06%	52.49%	85.71%	312.56m	78.08%	98.97%	82.74%	94.00%	55.49m
InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$, w/. MES)	63.90%	78.64%	53.94%	84.86%	309.55m	82.05%	97.95%	85.36%	94.02%	53.95m
Ours ($\mathcal{L}_{\text{weighted-InfoNCE}}$, w/. MES)	77.72%	**90.61%**	69.53%	**89.02%**	**236.02m**	91.41%	**99.62%**	92.94%	94.17%	**50.90m**

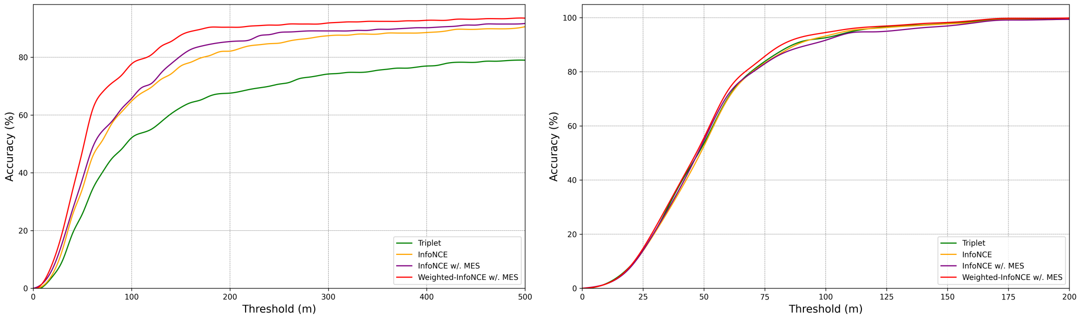


Figure 4: Meter-level localization accuracy of different methods (S2D) on cross-area (left) and same-area (right).

E Visualization

Feature Visualization

Human reasoning for cross-view geo-localization is typically based on landmarks or regions of interests within FOV. To further illustrate how the model performs drone-view to satellite-view matching, we extract the last hidden state of the ViT model (Dosovitskiy et al. 2021) and visualize it by averaging and unpatching, as shown in the Fig. 5. In both positive and semi-positive pairs, we could see that the model primarily focuses on regions with distinguishable features, such as buildings, vegetation, and roads.

Retrieval and Localization Examples

Here we provide some examples of the retrieval and localization results of GTA-UAV to show a better understanding of the task design and the data content in Fig. 6 ∼ 21.

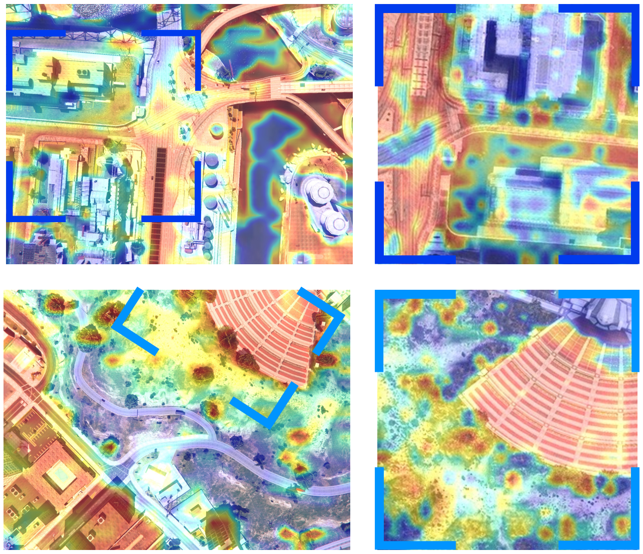


Figure 5: Feature heatmap for (positive) pair and (semi-positive) pair. The two pictures on the left side shows the drone-view query, and the pictures on the right side are the corresponding positive/semi-positive satellite-view.

F Limitations and Discussion

Our dataset aims to align with real-wolrd UAV geo-localization tasks, moving beyond the perfect one-to-one matching strong assumption in existing datasets and covering multi-altitude, attitude, and scene data that is difficult to capture in real datasets. Nonetheless, the scenes in current game maps are still highly homogeneous, filled with a large amount of similar data, and lack the diversity found in the real world. On the other hand, the imaging effects, object scales, and visual styles within the game world still exhibit certain biases compared to the real world. To further align with real-world tasks, the domain gap between synthetic and real data is an issue that needs to be considered. On the other hand, we still follow the retrieve-to-locate paradigm from existing research. However, such a coarse retrieval approach naturally leads to significant localization errors, especially in our partially matching paried dataset.

As a special setting of visual place recognition, UAV-based visual localization is constrained by the scale of the data. Some existing works based on visual foundation models (e.g., DINOv2) demonstrate generalizable performance across different scenes. However, in our tests, its performance on GTA-UAV is still suboptimal. This indicates that expanding the training paradigm of general visual place recognition models (achieving a ”GPT moment”) remains an important open problem to be addressed.

Finally, single-modality visual methods are constrained by the limitations of a single-sensor system. How to integrate multimodal sensor information (e.g., text, point clouds, etc.) is also an interesting and meaningful research direction.

References

Dai, M.; Zheng, E.; Feng, Z.; Qi, L.; Zhuang, J.; and Yang, W. 2023. Vision-based UAV self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*.

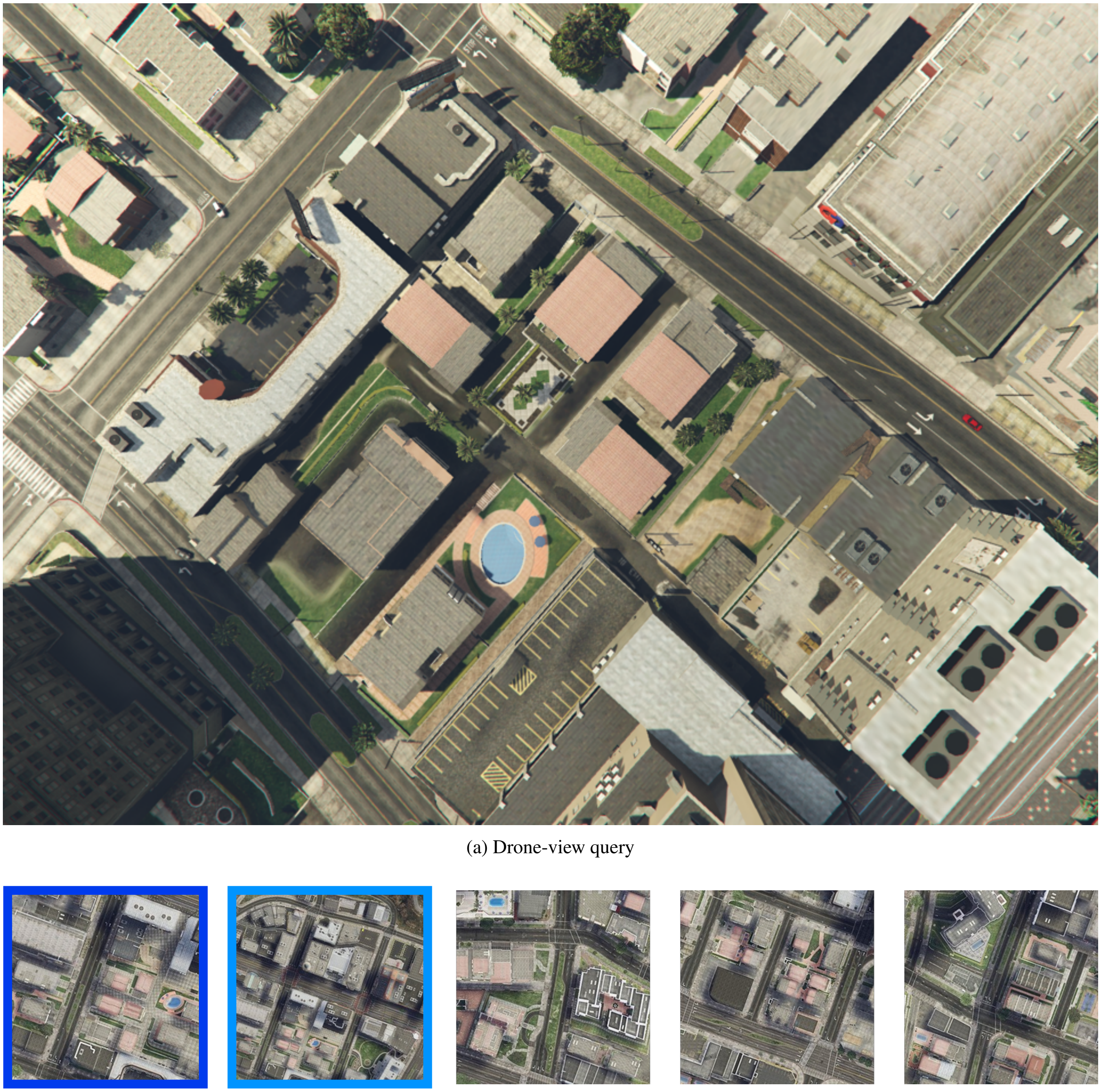
Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Kiefer, B.; Ott, D.; and Zell, A. 2022. Leveraging synthetic data in object detection on unmanned aerial vehicles. In *2022 26th international conference on pattern recognition (ICPR)*, 3564–3571. IEEE.

Xu, W.; Yao, Y.; Cao, J.; Wei, Z.; Liu, C.; Wang, J.; and Peng, M. 2024. UAV-VisLoc: A Large-scale Dataset for UAV Visual Localization. arXiv:2405.11936.

Zheng, Z.; Wei, Y.; and Yang, Y. 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, 1395–1403.

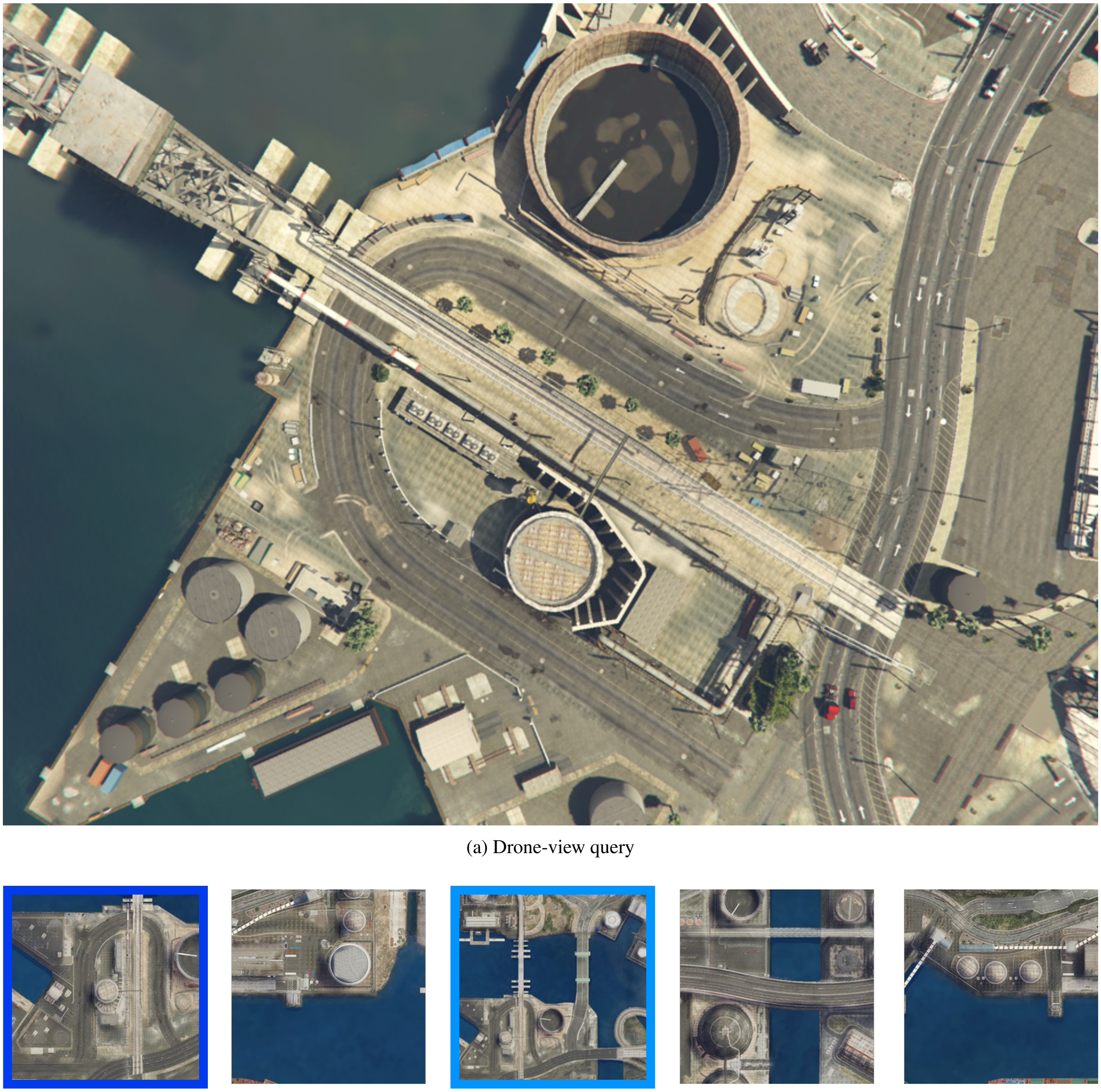
Zhu, R.; Yin, L.; Yang, M.; Wu, F.; Yang, Y.; and Hu, W. 2023. SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 4825–4839.



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

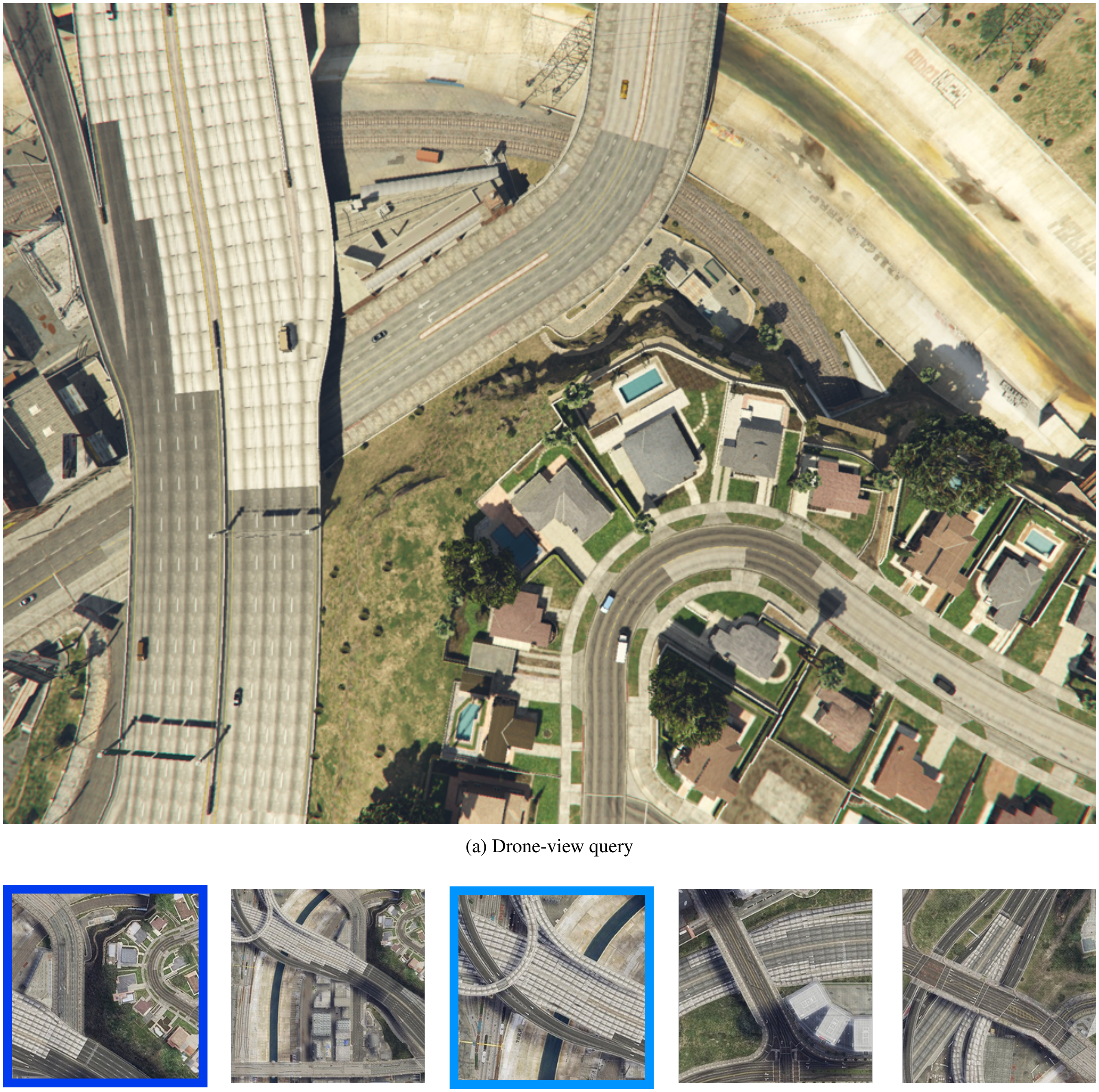
Figure 6: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

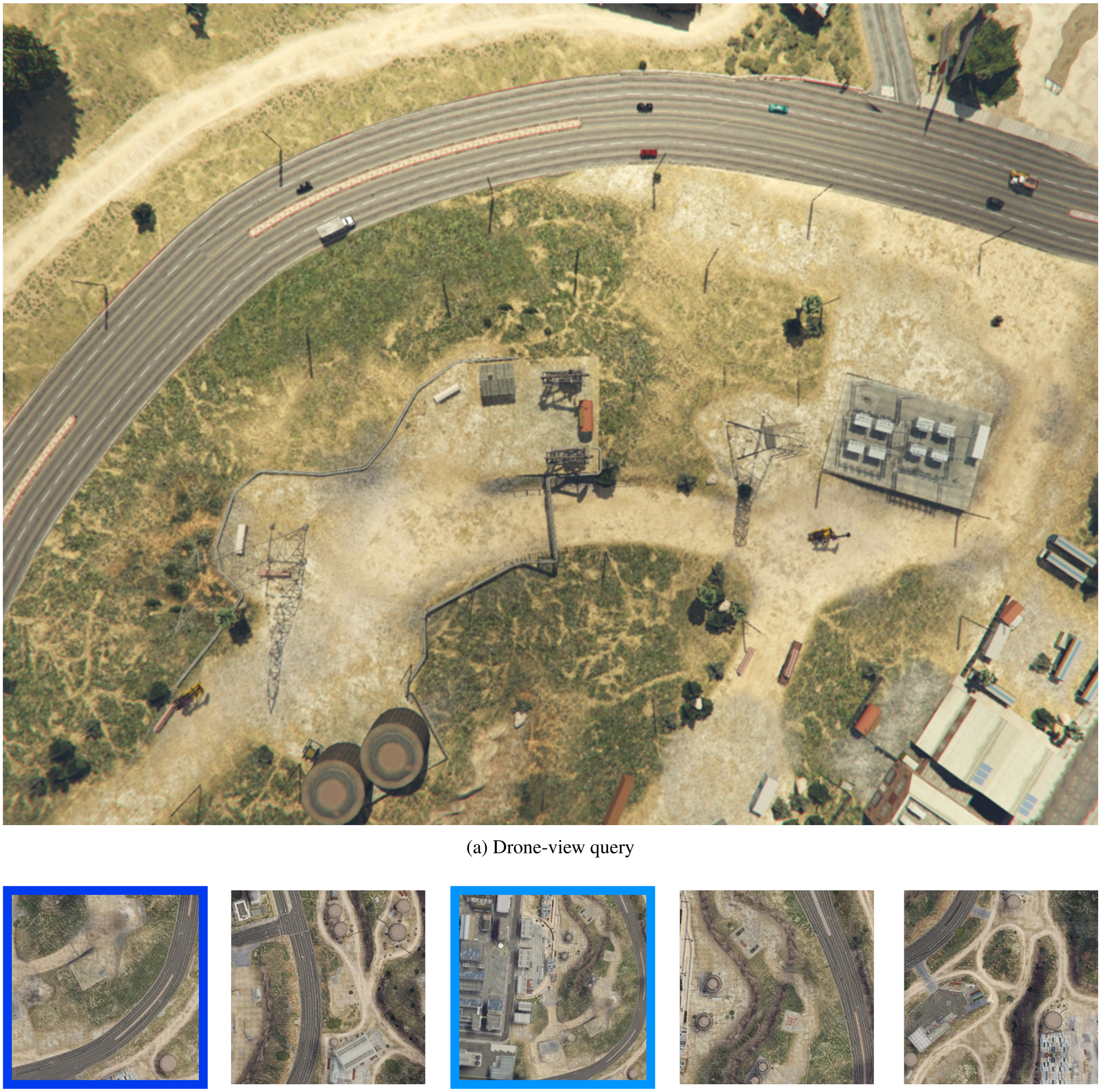
Figure 7: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

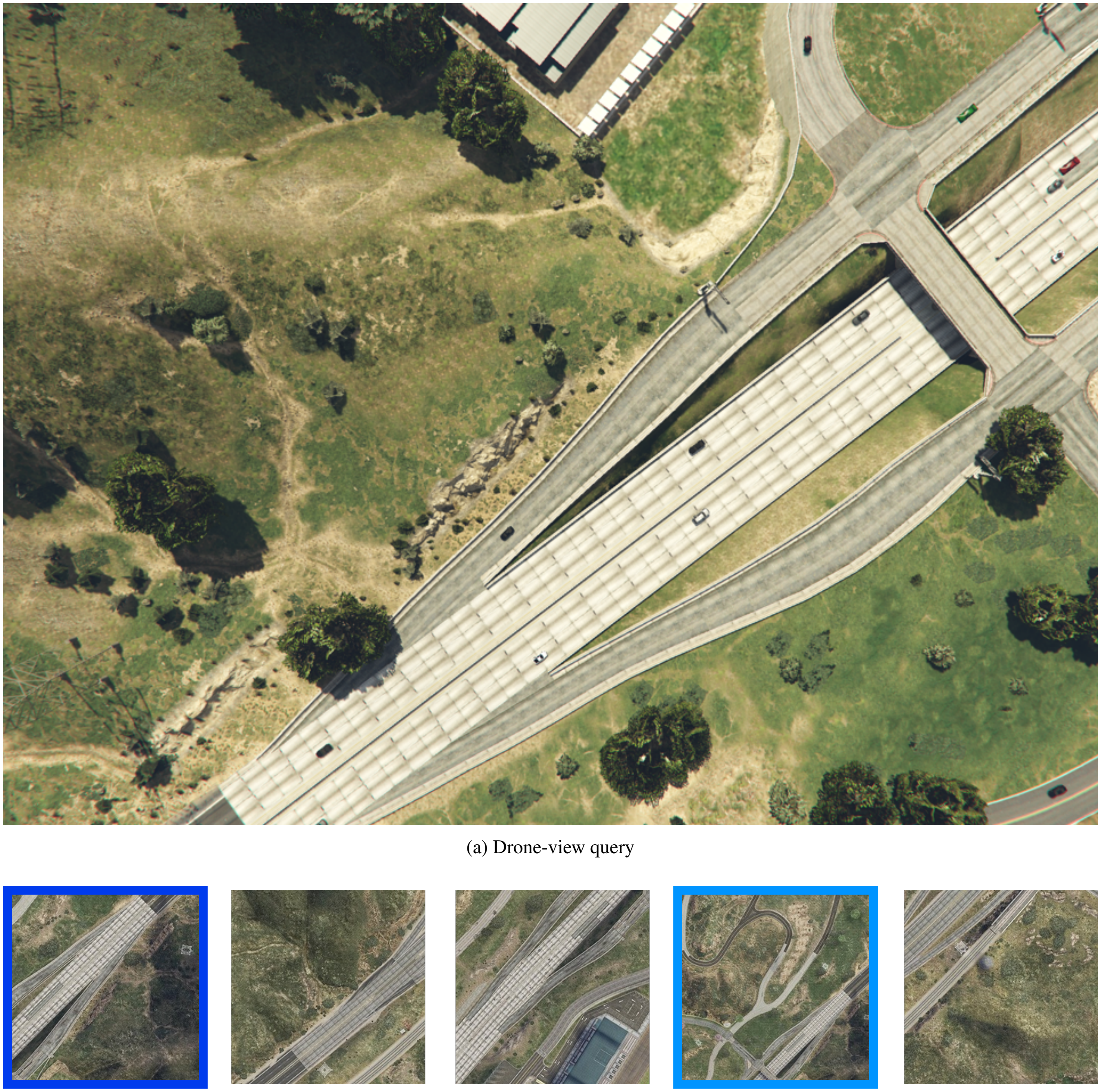
Figure 8: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

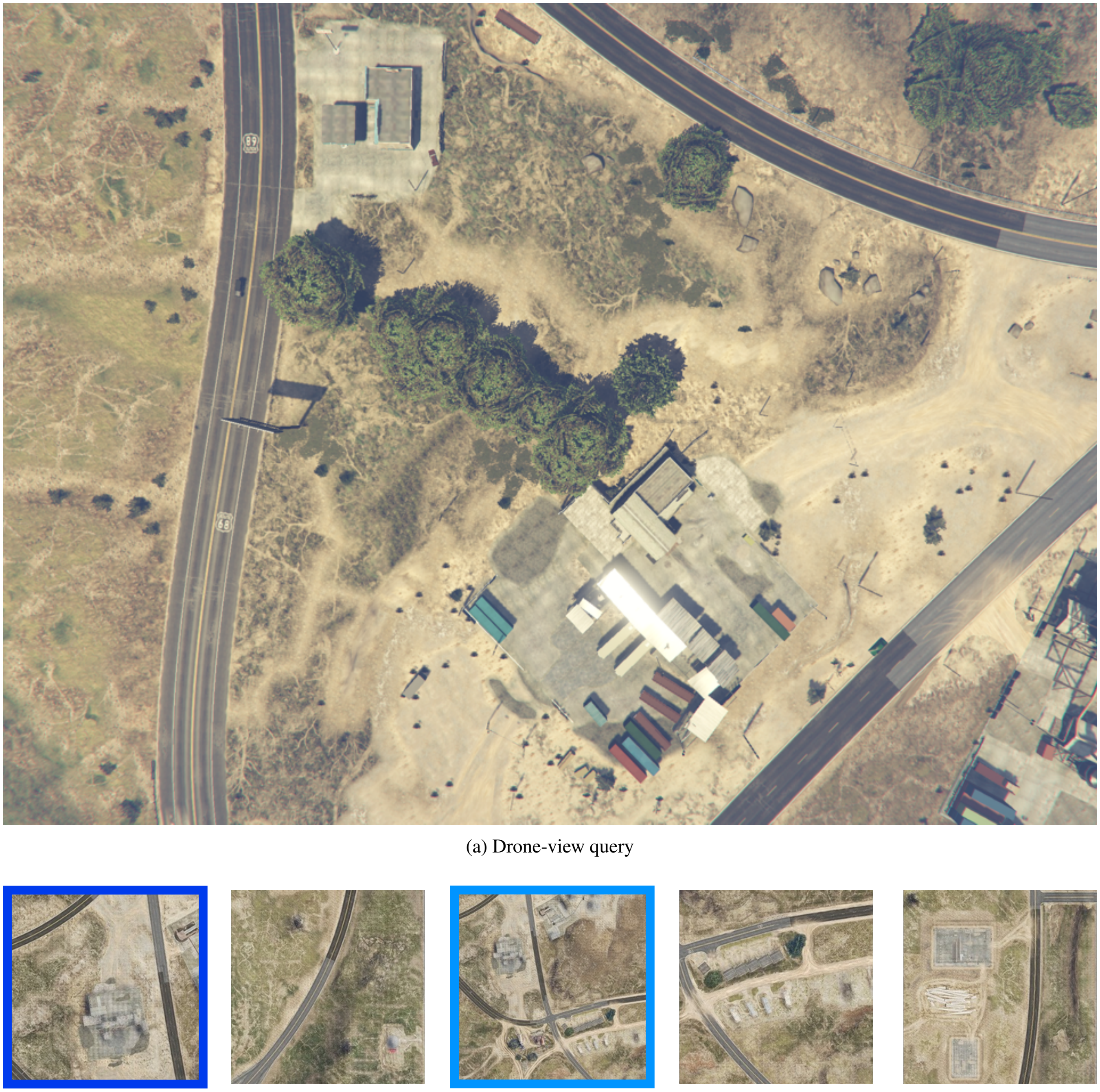
Figure 9: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

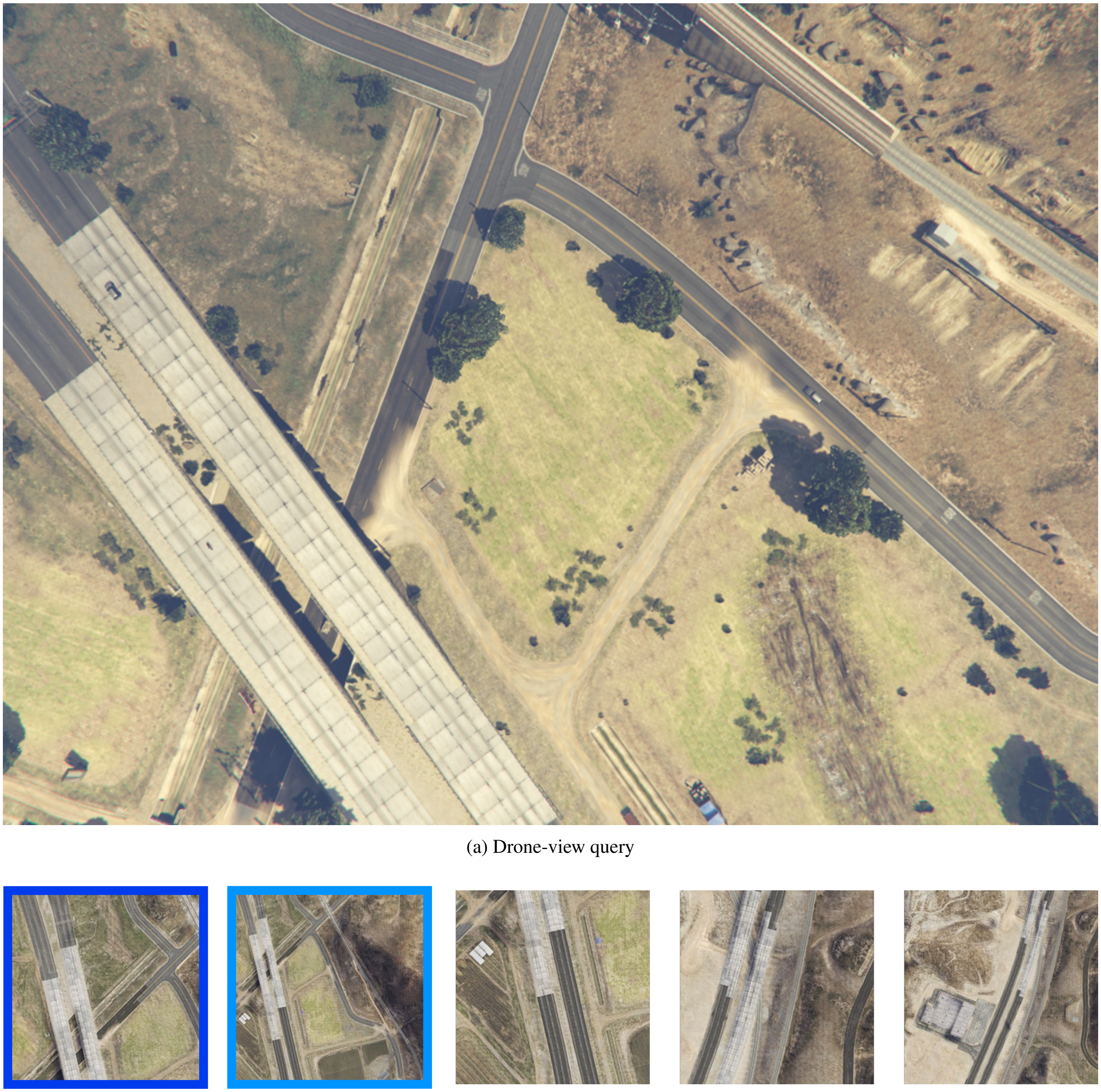
Figure 10: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

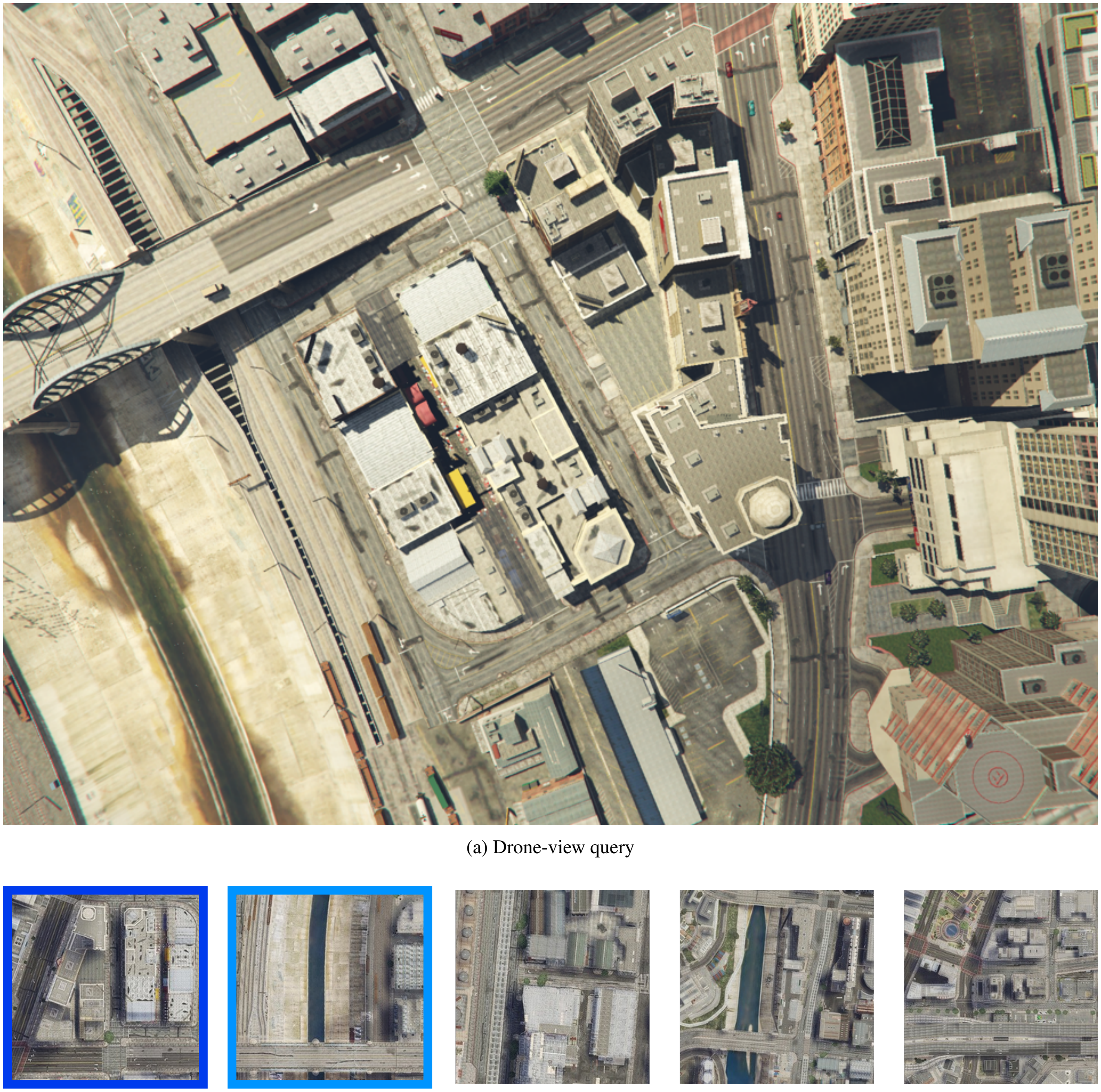
Figure 11: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

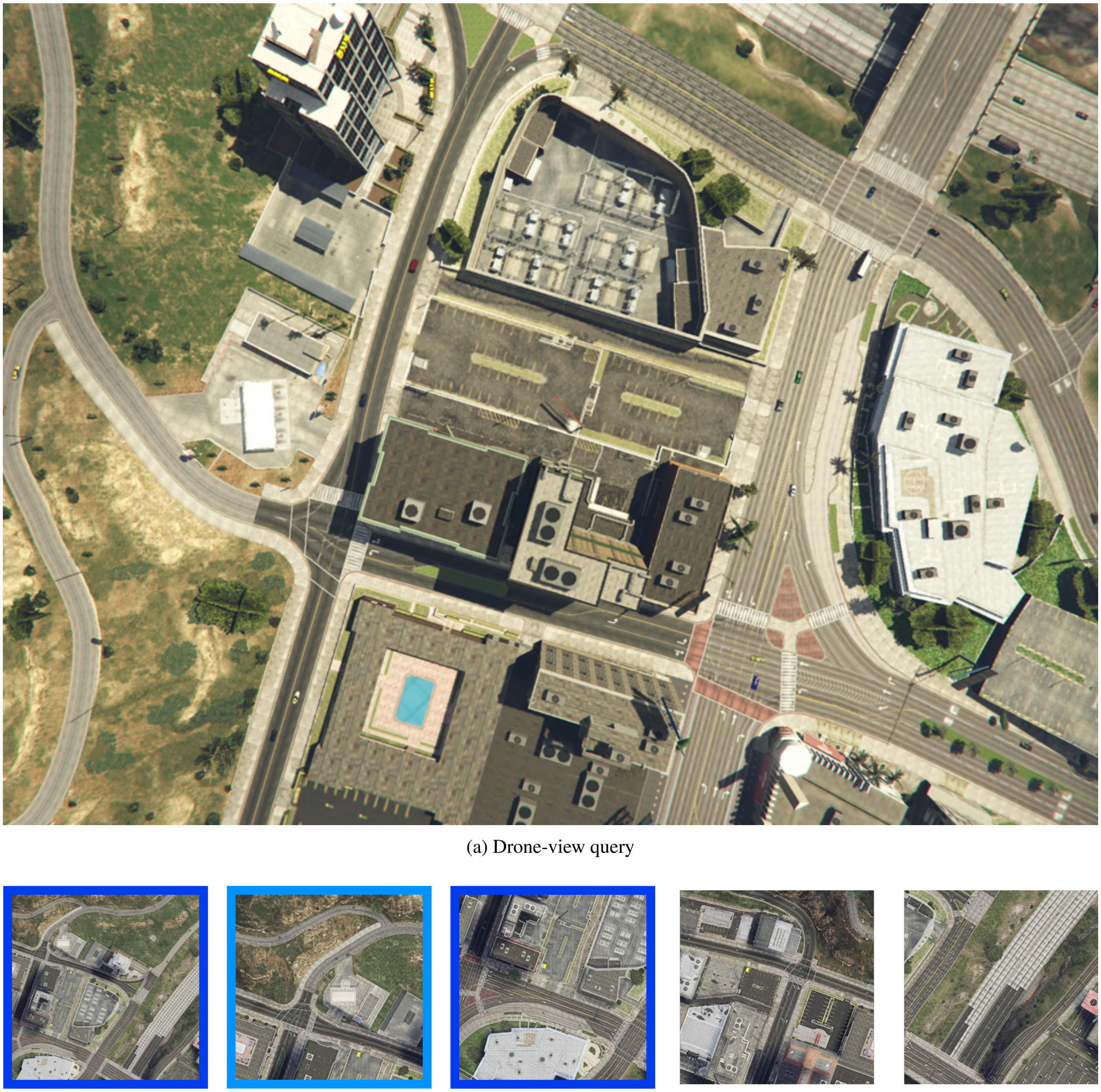
Figure 12: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

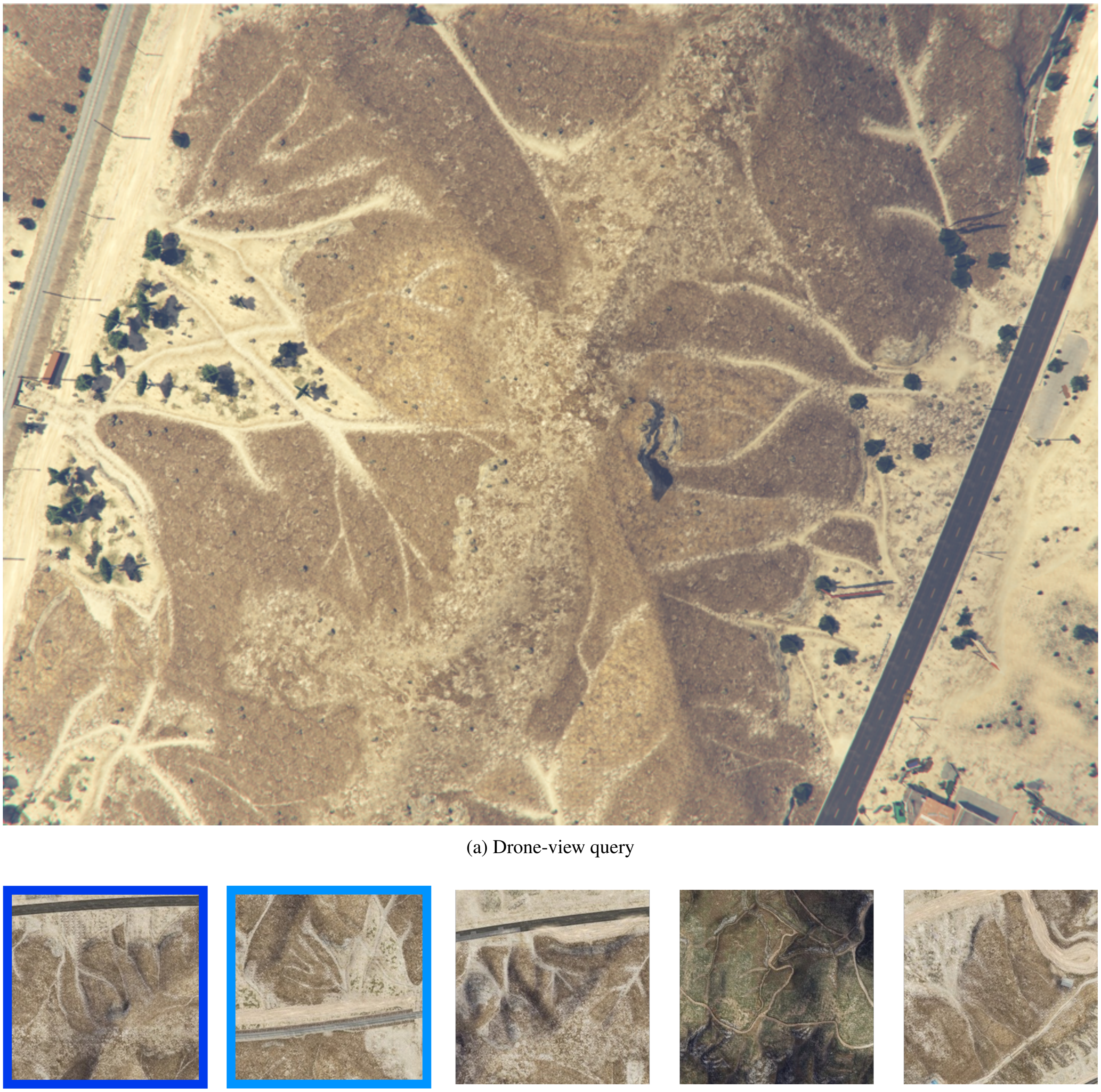
Figure 13: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

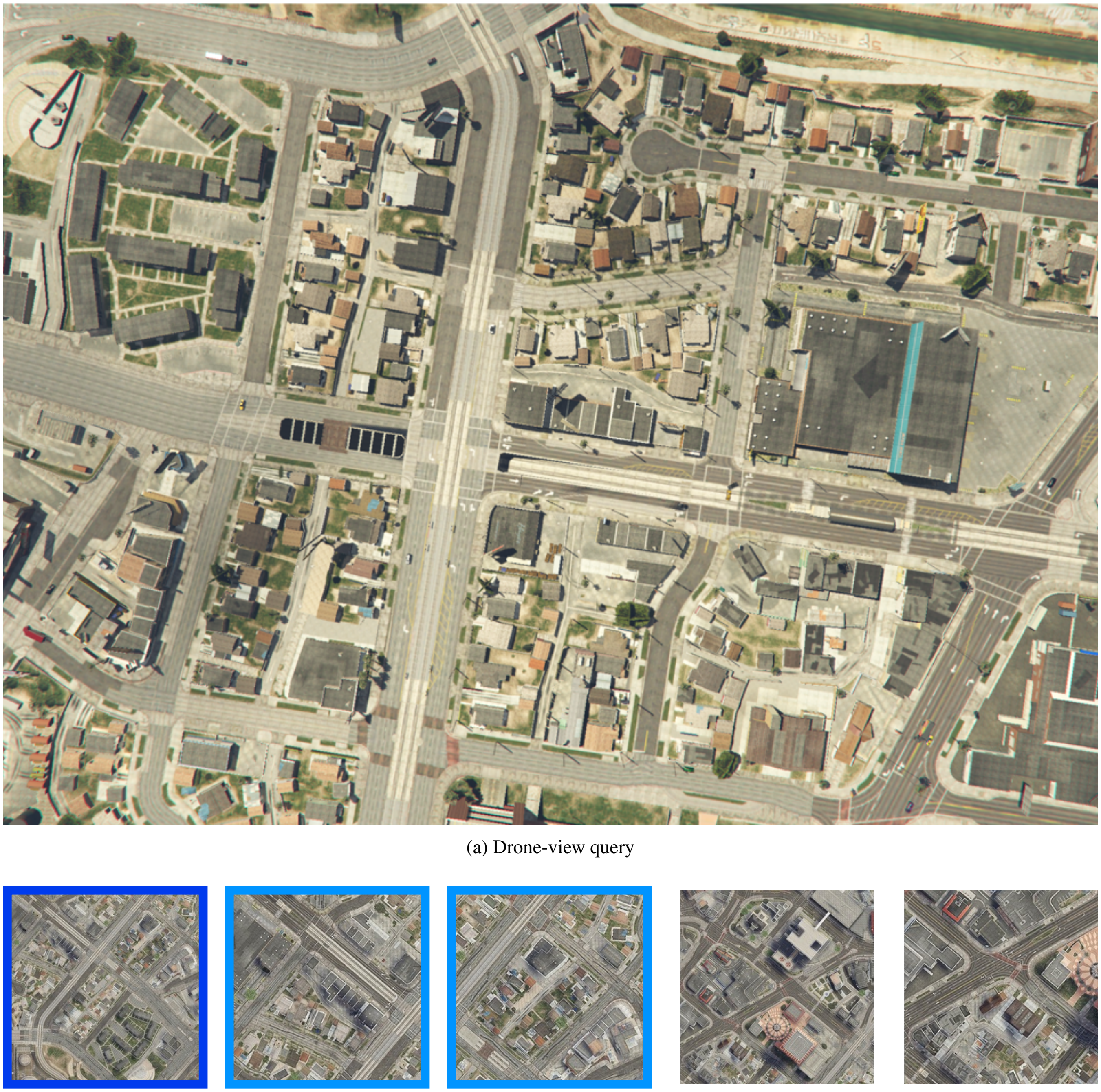
Figure 14: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

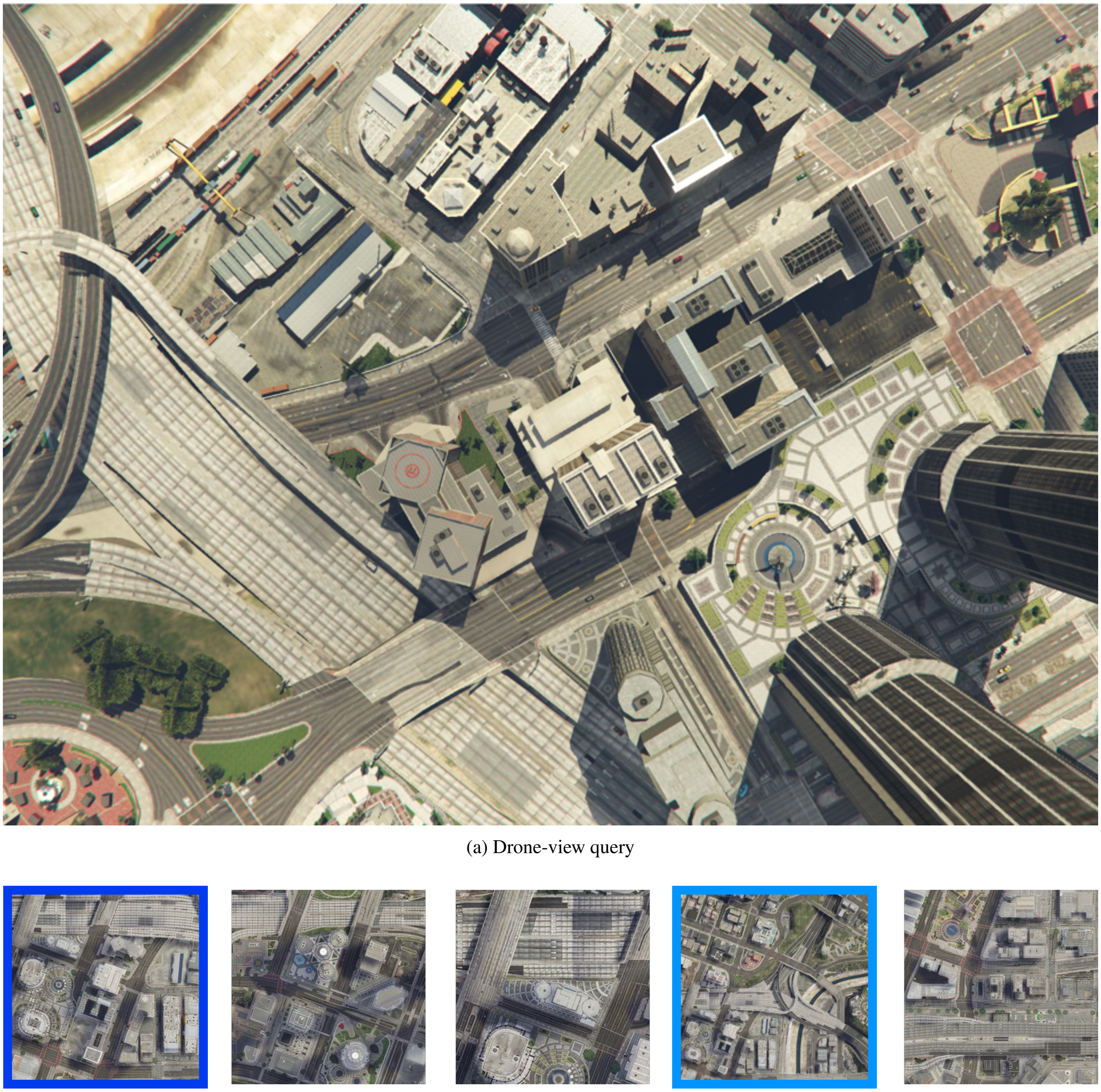
Figure 15: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

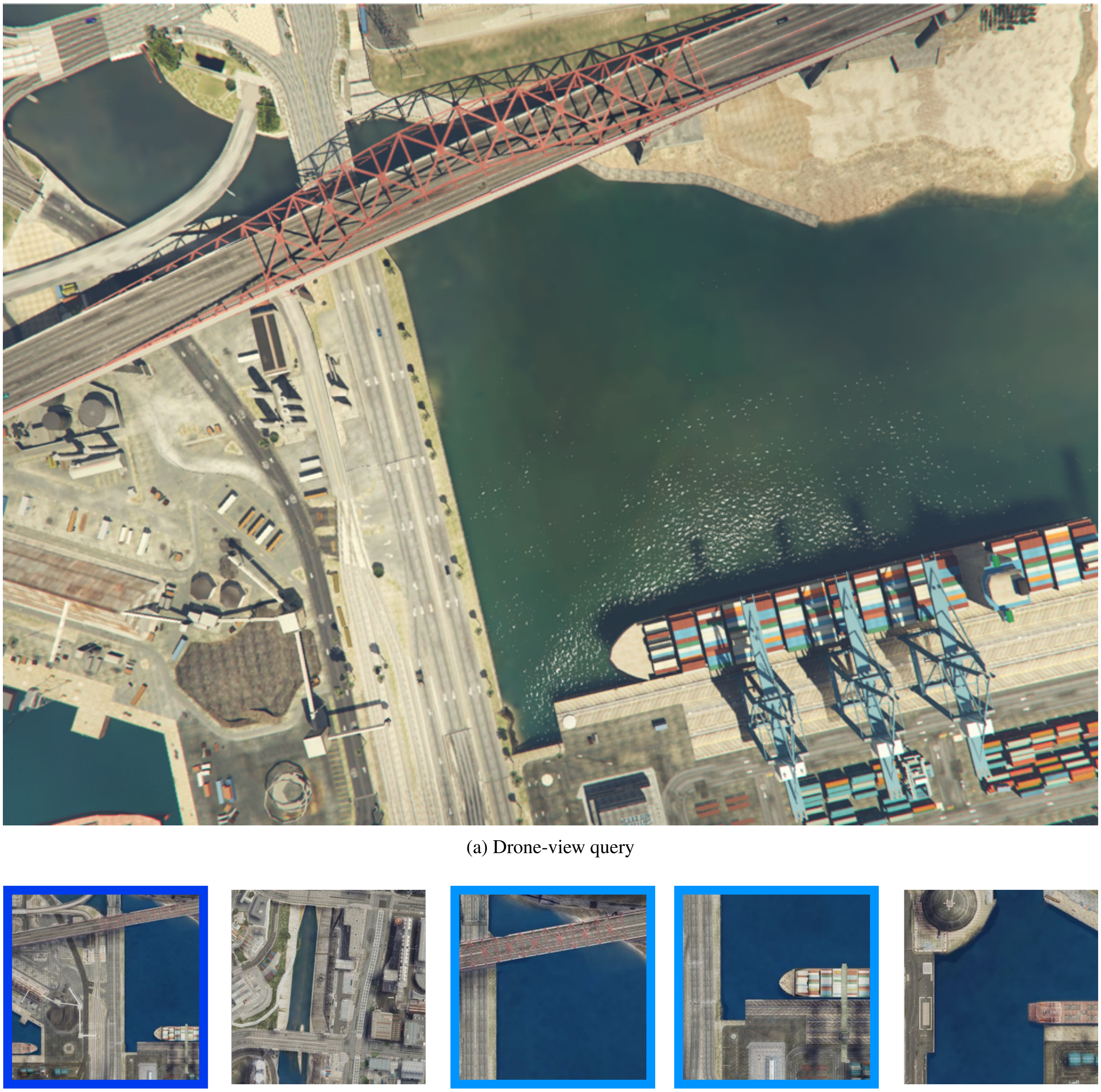
Figure 16: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

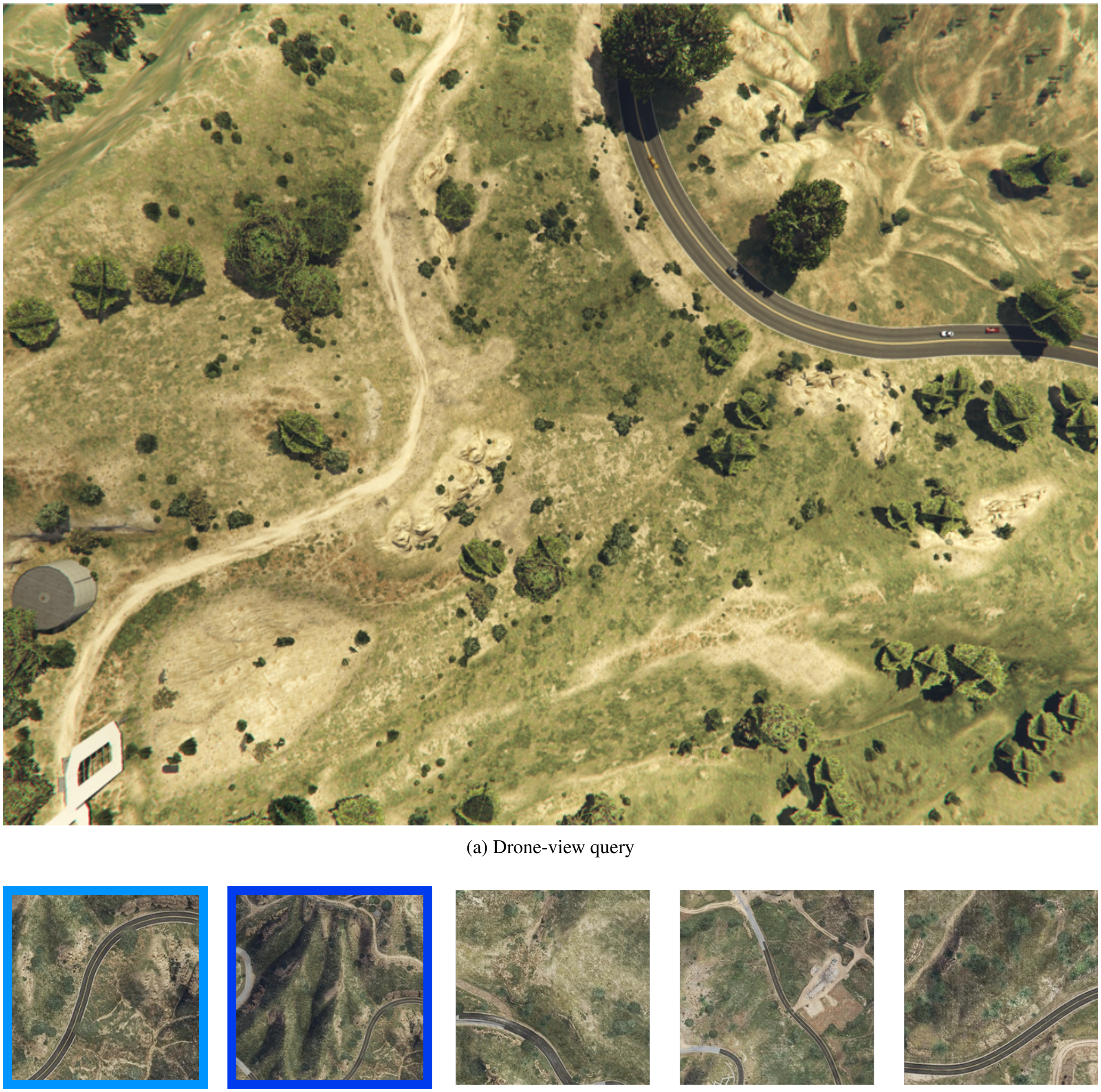
Figure 17: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

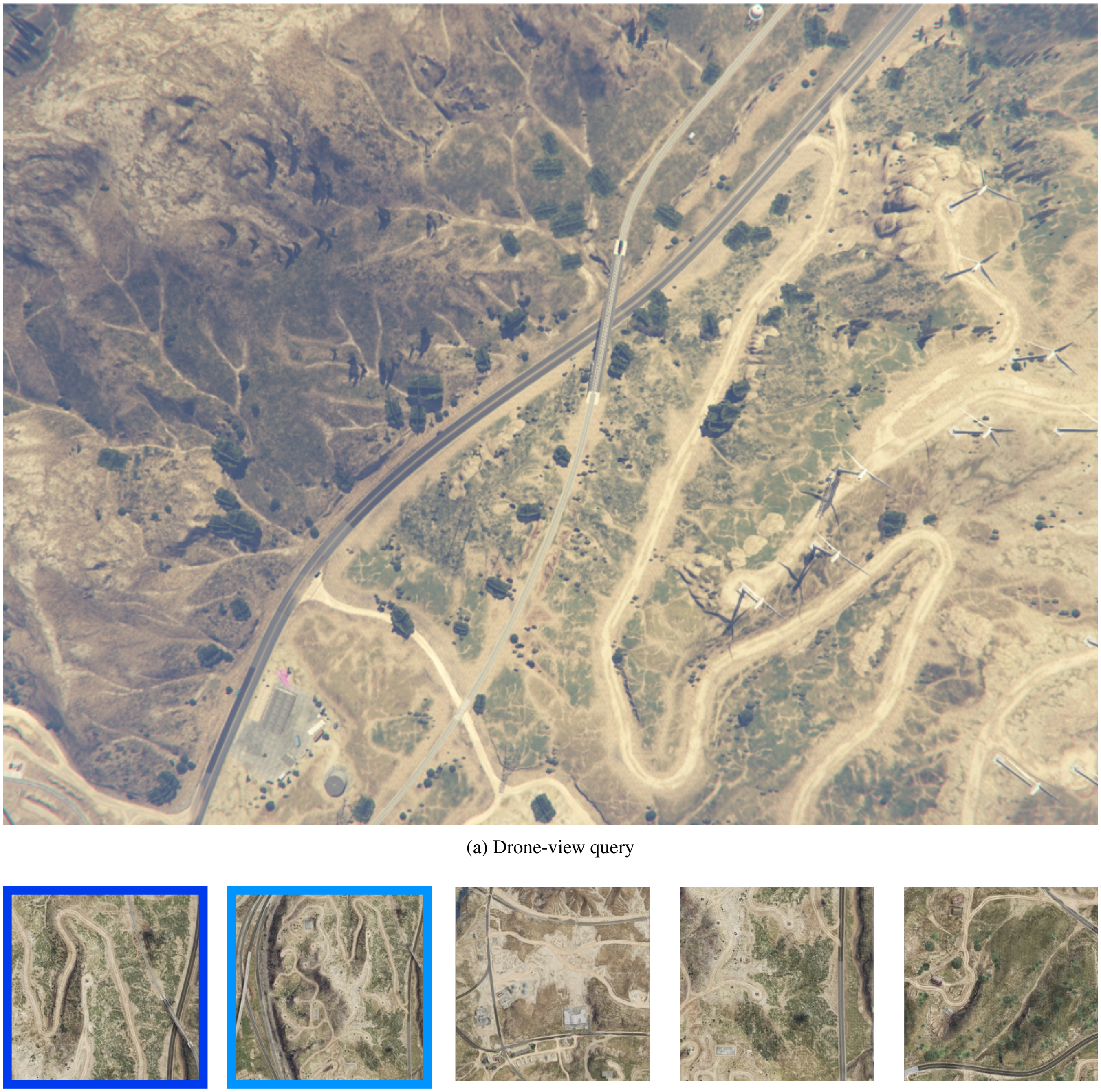
Figure 18: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

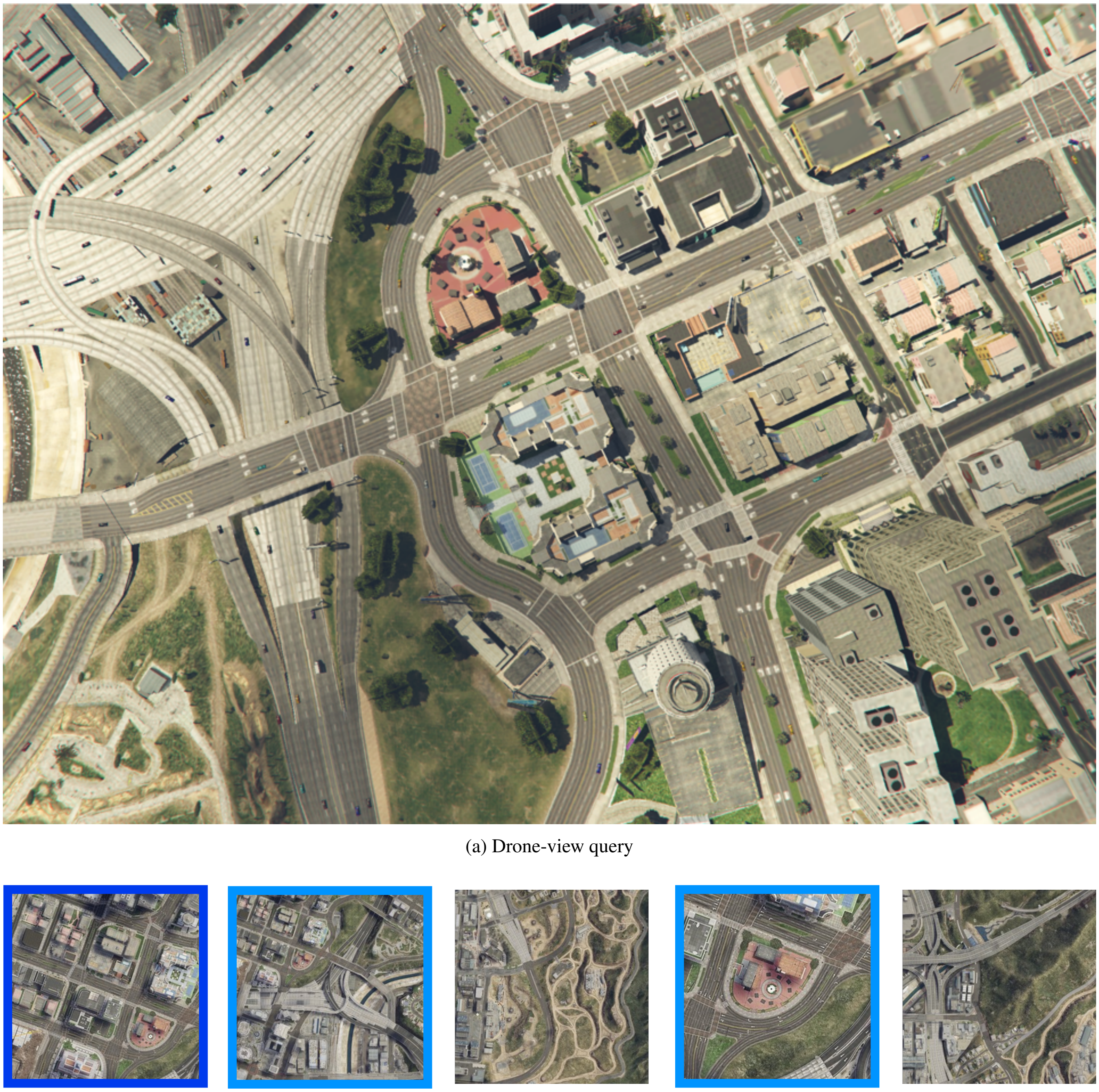
Figure 19: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

Figure 20: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)



(a) Drone-view query

(b) Top-5 satellite-view retrieval results.

Figure 21: Query example and retrieval results under cross-area training setting. (positive matched, semi-positive matched)