Video2BEV: Transforming Drone Videos to BEVs for Video-based Geo-localization

Supplementary Material

Outline. This supplementary material includes 4 aspects:

1. Visualization:
   - more visualizations of the Video2BEV transformation:
     - comparisons of the Video2BEV transformation at different elevation angles on the UniV;
     - visualizations of drone-view videos, BEVs, and satellite-view images on the UniV.
   - more visualizations of the UniV dataset;
   - more visualizations of synthetic negative samples on the UniV dataset.
2. Out-of-Distribution scalability test in rainy weather.
3. Failure case analysis.
4. Implementation details.
5. Inference efficiency.
6. Additional ablation study:
   - ablation study for loss weights;
   - ablation study for FPS;
   - visualizations of different trajectory lengths.

1. Visualization

1.1. Visualizations of the Video2BEV Transformation

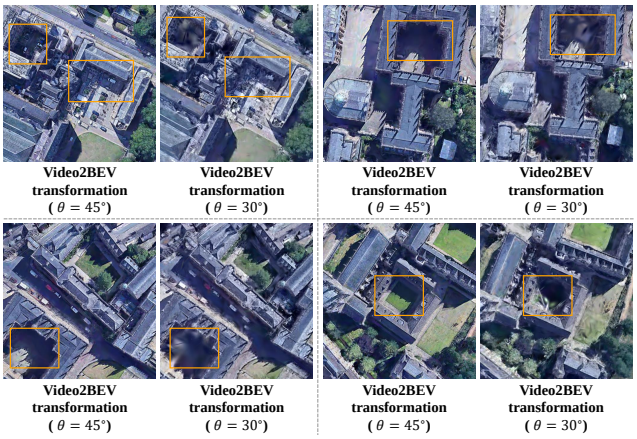


Figure 8. The transformed BEV comparison of videos with different evaluation θ on the UniV. We highlight the challenging regions.

Visualizations of the Video2BEV transformation at different elevation angles. Compared to the $45°$ subset, the $30°$ subset of the UniV dataset presents more occluded cases. We analyze the impact of occlusions and other environmental constraints on the proposed Video2BEV transformation. As shown in Fig. 8, the proposed Video2BEV transformation produces satisfactory BEVs at a $45°$ elevation angle, especially in areas between tall buildings. At the $30°$ elevation angle, some regions reconstructed by the Video2BEV transformation exhibit imperfections. These imperfect regions are primarily located between buildings, where it is challenging for drones to capture clear images at a relatively low elevation angle. Despite the imperfectly reconstructed regions, the proposed Video2BEV transformation significantly narrows the disparity between the drone view and the satellite view.

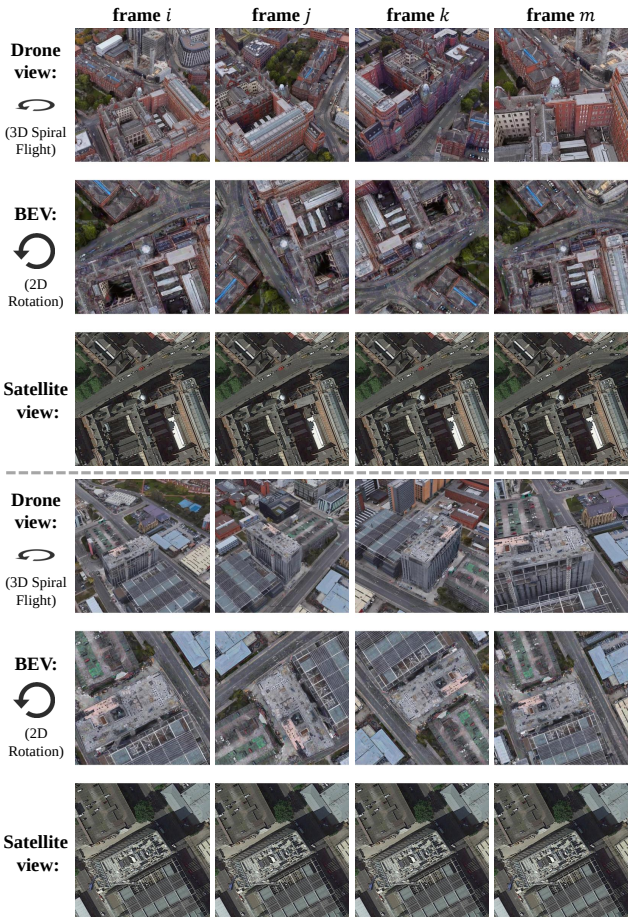


Figure 9. Visualizations of drone-view videos, BEV videos, and satellite-view images on the UniV dataset. i, j, k, m are frame indices, and $i < j < k < m$.

Visualizations of Drone-view Videos, BEVs, and Satellite-view Images. We provide visualizations of images from different platforms. For each building, both drone-view and Bird's Eye View (BEV) data are in video format and satellite-view data is in image format. Specifi-

Figure 10. Elevation angles θ illustration on the UniV dataset. For each case, top row shows $\theta = 45°$ and bottom row displays $\theta = 30°$. With a lower elevation angle, the new flight captures the target building with *wider Field of View (FoV)* but more *occlusions*, thereby posing more challenges for drone geo-localization.

cally, drones follow a spiral path around the target building, completing three circular flights. For BEVs, in the training set, we incorporate rotation angles and varying heights into BEV camera poses, generating a sequence of rotating and scaled-down BEVs (see Fig. 9). The 2D rotation is designed for data augmentation. In the test set, we only increase the height of the BEV camera poses and render a sequence of scaled-down BEV images. The satellite view contains one image for each location. After the proposed Video2BEV transformation, the BEVs align with the same viewing direction as the satellite view and exhibit a similar color pattern to the drone view.

1.2. Visualizations of the UniV Dataset

We provide additional visualizations of 45° and 30° elevation angles in the UniV dataset (see Fig. 10). Although both videos capture the same building, they differ significantly between the two elevation angles. Videos captured at a 45° elevation angle provide overall views of the core areas of the target building, with these areas visible in most cases. In contrast, at the relatively lower elevation angle of 30°, drone-view videos offer a wider field of view but also introduce more occlusions. Consequently, core areas of the target building are occluded in some frames while remaining visible in others (see Fig. 10), effectively simulating outputs from real-world drone flights. More visualizations can be found in the ***UniV-dataset.mp4***.

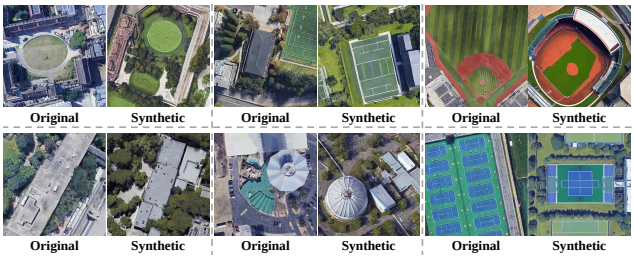


Figure 11. Visualizations of original images and synthetic hard negative samples on the UniV dataset.

1.3. Visualizations of Synthetic Negative Samples

We provide additional visualizations of original and synthetic images (see Fig. 11). Synthetic negative samples exhibit similarities to the original samples in terms of the architectural features and color patterns of the buildings. For cases in the first row, the synthetic samples have architectural features resembling the original images, such as the circular lawn, the green sports field, and the oval stadium. In the second row, the synthetic samples exhibit similar color patterns to those of the original images. Despite the similarities, the architectural details differ between the original and synthetic samples, making the synthetic samples suitable for serving as negative samples.

2. Out-of-Distribution (OOD) scalability test in rainy weather

Here, we propose a new research direction, namely, video-based geo-localization under rainy weather. Specif-

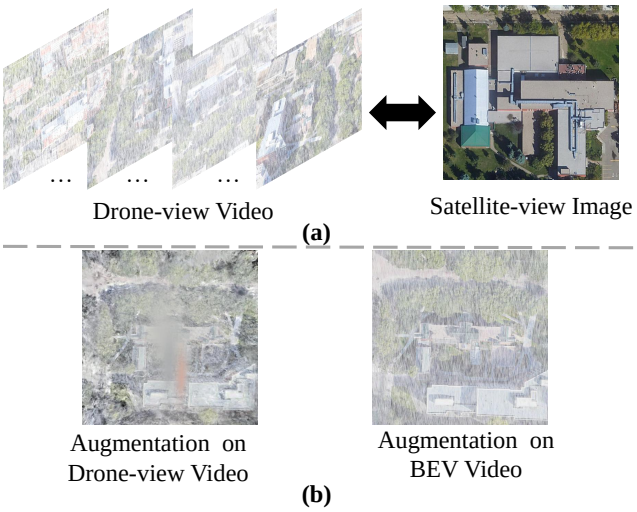


Figure 12. Visualizations of **(a) video-based geo-localization under rainy weather setting** and **(b) comparisons of rainy data augmentations**.

ically, this setting aims to match drone-view videos (in rainy weather) with geo-tagged satellite images (in clean weather). See Fig. 12a for more details. We perform rainy data augmentation and out-of-distribution (OOD) testing on the 45° & 2 fps test set (see Tab. 4). Compared to Tab. 2a, all methods, including ours, show suboptimal performance, indicating that this setting is a challenging one. We explore two types of combinations of rainy data augmentation and the Video2BEV transformation in our method. The first type applies rainy data augmentation to the clean drone-view videos (Ours† in Tab. 4), meaning the augmentation is applied **before** the Video2BEV transformation. The second type applies rainy data augmentation to the clean BEV videos (Ours* in Tab. 4), meaning the augmentation is applied **after** the Video2BEV transformation. The performance of these two combinations varies, which we attribute to the sensitivity of the vanilla version of 3DGS to the input in the Video2BEV transformation. Predicting accurate camera poses and point clouds from rainy videos remains challenging, leading to the imperfect visual quality of BEVs from rainy inputs (see Fig. 12b). This finally hinders accurate matching. However, we are confident that future advances in efficient 3DGS may help solve this problem. We leave this as a future research direction.

3. Failure Case Analysis

We provide additional qualitative visualizations of retrieval results, with a particular focus on the failure cases on the UniV (Fig. 13a) and SUES-200 (Fig. 13b) datasets. In these cases, the proposed method fails to recall the matched image in top-1. We observe that it is challenging because the recalled top-1 image has a very similar pattern to the query

Table 4. Comparisons in terms of an out-of-distribution testing on the rainy weather setting (45° test set). † and ∗ denote rainy data augmentation on drone-view videos and BEV-view videos, respectively.

Method	D→S		S→D	
	R@1	AP	R@1	AP
LPN	2.28	3.58	24.82	29.64
FSRA	16.83	21.18	44.08	49.42
DWDR	33.67	38.91	55.63	60.85
Sample4Geo	60.48	64.62	72.32	75.71
Ours†	30.10	33.75	36.80	40.98
Ours*	65.05	67.49	83.02	84.83

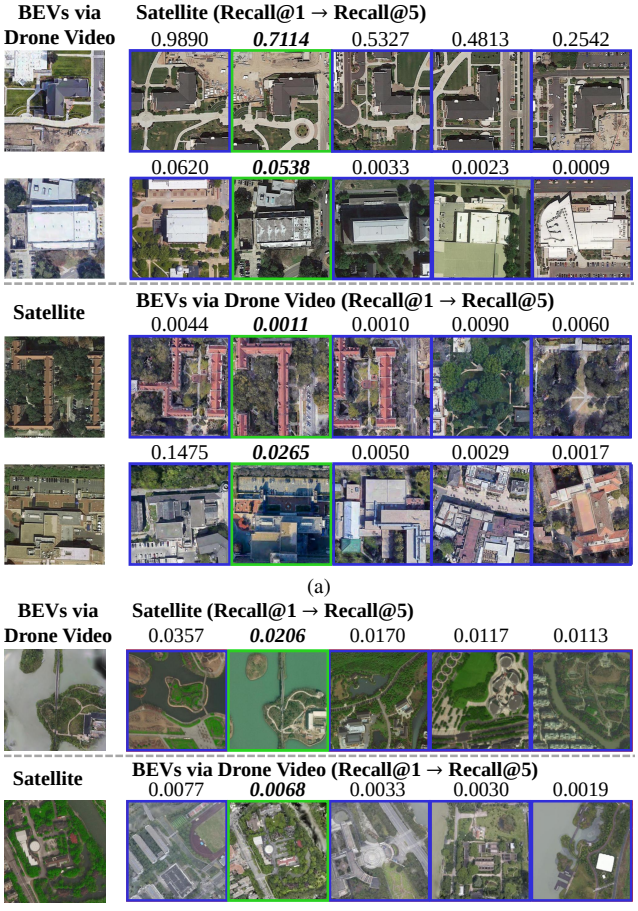


Figure 13. Typical failure cases for Drone → Satellite and Satellite → Drone on the UniV (a) and SUES-200 (b) datasets. We observe that the failures are mainly due to two factors. First, some buildings were under construction, which is quite different from the current view. Second, some satellite-view photo color is not accurate, and some similar buildings are false-matched. Given queries (left) from different platforms, matched galleries are in green box, and mismatched galleries are in blue box. Scores on the top are similarity scores estimated from our method.

Table 5. Inference efficiency analysis.

Method	LPN	FSRA	DWDR	Sample4Geo	Ours
Backbone	ResNet50	ViT-S	Swin-B	ConvNeXt-B	ViT-S
# Params./M	211.00	197.66	332.90	334.03	215.26
GFLOPs	16.35	24.60	30.38	90.54	28.38
Time/ms	6.76	3.77	14.23	10.88	6.26

image, particularly in terms of the appearance and structure of the geographic target in the two images. In the first case, all recalled images share a similar structure, and the predicted scores are relatively high. In the second case, all recalled images have a white rectangular roof. The roof of the ground truth image turns partly gray, which affects the retrieval prediction of our method. In the third case, recalled top-3 results have similar red roofs, making it challenging for the proposed method to accurately retrieve the ground truth building. In the fourth case, the recalled top-1 image has a similar architectural style to the query, and the ground truth image is in shadow. Both factors contribute to an inaccurate retrieval result. In the last two cases, satellite-view photo color is not consistent with that of drone-view, resulting in false positive results.

4. Implementation details

Data processing. (**i**) Video2BEV transformation. We estimate camera poses from drone-view videos using 8 NVIDIA GeForce RTX 4090 GPUs. The subsequent 3DGS training and BEV rendering are also conducted using the same computing resources. It takes less than 1 second to render 50 images with 512 $\times$ 512 with vanilla 3DGS. (**ii**) Hard Negative Sample Synthesis. It consists of caption generation, fine-tuning of the Stable Diffusion network, and hard sample synthesis, carried out on a single RTX 4090 GPU. All data will be released upon acceptance.

Model training. We adopt a two-stage training strategy. In the first stage, the encoder is based on the ViT architecture and supervised with instance loss and contrastive loss. In the second stage, the encoder is frozen and MLPs are optimized with the matching loss. All these experiments are conducted on 1 NVIDIA A800.

5. Inference efficiency

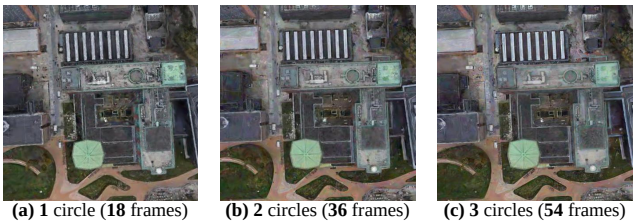
We provide details of two modules and an overall inference efficiency. (1) Hard negative sample synthesis is only used during training and does not impact the inference. (2) Our optimized 3DGS [1] renders Video2BEV transformation for 0.15s with 1 Nvidia 4090. It can be executed on a remote server in advance, thus not affecting on-device latency. Future advances in efficient 3DGS may facilitate our work to on-device rendering. The rendering efficiency is out-of-the-scope of our work, and we leave it as the future work. (3) The overall on-device ranking time is in Tab. 5. Ours takes slightly more time per sample than FSRA, but is more efficient than the rest of the competitive methods.

Table 6. Ablation study for loss weights λ.

λ	D→S		S→D	
	R@1	AP	R@1	AP
0.10	94.58	95.25	92.87	93.81
0.50	95.01	95.62	93.01	94.05
1.00 (Ours)	**95.01**	**95.64**	**93.44**	**94.44**

Table 7. Ablation study on the 45° subset for different FPS, **best** and *second best*. Overall, 5 and 10 FPS outperform a little bit over 2 FPS.

Method	D→S						S→D					
	FPS = 2		FPS = 5		FPS = 10		FPS = 2		FPS = 5		FPS = 10	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
LPN	86.31	88.34	*86.45*	*88.47*	**86.59**	**88.56**	83.31	85.60	*84.17*	*86.26*	**84.20**	**86.33**
FSRA	88.59	90.25	*88.73*	*90.36*	**88.87**	**90.45**	*87.30*	*89.17*	*87.30*	89.11	**87.59**	**89.33**
DWDR	*91.73*	*92.96*	**91.87**	**93.06**	*91.73*	*92.96*	89.87	91.45	**90.01**	**91.57**	**90.01**	*91.56*
Sample4Geo	**96.29**	**96.75**	*96.14*	*96.62*	*96.14*	96.61	**95.29**	95.99	**95.29**	**96.00**	**95.29**	**96.00**
Ours	*96.29*	*96.80*	**96.43**	**96.92**	96.15	96.70	**96.01**	**96.57**	*95.58*	*96.21*	94.58	95.42



(a) 1 circle (**18** frames) **(b) 2** circles (**36** frames) **(c) 3** circles (**54** frames)

Figure 14. Visualizations of different trajectory lengths in Video2BEV transformation. We generate BEV by 3DGS via 3 different trajectory lengths. We find 1/3 trajectory (a) providing competitive visual quality.

6. Ablation study

6.1. Ablation study for loss weights

We employ a two-stage training strategy. In the first stage, the losses are the instance loss $\mathcal{L}_I$ and the contrastive loss $\mathcal{L}_C$, resulting in the total loss $\mathcal{L} = \mathcal{L}_I + \lambda\mathcal{L}_C$. The second stage utilizes the matching loss $\mathcal{L}_M$ only. By default, we set weights=1 following previous works. We add an ablation study on λ in the first stage (Tab. 6). We observe that $\lambda = 1$ achieves the best performance.

6.2. Ablation study for different FPS

UniV's necessity lies in **multi-view video** and **occlusion-heavy elevation angle**, not specific FPS. Previous experiments all focus on the 2-FPS subset in the UniV, and we additionally provide an ablation study for different FPS (see Tab. 7). Compared to a single image, as long as videos cover more areas, videos offer better robustness to occlusion and viewpoint changes. High FPS is not a key, but it provides more chances to see unoccluded areas. Therefore, 5/10 FPS outperforms a little bit over 2FPS in Tab. 7. How to effectively and efficiently process high-fps videos is also a future research direction.

6.3. Visualizations of different trajectory lengths in Video2BEV transformation

We adopt all frames, *i.e.*, 54 frames, in the 2-fps video by default. Here we provide visualizations of different trajectory lengths (see Fig. 14). We find 1/3 trajectory, *i.e.*, 18 frames, is enough to train 3DGS, achieving competitive visual quality.

References

[1] Alex Hanson, Allen Tu, Geng Lin, Vasu Singla, Matthias Zwicker, and Tom Goldstein. Speedy-splat: Fast 3d gaussian splatting with sparse pixels and sparse primitives. In *CVPR*, pages 21537–21546, 2025. 4