Beyond Matching to Tiles: Bridging Unaligned Aerial and Satellite Views for Vision-Only UAV Navigation

Supplementary Material

A.1. List of Acronyms

For clarity, the main acronyms used in this paper are grouped into four categories and summarized in Tab. 1.

Table 1. List of main acronyms used in this paper.

Type	Acronym	Description
Concept	UAV	Unmanned Aerial Vehicle
	GNSS	Global Navigation Satellite System
	CVGL	Cross-View Geo-Localization
	M2T	Match-to-Tile
Data	UVP	UAV-View Patch
	RST	Remote-Sensing Tile
	RSB	Remote-Sensing Block
	U-S	UAV-Satellite
	G-S	Ground-Satellite
Module	GLUF	Global-Local Unity Feature
	RCE	Relative Coordinate Encoder
	PSG	Patch Similarity-Guided
	CA	Cross-Attention
Metric	Recall@1	Recall at 1
	LSR	Localization Success Rate
	HSR	Heading Success Rate
	MLE	Mean Localization Error
	MHE	Mean Heading Error
	MedLE	Median Localization Error
	MedHE	Median Heading Error
	SR@20	Success Rate at 20 m
	SPL	Success Weighted by Path Length
	NE	Navigation Error

A.2. Localization and Heading Performance with Different Backbones

We evaluate Bearing-UAV with four different backbone networks, as shown in Tab. 2. Here, we define Recall@1 as the accuracy of retrieving the RST whose location is closest to the UVP from the four adjacent RSTs. Each metric is reported in two forms, "Sat." and "UAV", denoting performance under the satellite view (Sat.) and the UAV view (U-S cross-view), respectively. The former serves as a near-ideal reference benchmark, while the latter corresponds to our target cross-view localization task.

By comparing the four rows in Tab. 2, we observe that using ResNet18 or ViT-Small as the backbone leads to noticeably worse performance, MobileNet-V3-Small is slightly better than ViT-Small, whereas VGG-16 achieves a clear advantage on all metrics.

With comparable model sizes, ViT-Small and ResNet18 are better at capturing high-level global semantics, which is advantageous when features are well aligned, but less suitable under misaligned cross-view conditions where fine local cues are critical. In our localization and navigation setting, the feature maps produced by the backbone are further processed by the GLUF module to extract local features for misalignment-robust retrieval and localization. The reduced low-level structural information in these deeper, more global backbones limits their localization accuracy and thus constrains the overall performance. Moreover, ViT-Small is known to be data-hungry; given the moderate scale of our dataset and the domain gap between ImageNet-style pretraining and overhead/aerial imagery, its potential is not fully realized.

In contrast, simpler CNN backbones such as VGG-16 (and MobileNet-V3S) preserve richer fine-grained spatial details, supplying the GLUF module with stronger local structural cues that are essential for robust UAV-satellite feature matching, especially under misaligned and varying-IoU conditions. This suggests that classical convolutional backbones can still offer distinct advantages in cross-view localization-orientation scenarios.

Another noteworthy observation is that MedLE and MedHE are consistently lower than the corresponding MLE and MHE, and this gap is more pronounced in the cross-view setting, where the median localization error is 1.3 m lower than the mean and the median heading error is 5.7° lower than the mean. This phenomenon is closely related to the long-tailed distribution of localization and heading errors, especially for heading, where a small portion of unseen samples with large errors substantially inflates the mean, and the trend is as illustrated in Fig. 6.

A.3. Cross-view Multi-city Dataset Design

All data of our Bearing-UAV-90K were collected from Google Earth. Our use of Google Earth maps strictly follows its terms for non-commercial academic research, consistent with prior datasets such as DOTA (CVPR 2018) and OmniCity (CVPR 2023). Bearing-UAV-90K is publicly available at https://huggingface.co/datasets/HaoyZhou/bearinguav/tree/main.

Table 2. Comparison of localization and heading performance with four different backbones in both satellite views and UAV views (U-S cross-view). Bearing-UAV with the VGG-16 backbone achieves the best overall results. Mobile-V3S = MobileNet-V3-Small.

Method	**Recall@1**$^{\uparrow}$		**LSR@15**$^{\uparrow}$		**HSR@15**$^{\uparrow}$		**MLE**$^{\downarrow}$		**MedLE**$^{\downarrow}$		**MHE**$^{\downarrow}$		**MedHE**$^{\downarrow}$	
	Sat.	UAV	Sat.	UAV	Sat.	UAV	Sat.	UAV	Sat.	UAV	Sat.	UAV	Sat.	UAV
Ours ResNet18	83.88	75.34	88.38	71.83	73.20	46.83	8.83	12.09	7.89	10.56	14.59	26.52	8.54	16.16
Ours ViT-Small	86.38	81.39	93.35	85.47	77.58	51.57	7.53	9.46	6.73	8.02	12.02	24.89	7.69	14.45
Ours Mobile-V3S	86.41	79.76	93.52	81.20	83.25	64.94	7.49	10.34	6.73	8.72	10.57	19.53	6.28	10.14
Ours VGG-16	**90.76**	**83.17**	**98.33**	**89.36**	**98.12**	**77.21**	**5.66**	**8.61**	**5.05**	**7.30**	**4.15**	**12.90**	**2.96**	**7.20**

Figure 1. Satellite image of four cities with distinct landscapes. Below the satellite images are their selected U-S cross-view image pairs (UAV views left). City A is dominated on the right side by vegetation; City B consists mainly of densely packed low-rise buildings; City C contains many tall buildings and a wide river; City D lies in mountainous terrain with sparse, highly similar buildings. They exhibit prominent cross-view appearance gaps, illumination variations, and scene changes, covering diverse landforms such as buildings, roads, forests, mountains, and rivers.

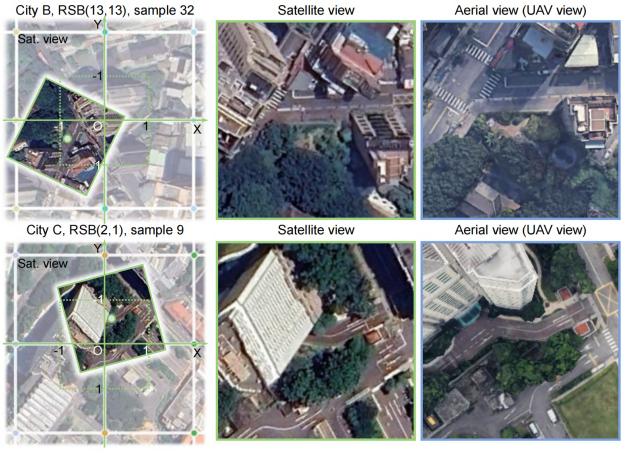

Figure 2. Two examples of UAV-satellite cross-view UVP sampling during dataset construction. In each row, the left panel shows the sampled patch in the RSB, and the middle and right panels show the corresponding satellite-view patch and UAV-view patch.

A.3.1. Dataset Construction Process

We construct our dataset from four cities, each represented by a square, contiguous satellite image, with details illustrated in Fig. 1. We first describe the data collection process, followed by the dataset structure.

Sampling Process. For each city, we first download the entire satellite image in satellite-view and then extract cross-view samples block by block from the corresponding remote-sensing blocks (RSBs). As illustrated in Fig. 2, the top row shows an example RSB in City B indexed by $(13, 13)$. The X-axes and Y-axes passing through the block center partition this RSB into four adjacent RSTs.

The process of constructing cross-view samples can be summarized in the following two steps: **(1)** In Google Earth 2D mode (satellite-view mode), within the effective sampling area (green dashed box) defined in the local XOY coordinate system of the RSB, we randomly sample relative coordinates and headings to obtain a satellite-view patch (the 32nd sample). The selected patch, shown as the dark patch with a green border centered at the green dot (the top-middle panel), serves as an ideal reference. **(2)** Using the same geodetic coordinates and heading angle, we then query Google Earth 3D mode (UAV-view mode) and extract the corresponding UAV-view patch (UVP), as illustrated by the blue-bordered image in the top-right panel.

The second row shows an analogous process for the RSB indexed by $(2, 1)$ in City C, where we obtain the four RSTs together with the 9th satellite-view patch and its paired 9th UVP. Unlike satellite-view, the UAV-view mode in Google Earth renders a photogrammetry-based 3D reconstruction of the entire city. All buildings, vegetation, and terrain are represented as textured 3D meshes, enabling oblique viewpoints that closely approximate what a real UAV camera would observe at the same position. Consequently, UVPs provide a more realistic approximation of true UAV perspectives compared with satellite-view patches.

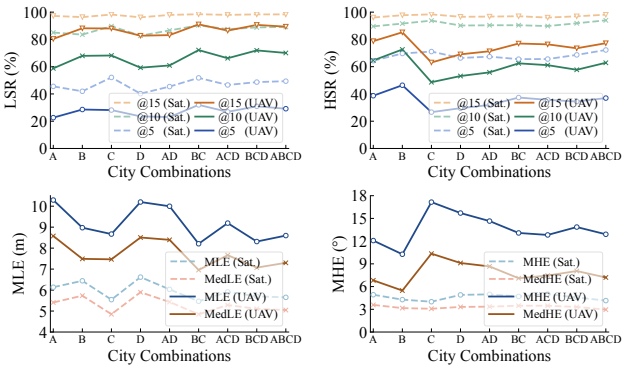

Figure 3. Localization and orientation performance under different city combinations.

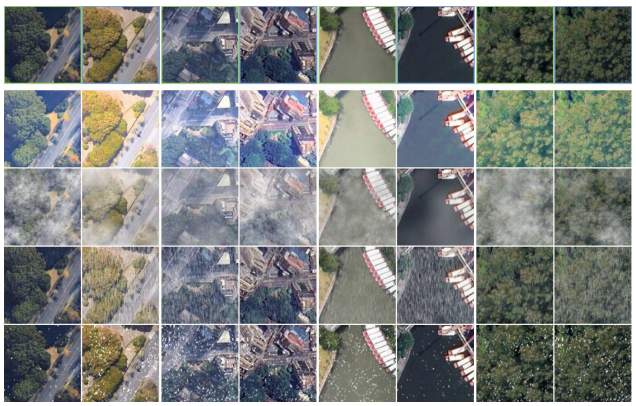

Figure 4. Four types of weather augmentations. The first row shows four original cross-view sample pairs from the four cities, where the green and blue boxes denote the UVP and its corresponding satellite-view patch. The four rows below visualize the corresponding illumination, fog, rain, and snow augmentations.

Basic Composition of the Dataset. Following this procedure, we sample 100 cross-view pairs from every RSB. Each city's image contains 15×15 overlapping RSBs whose effective sampling areas are designed to seamlessly tile the satellite image (except for an outer margin of 128 pixels that cannot be fully covered). As a result, the four city images yield a total of 90k samples per view (i.e., 90k satellite-view patches and 90k UVPs).

For each sample, the basic metadata include one UVP (paired with the corresponding satellite-view patch), its relative coordinates and heading angle (encoded as a cosine-sine vector), the four adjacent RSTs which are shared by all samples in the same RSB block, and the file paths to all associated image tiles. These fields are stored in a single `metadata.csv` file, which can be directly used for training and evaluation. As part of the dataset, `metadata.csv` provides the paths to the corresponding entity images (all UVPs and RSTs) during training and testing.

In addition, every satellite image and UVP is associated with a JSON file that records auxiliary information such as the image name, center latitude and longitude, pixel resolution, spatial extent, camera height, tilt angle, and basic conversion factors, ensuring that the dataset is well documented and easily reusable.

A.3.2. Impact of Different City Combinations

We employ more fine-grained metrics to evaluate the effect of city combinations on UAV bearing estimation. As shown in Fig. 3, the combinations include four single-city settings, two two-city pairs, two three-city mixtures, and one four-city configuration that exposes the model to the entire set.

The models trained on single-city subsets exhibit noticeable differences in localization and heading performance on their corresponding satellite images, especially in the UAV-view setting. This suggests that variations in urban morphology have a non-negligible impact on purely vision-based localization and navigation.

Specifically, City B achieves relatively strong overall performance: it features a dense, nearly grid-like street network and a large number of small buildings, providing rich fine-grained structural information and strong, consistent cues for both localization and orientation. Although City C also contains many buildings that offer abundant structural texture, the tall high-rise buildings induce larger cross-view appearance discrepancies, and the presence of a long river corridor with relatively sparse visual features is likely a major factor behind the degradation in heading accuracy. City A and City D show comparatively worse localization and heading performance, which correlates with the presence of large green areas, mountainous terrain, and many visually similar buildings.

From a global perspective, a particularly noteworthy trend is that the LSR and MLE do not degrade when the number of cities increases and the sample diversity grows; instead, they exhibit a clear upward trend. Meanwhile, the heading success rate and heading accuracy remain roughly at an average, stable level. This indicates that, as more cities are included and the training data become more diverse, Bearing-UAV is able to learn more generic orientation cues that generalize across heterogeneous city layouts, rather than overfitting to any single satellite image.

A.4. Effect of Weather Augmentation

We apply weather augmentation to 20% of the training samples for each of the four weather types and evaluate the augmented model under six test settings (four single-weather conditions, a mixed-weather condition, and a no-augmentation normal condition). Four types of weather augmentation are shown in Fig. 4, and the fine-grained metric curves are shown in Fig. 5.

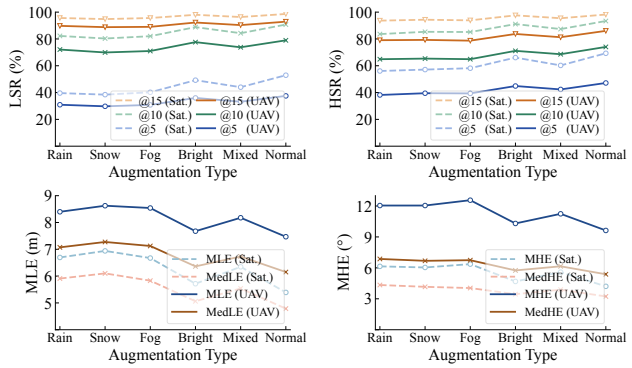

Figure 5. Effect of Weather Augmentation on Bearing-UAV. Mixed-weather training further boosts and stabilizes performance compared with the non-augmented setting.

Overall, the augmented model consistently improves both success rate and localization/heading accuracy on unseen samples. More importantly, the performance curves exhibit similar trends across different weather settings, indicating that the model learns a shared, weather-robust representation rather than overfitting to a specific appearance pattern.

Among the four types, illumination augmentation provides the most significant gain. Considering the large brightness discrepancies between UAV and satellite views, this suggests that illumination augmentation effectively mitigates cross-view appearance gaps and is particularly beneficial for cross-view localization and navigation.

A.5. Visual Analysis of Localization and Orientation Performance

To more comprehensively analyze the factors that affect the localization and orientation performance of Bearing-UAV at different levels, and to provide more intuitive support for this analysis, we perform multi-granularity visualization and statistical analysis of the evaluation metrics at three levels: city-level, RSB-level, and sample-level.

A.5.1. City-level Statistical Analysis

For the VGG-16 backbone version of Bearing-UAV, the localization error (LE) and heading error (HE) distributions and scatter plots for the 9k unseen test samples (accounting for 10% of all samples) in the satellite and UAV views are shown in Fig. 6.

The two left columns show the localization and heading errors in the satellite view, while the two right columns correspond to the UAV view. The first row presents the error distributions, and the second row shows the error scatter plots. We can first observe that, in both views, the localization and heading errors follow a long-tailed distribution that is highly concentrated around small values, with the median errors consistently lower than the corresponding means. This effect is particularly pronounced for the heading errors in the UAV-view setting. Moreover, compared with the satellite view, the UAV view (U-S cross-view) setting exhibits noticeably more outliers, which is closely related to the increased complexity of cross-view localization.

In the bottom row, the scatter plots show that the vast majority of samples stay close to the horizontal median-error lines, and large-error outliers are relatively sparse and dispersed across the sample index, without obvious structural bias or drift. Distance errors rarely exceed 20–25 m, whereas heading errors exhibit a few extreme outliers above $45°$, suggesting that heading estimation is more susceptible to rare failure cases than localization. Overall, Fig. 6 indicates that Bearing-UAV achieves stable and accurate localization and orientation on most unseen samples in both views, with only a small fraction of challenging cases contributing to the long tails of the error distributions.

A.5.2. RSB-level Statistical Analysis

During our evaluation, we observed that the accuracy of UAV localization and orientation is significantly influenced by terrain, and that this effect differs between the satellite view and the UAV view.

To analyze this phenomenon, we compute the mean localization and heading errors for each RSB and visualize them as RSB-level heatmaps, as shown in Fig. 7, which depict the spatial distribution of the model's accuracy across different terrain types. From top to bottom, the rows show the RSB-level MLE heatmaps in the satellite view and UAV view, satellite images of the four cities, and the RSB-level MHE heatmaps in the satellite view and UAV view. From left to right, the four columns correspond to City A–D. To enhance the visualization, we center the color scale of each heatmap at the mean error and set a symmetric range around this mean value.

We mainly analyze the relationship between terrain and errors under the UAV-view setting. From the RSB-level localization error heatmaps, we observe that dark-blue regions in the lower-left of City A, the right side of City B, and light-blue region in City C mostly correspond to tall building areas; these regions also exhibit relatively large heading errors, indicating that strong cross-view perspective changes increase both localization and orientation errors. In contrast, the central-upper area of City B and the central area of City A show lower localization and heading errors; these regions are dominated by low-rise buildings and small green spaces with rich fine-grained structural textures and small appearance differences, which are favorable for accurate localization and orientation. By comparison, the upper-right region of City A and the lower-left region of City D are mainly covered by forests or open fields, where sparse features make localization and orientation more difficult. Regions with highly similar building appearances can also cause degrade

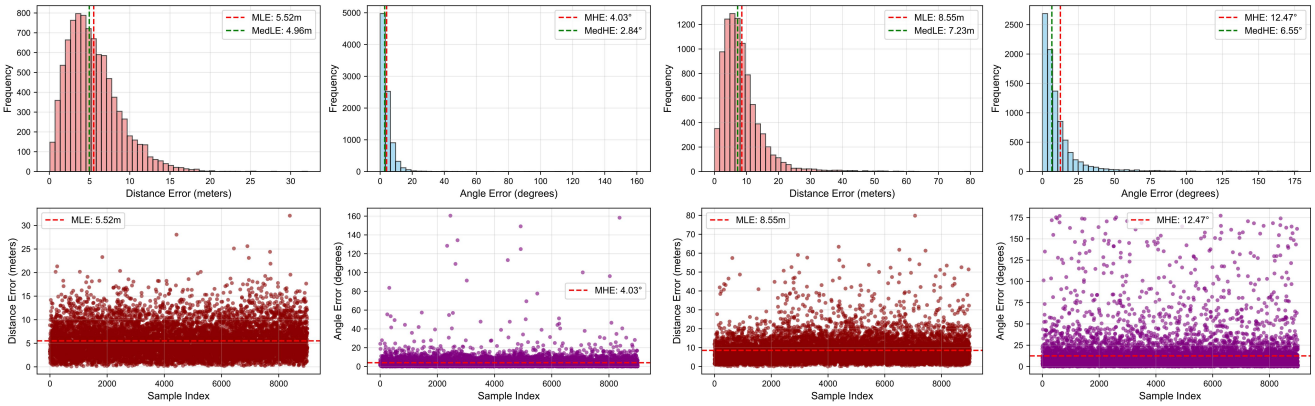

Figure 6. Localization and heading error distributions and scatter plots of 9k unseen test samples in satellite and UAV views.

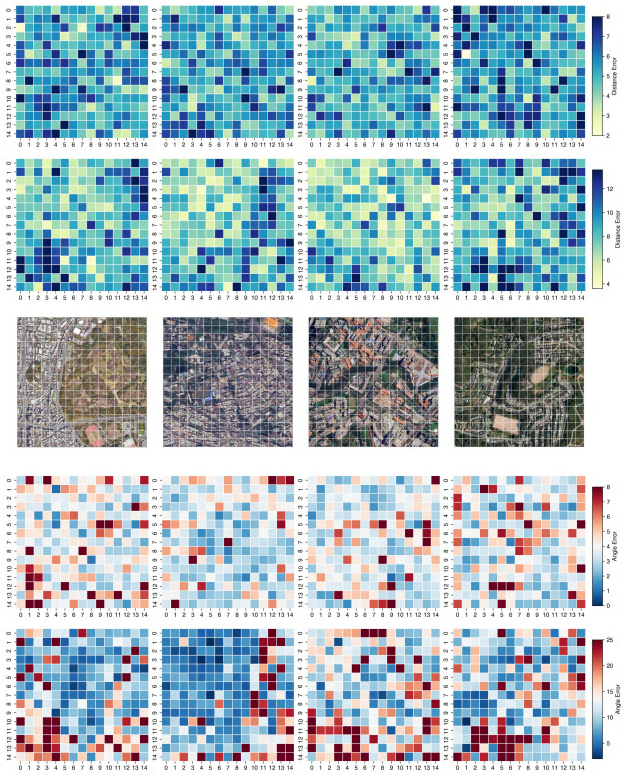

Figure 7. Heatmaps of mean localization and heading errors aggregated per RSB in two views across four cities.

performance, such as the plaza on the central-right of City C and the building cluster in the upper-right of City D.

In addition, the satellite-view heading-error heatmap of City C shows that the central river is associated with larger heading errors, consistent with the fact that rivers provide very limited texture cues. Moreover, since UVP samples are collected in different seasons and at different times than the satellite imagery, illumination changes, color shifts, and scene changes also affect localization and orientation. Comparing the two views, the spatial distributions of large errors vary, reflecting cross-view appearance gaps, especially perspective distortion around tall buildings, illumination discrepancies, and changes in surface objects.

Overall, these observations confirm that large cross-view visual discrepancies, regions with sparse structural textures, and areas with highly similar appearances all have a significant influence on the performance of the proposed algorithm.

A.5.3. Sample-level Visual Analysis

To support the analysis from a finer-grained perspective, we perform map-based sample-level visualization of localization and heading errors under U-S cross-view, revealing detailed error patterns and enabling intuitive visualization of localization errors (LE) and heading vectors (HE). For each of the four cities, we visualize LE and HE across 2250 unseen samples overlaid on corresponding satellite images, as shown in Fig. 8. This geographic visualization helps clearly illustrate how errors vary across different urban regions. To facilitate reading and cross-referencing with the figure, we include the relevant details in the caption for clarity.

A.6. More Navigation Experiments

We further comprehensively compare Bearing-UAV (red trajectories) with four baselines on seven additional navigation routes across four cities, as shown in Fig. 9. The eight routes (City A–D, Route #1 and #2) have lengths of 771, 1119, 644, 757, 524, 821, 721, and 1115 m, respectively, with 10–13 waypoints each and diverse scene types. While not covering the full satellite image, these routes span a wide range of scene types and scales (*e.g.*, large/small buildings, similar building clusters, roads, rivers, open fields, forests, and playgrounds), thus reflecting the model's navigation performance under different scene layouts.

Figure 8. Visualization of U-S cross-view localization and heading errors on real satellite images from four cities. The four satellite images in the **top-left, top-right, bottom-left, and bottom-right panels** correspond to **City A–D**, respectively, and overlay the localization and heading results of all unseen test samples under cross-view setting. In each image, *blue/red dots* denote the ground-truth/predicted positions, while *blue/red arrows* denote the ground-truth/predicted heading vectors. The *green line segment* connecting the endpoints of the two arrows represents the position error, whose magnitude is annotated by *green numerical values*, and the *purple numeric values* denote the heading error; smaller values indicate better accuracy, and perfect overlap between the red and blue arrows corresponds to the ideal case. This visualization enables precise, fine-grained inspection of localization/heading errors under different scene types across all cities. The satellite images are divided into fine-grained RSTs, such that the surrounding terrain of any sample can be examined and analyzed in detail against the real-world surroundings. For example, in the high-rise building areas (*e.g.*, the lower-right of City A and the lower region of City C), in feature-sparse regions (*e.g.*, forests in the upper-right of City A, green fields and the central river in City C, and open grounds in City D), as well as in dense clusters of similar buildings (*e.g.*, City B and City D), many samples exhibit noticeably large localization and/or heading errors. These errors are likely due to the adverse effects of cross-view appearance discrepancies, feature sparsity, and feature similarity.

A.6.1. Analysis of Multiple Flight Trajectories

From these visualizations, Bearing-UAV shows the strongest adaptability to multi-city scenarios and complex navigation routes, consistently achieving stable performance across diverse scenes, such as dense urban blocks, irregular river boundaries, and mixed residential–commercial areas. It successfully completes four routes and, on the remaining ones, still follows most segments before failure. These failures typically occur at later stages and are mainly caused by severe appearance ambiguities or rapid structural changes, rather than early-stage drifts. This suggests that Bearing-UAV maintains stronger long-horizon consistency under cross-view misalignment.

In contrast, baseline methods struggle in structurally complex scenes or along routes with tight curves and multi-directional turns. GTA-UAV (green) trajectories are often longer but frequently diverge from the intended path and continue advancing in an incorrect direction, indicating unstable behavior. University-1652 (orange), DenseUAV (yellow), and SUES-200 (blue) exhibit more severe issues, often drifting shortly after the starting point. In several cases, they fall into local loops, circling within a small region instead of progressing along the intended route.

A.6.2. Analysis of Flight Frame Sequences

To provide an intuitive illustration of the cross-view UAV navigation process, we visualize step-by-step trajectories for two representative routes, including the sequence of visited satellite-view RSBs and UVP frames, enabling analysis of how the system progressively adjusts heading, updates relative position, and aligns with the target route under varying scene structures and cross-view discrepancies.

Corresponding to trajectory # 1 of City D in Fig. 6 of the main paper, Fig. 10 shows a successful navigation case, in which Bearing-UAV completes the route in 45 steps while maintaining stable alignment with the designated trajectory. Frames 10–14 and 30–35 show smooth transitions. Noticeable heading changes occur at frames 8–9, 14–15, 24–25, 35–36, and 44–45, which are near waypoints and correspond to normal turns, demonstrating that the model can effectively handle turning maneuvers. In contrast, consecutive large heading changes appear at frames 28–30 and 38–40. At these positions, the field of view contains repeated patterns (*e.g.*, rows of similar buildings and vegetation), indicating that feature similarity or repetition can introduce stochastic disturbances to heading estimation.

Fig. 11 shows a more challenging case from trajectory #2 of City A, where navigation eventually fails, but the UAV still manages to follow a substantial portion of the planned path before drifting away. Together, these visualizations offer insights into the model's decision-making behavior, the challenges caused by viewpoint inconsistencies, and the contrast between successful and near-successful trajectories.

Table 3. Ablation study under satellite-view settings.

GLUF	RCE	PSG	CA	R@1↑	LSR↑	HSR↑	MLE↓	MHE↓
✗	✓	✓	✓	67.33	58.56	83.67	14.55	10.41
✓	✗	✓	✓	90.55	98.10	95.11	5.74	5.51
✓	✓	✗	✓	90.65	**98.40**	96.96	5.70	4.72
✓	✓	✓	✗	89.88	98.28	96.30	5.68	5.12
✓	✓	✓	✓	**90.76**	98.33	**98.12**	**5.65**	**4.15**

A.7. More Ablation Experiments

We also take ablation study under the satellite-view settings. As shown in Tab. 3, the satellite-view results follow the same trend as in the UAV view: adding the GLUF module yields a clear performance gain in both views. Moreover, the RCE and PSG modules further enhance localization and heading accuracy, with particularly strong improvements in heading estimation. The RCE, PSG, and CA exhibit slightly better orientation than localization performance. This consistent observation across views validates the effectiveness of each component.

A.8. Limitations

Vision-only UAV localization remains an open problem and is still far from mature. Vision-only performance is a critical complement to sensor-vision fusion under GNSS denial or long-term inertial drift, and also underpins geometric reasoning methods (*e.g.*, PnP) that estimate multi-DoF UAV poses. However, M2T/retrieval paradigm is limited by grid density and typically ignore heading estimation, restricting long-range localization and autonomous navigation. Moreover, repeated encoding of satellite tiles makes M2T-based methods time-consuming even when with onboard satellite imagery. In contrast, our paradigm avoids this cost and achieves state-of-the-art vision-only localization while jointly estimating heading, enabling more complete conditions for long-distance autonomous navigation.

Nevertheless, although our method focuses on challenging scenarios such as misalignment, sparse features, and varying IoUs, real-world flight environments involve many additional temporal and spatial variations. Thus, our approach still has several limitations. First, its generalization to unseen cities remains underexplored. The model is trained and evaluated on satellite images from four cities, where supervised learning allows effective regression of position and heading. However, its cross-city transfer ability requires further validation and improvement. Second, the current framework does not explicitly address dynamic scene changes, such as moving vehicles, seasonal vegetation variation, or disaster-induced appearance changes. Handling such dynamic factors will be the focus of future work.

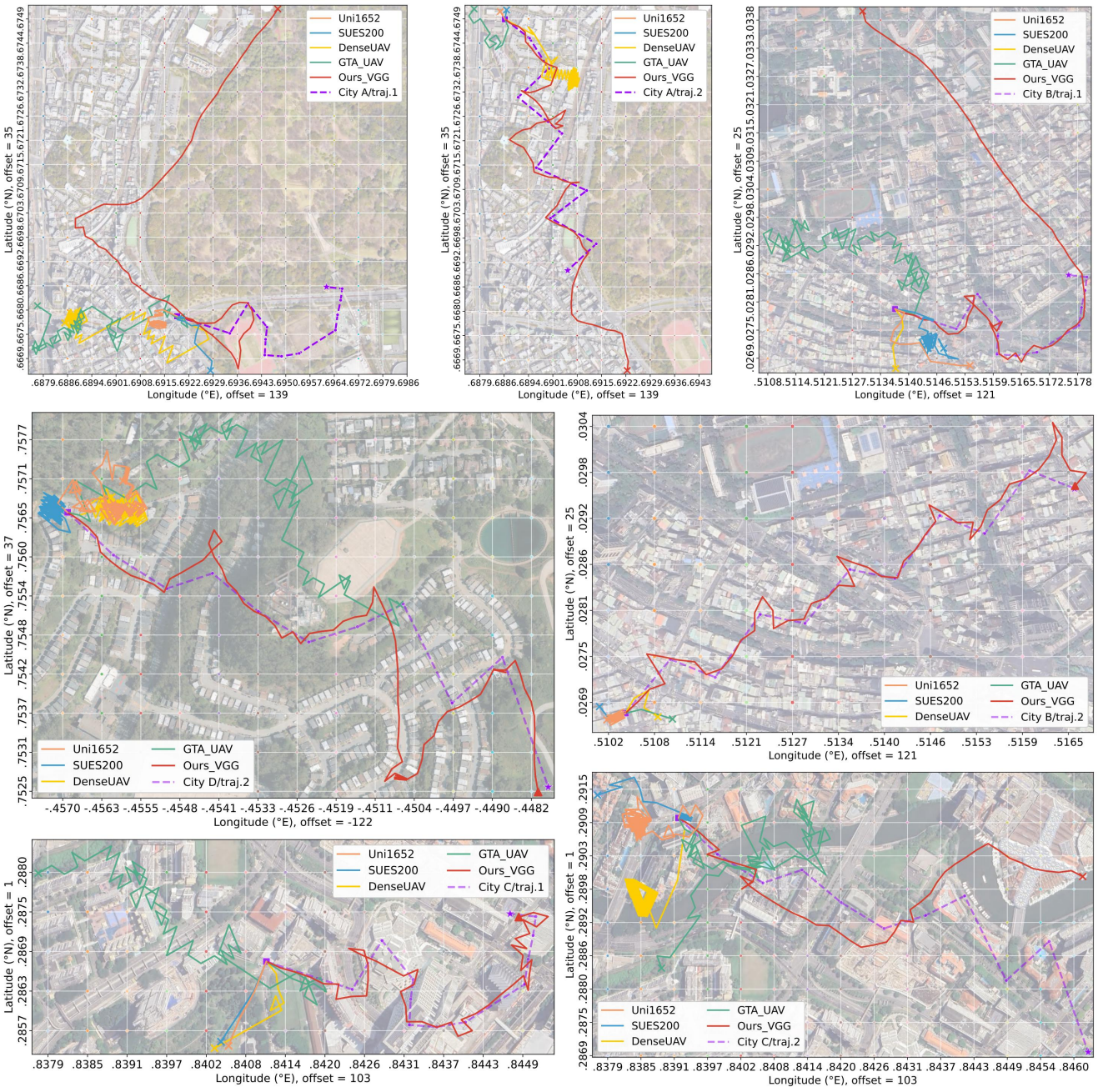

Figure 9. Navigation performance comparison on seven additional routes. We further provide a comprehensive comparison between Bearing-UAV (red trajectories) and four baseline methods on seven additional navigation routes across the four cities. The UAV step size is 25 m, and the waypoint arrival threshold is 20 m. All visual elements follow those in Fig. 6 of the main paper. The eight routes across Cities A–D range in length from 524 m to 1119 m. Each route contains 10–13 waypoints and covers diverse scene types, including buildings of varying scales, green areas, rivers, and playgrounds. Overall, Bearing-UAV exhibits the strongest adaptability to different urban layouts and highly winding paths. It successfully completes four of the eight routes and, for the remaining ones, still follows most of the designated path before failure. In contrast, baseline methods struggle in these complex scenes and along curved routes. The green GTA-UAV trajectories are often longer but show significant deviation from the reference path. The orange Uni-1652, yellow DenseUAV, and blue SUES-200 trajectories tend to drift away near the starting point, with some falling into local loops and circling within a small region. Notably, large positional deviations and flight drift mainly occur in vegetation areas (traj. #1 City A, traj. #2 City A, traj. #2 City D), similar building areas (traj. #1 City B, traj. #2 City D), and tall building areas (traj. #2 City C). These regions often suffer from feature sparsity, high feature similarity, and significant cross-view appearance discrepancies, all of which degrade localization and heading estimation performance.

Figure 10. Visualization of cross-view navigation frames of Bearing-UAV. We present a successful case for trajectory #1 in City D from Fig. 6 of the main paper. Frames are arranged top-to-bottom and left-to-right. In each step, the left panel shows the current RSB, where the blue rectangle indicates the UAV field of view and the airplane icon indicates the heading. The top-right inset shows the UVP, and the bottom-right inset reports outputs including heading angle, relative coordinates, direction vector, geographic coordinates, and the RSB index.

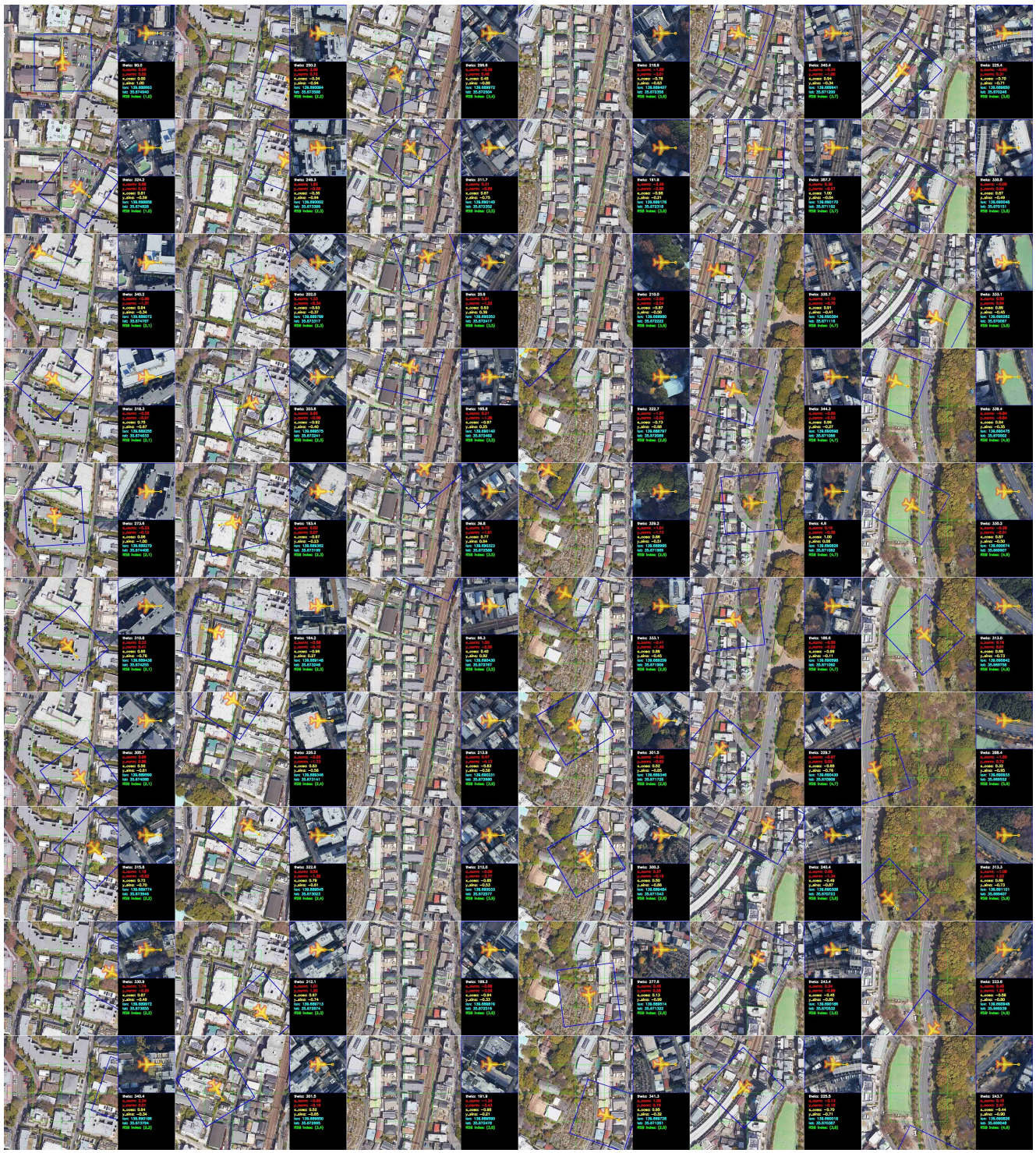

Figure 11. The UAV cross-view navigation process corresponding to trajectory #2 in City A, as shown in Fig. 9 (first row, middle). This is a more challenging case in which navigation eventually fails. However, the UAV still follows a substantial portion of the planned path before drifting away. We visualize overlapping portion (first 60 steps) to highlight cross-view discrepancies. Significant spatial differences between UVP and satellite patches arise from 3D parallax and illumination variations at low altitudes. Deviations occur at frames 26–34, likely due to waypoint turns and repetitive building patterns. Despite these challenges, Bearing-UAV remains robust along winding trajectories.