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

Hao Ju[1] Shaofei Huang[1] Si Liu[2] Zhedong Zheng[1†]
[1]Faculty of Science and Technology and Institute of Collaborative Innovation, University of Macau
[2]Institute of Artificial Intelligence, Beihang University
{yc47429,zhedongzheng}@um.edu.mo, shaofeihuang.ai@gmail.com, liusi@buaa.edu.cn

Abstract

Existing approaches to drone visual geo-localization predominantly adopt the image-based setting, where a single drone-view snapshot is matched with images from other platforms. Such task formulation, however, underutilizes the inherent video output of the drone and is sensitive to occlusions and viewpoint disparity. To address these limitations, we formulate a new video-based drone geo-localization task and propose the Video2BEV paradigm. This paradigm transforms the video into a Bird's Eye View (BEV), simplifying the subsequent ***inter-platform*** *matching process. In particular, we employ Gaussian Splatting to reconstruct a 3D scene and obtain the BEV projection. Different from the existing transform methods,* e.g.*, polar transform, our BEVs preserve more fine-grained details without significant distortion. To facilitate the discriminative* ***intra-platform*** *representation learning, our Video2BEV paradigm also incorporates a diffusion-based module for generating hard negative samples. To validate our approach, we introduce UniV, a new video-based geo-localization dataset that extends the image-based University-1652 dataset. UniV features flight paths at* 30° *and* 45° *elevation angles with increased frame rates of up to 10 frames per second (FPS). Extensive experiments on the UniV dataset show that our Video2BEV paradigm achieves competitive recall rates and outperforms conventional video-based methods. Compared to other competitive methods, our proposed approach exhibits robustness at lower elevations with more occlusions. The code is available at: https://github.com/HaoDot/Video2BEV-Open.*

1. Introduction

Drone visual geo-localization aims to retrieve images of the same location from another platform, such as satellite, using visual information captured by the drone. This process is typically supported by off-line GNSS metadata [74], enabling drones to self-localize even in GNSS-denied environments, such as urban canyons or rural areas. The prevailing approach follows an **image-based** matching paradigm [6, 8, 31, 58, 59, 74], where a single drone-captured snapshot serves as the query to retrieve the corresponding location from the satellite-viewed candidate pool. However, despite the advancements in image-based paradigms, two critical limitations persist. Both are due to drone flight height regulations [2, 19, 52]. **(1)** Drones typically operate at lower altitudes in cluttered environments, resulting in significant occlusions in the captured images from buildings, trees, and other foreground objects. Such occlusion can lead to a substantial loss or degradation of visual information in the drone image captured from a single viewpoint, making it difficult to establish accurate correspondences with satellite imagery. As shown in Fig. 1 (*left*), the core areas in the query are entirely obstructed by surrounding buildings. **(2)** Similarly, due to height limitation, drones typically capture images at oblique angles, while satellite images are predominantly acquired from a top-down perspective. The significant viewpoint disparity between the drone and satellite perspectives further increases the difficulty of matching.

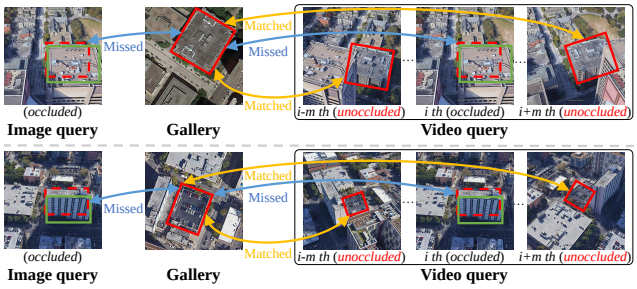
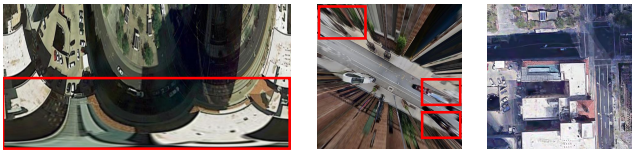


Figure 1. **Typical failure cases for image-based drone geo-localization.** For image queries (*left*), the core areas in ground-truth are occluded by another building, largely compromising the spatial matching. In contrast, video queries (*right*) usually contain unoccluded frames in a circling flight, and thus could reflect a more comprehensive view of the target location.

†Corresponding author.



(a) Satellite to Ground (*distorted and blurry*) (b) Ground to Satellite (*distorted and blurry*) (c) Drone video to Satellite image (Ours)

Figure 2. Prevailing image-based geometric transformation (a) Satellite to Ground transformation by the polar transformation [41], (b) Ground to Satellite transformation by the spherical transformation [61]. Our Drone video to Satellite transformation is shown in (c). Compared to image-based approaches, our method, fully leveraging the comprehensive view from the free-of-lunch drone videos, mitigates severe distortion and blurring.

To address the limitations of the image-based geo-localization paradigm, we formulate a new **video-based** drone geo-localization task and propose a corresponding paradigm named **Video2BEV**, which leverages drone videos and transforms them into Bird's-Eye View (BEV) representations for drone-satellite matching. **(1)** Different from traditional single-view image approaches, our Video2BEV paradigm resorts to the multi-view nature of video to recover occluded regions and improve matching robustness. As shown in Fig. 1 (*right*), even though core areas of the target location are occluded in a certain drone-captured frame, we still can recover such core areas from other frames with different viewpoints. **(2)** In this work, we reconstruct the view in the BEV format, since BEV representation aligns with the satellite's top-down viewpoint, thus reducing the **inter-platform** discrepancy. To convert input image into a calibrated format, existing methods usually apply 2D geometric transformations [41, 61], but suffer from spatial distortion and blurring (see Fig. 2 (a, b)). Inspired by the success of 3D Gaussian Splatting (3DGS) in reconstruction [4], we introduce a new 3D-aware transformation. In particular, we leverage 3DGS to reconstruct the 3D scene based on the multi-view snapshots from the drone video, and then obtain the BEV representation via projection. As shown in Fig. 2 (c), the BEV generated by our Video2BEV transformation exhibits fine-grained textures with minimal distortion or blurring, thus facilitating the subsequent **inter-platform** matching. Furthermore, considering the nearby location with a similar visual appearance, the proposed Video2BEV paradigm further incorporates a diffusion-based hard negative synthesis module. This module generates BEV representations that retain original semantic content but with different fine-grained discrepancies, serving as hard negatives during training. By incorporating these challenging samples, the model learns to discriminate **intra-platform** samples from highly similar yet geographically distinct locations.

Finally, to support the video-based geo-localization task, we introduce a new dataset UniV with 2 drone videos per location accompanying with 16 ground-view snapshots and

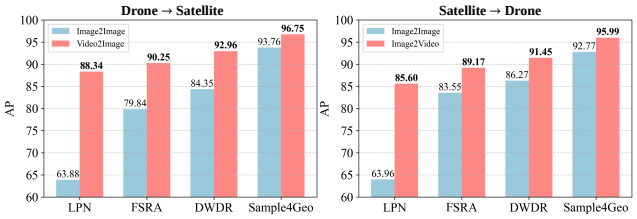


Figure 3. Performance comparisons of leveraging image data (Image2Image) or video data (Video2Image or Image2Video) with different methods including LPN [58], FSRA [6], DWDR [59], and Sample4Geo [8]. We report the Average Precision (AP) metric. For a fair comparison, we keep the same number of data in the gallery. We could observe that all re-implemented methods achieve better performance when adopting video query or gallery.

1 satellite-view image, which is closer to the real-world deployment. With the help of unobstructed frames, the video input significantly reduces the impact of occasional occlusions present in the single image, thereby improving overall performance on all our re-implemented methods. For instance, LPN [58] receives +24.46% AP increment (see Fig. 3). In brief, our contributions are:

- We formulate a new video-based geo-localization task and propose the Video2BEV paradigm that transforms drone-view videos into BEV representations with the assistance of 3DGS, simplifying the subsequent **inter-platform** matching process. To further enhance the **intra-platform** representation learning, we introduce a diffusion-based hard negative sample synthesis module, which generates challenging training samples to expand data diversity and improve discriminative capability.
- To validate the video-based drone geo-localization task, we introduce a new dataset, called UniV, with 3,304 drone flight videos with corresponding satellite and ground-view images. Our experiments reveals two insights: (1) Video-based data shows consistent performance advantages over single-frame image retrieval across multiple metrics (see Fig. 3). (2) The proposed Video2BEV achieves 96.80 AP on Drone Video → Satellite, outperforming other competitive methods. Furthermore, the trained model shows strong generalization capabilities, maintaining 91.50 AP when tested on the unseen real-world dataset, *i.e.*, SUES-200, without fine-tuning.

2. Related Work

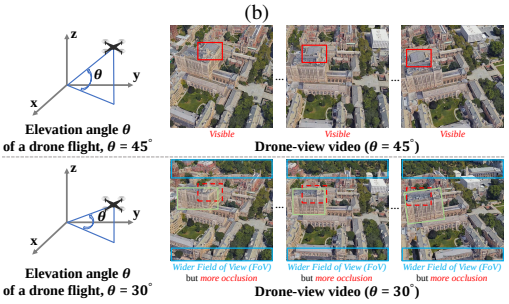
Image-based Geo-localization. Image-based geo-localization, which is usually regarded as a sub-task of image retrieval, applies image query to determine locations [63, 71]. The primary challenge of this task is the large appearance discrepancy due to different viewpoints across platforms, including ground [17], satellite [11], and drone [9, 76]. Previous methods can be coarsely divided into two families: image-level alignment and feature-level

Table 1. (a) Dataset comparisons between UniV and other visual geo-localization datasets. G, S, and D denote ground-view, satellite-view, and drone-view, respectively. We enable video modality and add another common elevation angle of drone flight. (b) Elevation angles θ illustration. Top panel shows $\theta = 45^\circ$ and bottom panel displays $\theta = 30^\circ$. With a lower elevation angle, the new flight captures the target location with wider *Field of View (FoV)* but more *occlusions*, thereby posing more challenges for drone visual geo-localization.

(a)

Datasets	Platforms	#data per location	Modality	Elevation
CVUSA [64]	G, S	1 image + 1 image	Image	N/A
Lin *et al.* [27]	G, S	1 image + 1 image	Image	45°
Vo *et al.* [56]	G, S	1 image + 1 image	Image	N/A
Tian *et al.* [50]	G, S	1 image + 1 image	Image	45°
CVACT [28]	G, S	1 image + 1 image	Image	N/A
Vigor [78]	G, S	2 images + 1 image	Image	N/A
SUES-200 [77]	S, D	(1 + 50 × 4) images	Image	45° ∼ 70°
University-1652 [74]	G, S, D	(16 + 1 + 54) images	Image	45°
GeoText-1652 [5]	G, S, D	(16 + 1 + 54) images + 180 texts	Image + Text	45°
UniV	G, S, D	(16 +1) images + **2 videos**	Image +**Video**	**30°**, 45°

(b)



alignment. **(1) For image-level alignment**, Shi *et al.* [41] leverage polar transformation to warp satellite images to the ground view. Similarly, Wang *et al.* [61, 67] transform ground images to the satellite view. Regmi *et al.* [36] synthesize aerial images from ground images to facilitate matching with satellite view via Generative Adversarial Networks (GANs). Tian *et al.* [49] employ GANs to transform drone-view images into satellite-view images. Andrea *et al.* [54] convert drone-view images to ground views through 3D reconstruction but the output is with distortions. **(2) For feature-level alignment,** Dai *et al.* [6] and Wang *et al.* [58] establish feature alignment in a region-correspondence manner. Some methods [26, 46] focus on key-point alignment. Other methods aim to improve the discriminative ability of neural networks with tailored modules, such as lite-transformer encoder [62], layer-to-layer attention block [65, 79], adaptive integration module [47], strong backbones [66], adaptation information consistency module [23], spatially-adaptive denormalization [60] and well-designed loss functions, *e.g.*, semantic augmentation loss [70], contrastive loss [8, 22, 31, 71], dynamic weighted decorrelation regularization [59], peer learning [69], instance loss [74] and the optimal transport [43]. Additionally, other methods [28, 37, 42] fuse the extra orientation meta-information from GNSS with extracted features. However, previous methods overlook the viewpoint variation within drone-view videos, thinking in an image-based setting. Our method is among the early attempts to leverage the viewpoint variation of drone-view videos to transform drone-view video to Bird's-Eye View (BEV), thereby reducing the viewpoint disparity between drone and satellite views.

Video-based Geo-localization While video understanding has been a focus of the computer vision community for decades, the problem remains challenging due to the complexity added by the time dimension and the volume of data. Early works [3, 12, 40] leverage two-stream convolution networks to fuse spatial semantic information with motion information. Subsequently, attention mechanisms [7, 13, 34] have been introduced for long-term video understanding, such as vanilla self-attention [1, 55], shift window [29, 30], masked auto-encoder [14, 51], and local spatiotemporal attention [45]. Recently, large language models [48] have also shown their superiority in video understanding. In visual geo-localization, videos contain more visual information captured through the camera's trajectory, which can provide more comprehensive information compared to images. Vyas *et al.* [57] are the first to collect ground-view data in video format and propose a hierarchical approach for processing clips of ground-view videos. Regmi *et al.* [37] leverage the geo-temporal proximity between the ground-view videos and GNSS locations to extract coherent features from videos. Expanding to a global scale, Kulkarni *et al.* [21] introduce a large-scale ground-view video dataset for worldwide geo-localization. Different from the ground-view videos, drone-view videos typically contain multi-view and multi-scale information for the target location [32]. In this paper, we collect drone-view data in video format and propose a video-based geo-localization dataset. Rather than adopting the off-the-shelf video backbone, we propose a Video2BEV transformation to leverage the 3D geometric correspondences and enable a straightforward spatial alignment for matching.

3. The UniV Dataset

Given the lack of a video-based drone geo-localization benchmark, we collect a new dataset dubbed *UniV* involving the video modality. We follow the location information and the protocol of the existing University-1652 dataset [74]. The UniV dataset encompasses 1,652 locations in 72 universities from three platforms, *i.e.*, ground, satellite, and drone cameras. In particular, the UniV dataset contains 16 ground-view images and 1 satellite-view image for each location and the training set of UniV dataset contains 701 locations, while the test set in the UniV dataset includes other 951 locations. There are no overlapping locations between the training and test sets. The proposed UniV dataset is different from the image-based University-1652

and other datasets in two primary aspects, *i.e.*, modality and elevation-angle expansions (see Tab. 1a).

Modality Expansion. Existing datasets [28, 56, 64, 78] collect data from two platforms, *e.g.*, satellite and ground. Although some datasets [50, 74, 77] include drone views, collected data is still in an image format. We adopt similar operations as the University-1652 but collect drone-view data in a video format. Specifically, we leverage the 3D engine of google earth [33] to simulate the real-world movement of a drone equipped with a camera. To collect video data containing various scales and viewpoints, we leverage the dynamic viewpoints within the 3D engine and set the moving viewpoints along a spiral curve for moving around the target location in three circles, closely approximating real-world drone flights. All videos are collected in 30 frame rate. Considering the video redundancy, in practice, we subsample videos along the temporal dimension, resulting in frame rates of 2, 5, and 10 for further processing.

Elevation-angle Expansion. Conventional datasets [27, 50, 74] collect drone data in a fixed elevation angle, *i.e.*, 45°, which does not fully simulate the real-world cases. Therefore, we add one new synthetic flying path at another common setting, *i.e.*, a lower elevation angle 30°. The new flying path poses two new challenges (see Tab. 1b). First, drones flying at a 30° elevation angle capture scenes that include the target location and more surrounding areas, providing a wider Field of View (FoV), thus introducing disruptions for the center target location during matching. Second, flight paths at a lower elevation angle lead to more occluded cases, which lay over the core areas of the target location. It poses challenges for mining the discriminative frames in the video, whereas it becomes easier when captured at a 45° elevation. Therefore, the proposed dataset could further evaluate the robustness of methods against more disruptions and occlusions, which is closer to real-world drone geo-localization usage.

Discussion. The contribution to the community. Different from existing datasets [27, 50, 74], the proposed UniV expands the modality from image to video (see Tab. 1a), facilitating the development of robust drone visual geo-localization. A single image provides limited information about the corresponding location. When core areas of the location are occluded, single-image queries can not produce reliable matching results (see Fig. 1). In such cases, the video contains both occluded and unoccluded frames. One frame may contain core-area information to complement another frame and together they can provide robust and complete information required for drone visual geo-location. In this way, all re-implemented methods perform better when adopting video data (see Fig. 3). Moreover, the UniV dataset also introduces a new real-world challenge for drone visual geo-localization. The new elevation angle of 30° is typical in real-world flights*. The 30° elevation angle faces more occlusion cases (see Tab. 1b), simulating outputs of real-world drone flights.

4. Method

4.1. Video2BEV Transformation

During the flight around the target location, the viewpoints of the camera change dynamically, resulting in captured drone-view videos that contain rich multi-view information about both the target location and surrounding areas. We explicitly leverage the multi-view information and transform the drone-view video into Bird's Eye Views (BEVs). In doing so, we ease the learning process for the subsequent model. Instead of learning geometry correspondence and feature correspondences simultaneously [4], the subsequent model only needs to learn the feature mapping relationship between two views, thus significantly facilitating network convergence. As shown in the left part of Fig. 4, given the drone-view video containing multi-view images, we estimate corresponding camera poses by structure from motion [53] and reconstruct the scene containing the target locations utilizing 3D Gaussian Splatting (3DGS) [18]. After reconstructing the scene, we adopt the normalized camera pose and the unit vector in the world coordinate to calculate the BEV camera pose and render BEVs. In particular, the vanilla 3DGS takes less than 8 seconds to render 50 BEVs with the shape of 512×512 on a NVIDIA 4090 GPU. Since we usually foreknow the search area in practice, the BEV generation process can be regarded as an off-line pre-processing. BEVs rendered with our Video2BEV transformation module do not suffer from severe distortion (see Fig. 4), thereby aiding in the subsequent **inter-platform** matching process with the satellite images.

4.2. Hard Negative Sample Synthesis

Negative samples play a significant role in discriminative metric learning. Current negative sample mining strategies [8, 31, 71] cannot ensure the quantity and quality of negative samples, as challenging samples are inherently scarce, and the selected negative samples do not necessarily exhibit similar architectural styles or consistent semantic details as the original samples. In order to bypass these drawbacks, we propose to generate diverse BEV representations as hard negative samples through a fine-tuned diffusion model, which is shown in the right of Fig. 4. After transforming the drone-view video to BEVs x_{bev} via our Video2BEV transformation, we utilize a visual-LLM [16] to generate captions for both BEV and satellite-view images. After obtaining captions for the BEVs and satellite images, we fine-tune a stable diffusion network [39] with

*The United States and the United Kingdom allow drone flights up to 400 feet; China restricts drones up to 120 meters.

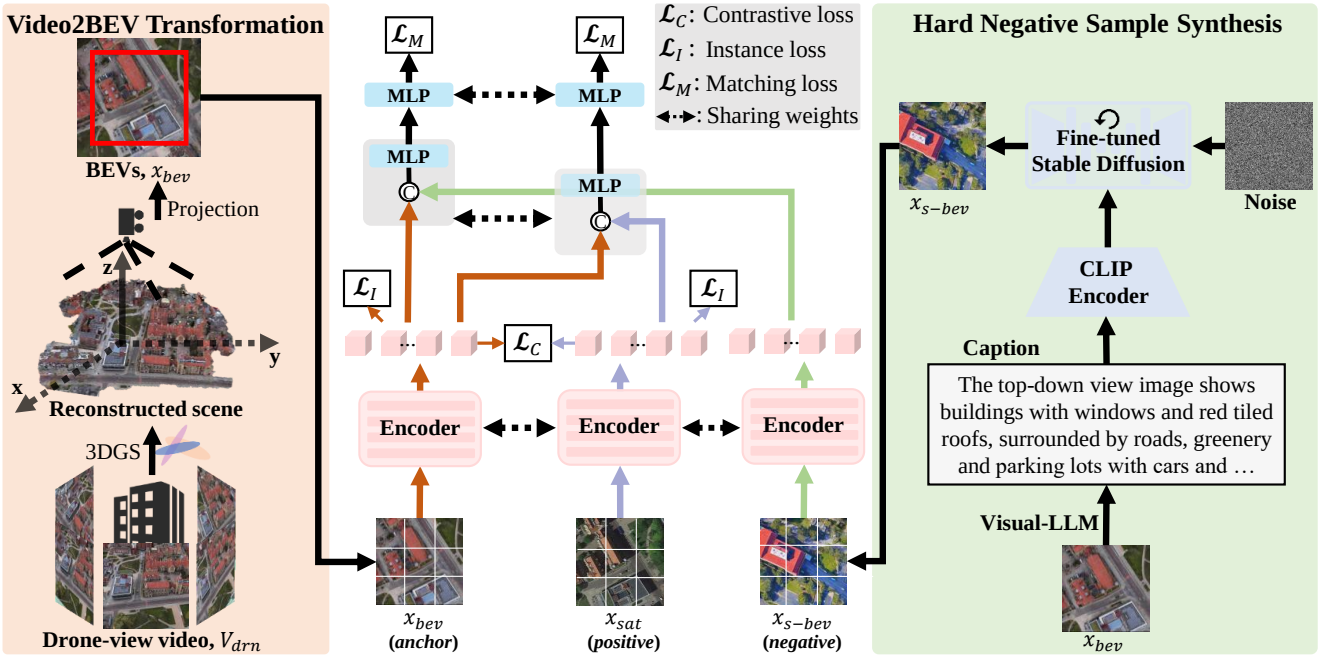


Figure 4. The overview of the Video2BEV paradigm. **Video2BEV Transformation (*left*).** Given drone-view video V_{drn} containing multi-view frames, we adopt 3D Gaussian Splatting (3DGS) to reconstruct the scene at first. Then we render the scene from a Bird-Eye-View to get the projection (BEVs). Considering the region of the core area, we further crop BEVs for training. We can observe that BEVs exhibit resemblances to the corresponding satellite-view images. **Hard Negative Sample Synthesis (*right*).** Given captions generated by an off-the-shelf visual-LLM [16], we fine-tune a stable-diffusion model [39] with LoRA [15], and conduct inference to synthesize samples which serve as negative samples for subsequent usage. **Model Architecture (*middle*).** Given outputs of the proposed Video2BEV transformation, we extract embeddings by a shared encoder for satellite images x_{sat} and BEVs x_{bev}, supervised by the contrastive loss $\mathcal{L}_C$ and the instance loss $\mathcal{L}_I$. Then we extract embeddings from synthetic BEVs $x_{s\text{-}bev}$ and adopt MLP to fuse both positive and negative samples, supervised by the matching loss $\mathcal{L}_M$. Similar operations for satellite-view images are omitted.

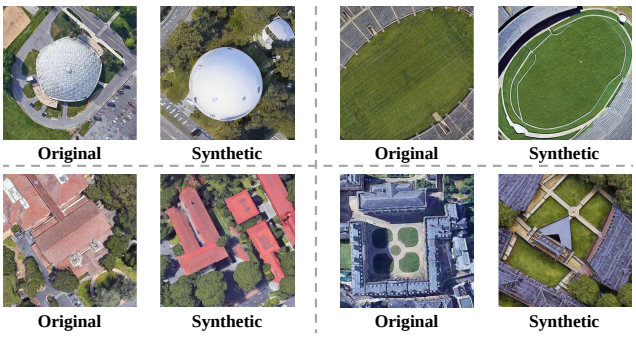


Figure 5. Visualizations of original images and synthetic hard negatives. Synthetic negatives exhibit similar colors and structures to original images, which assures the quality of negatives.

LoRA [15] to generate diverse synthetic images, during which we freeze the CLIP text encoder [35]. The outputs of this model are negative samples for the subsequent step. Since the transformed BEV and satellite images share the same top-down viewpoint and semantic content, we conduct inference on the same diffusion model to generate negative samples for both BEV and satellite-view images using corresponding captions. We provide visualizations of original and synthetic images in Fig. 5. Synthetic negative samples exhibit similar semantic contents as the original ones but with different fine-grained information, thus enhancing the discriminative **intra-platform** representation learning.

4.3. Model Optimization

We adopt a general architecture from vision-language models [24, 25], enhanced by BEVs and synthetic negative samples. The model architecture is shown in the middle of Fig. 4 and is optimized in a **two-stage** manner following [25, 68]. In the **first stage**, we transform the drone-view video to BEVs by the proposed Video2BEV transformation (see the left part of Fig. 4). Then, we adopt a shared encoder to extract embeddings from paired BEV and satellite-view images. The encoder is ViT-S [10] excluding the classifier. The supervisions of this stage are the instance loss $\mathcal{L}_I$ [75] with the square-ring partition [58] and the contrastive loss $\mathcal{L}_C$ [24]. We apply multiple classifier modules to each part of the embeddings (similar to LPN [58]), yielding the location probability of two views which are denoted as $\hat{p}_{sat}$ and $\hat{p}_{bev}$ respectively. The instance loss $\mathcal{L}_I$ is formulated as the location classification as :

$$\mathcal{L}_I = -log(\hat{p}_{sat}) - log(\hat{p}_{bev}). \quad (1)$$

Then we accumulate instance losses from multiple parts to form the final instance loss. For the contrastive loss, given a pair of satellite-view and BEV images, the satellite-to-BEV similarity is defined as:

$$S_{sat2bev} = \frac{exp(s(f_{sat}, f_{bev})/\tau)}{\Sigma_{j=1}^{N} exp(s(f_{sat}, f_{bev}^{j})/\tau)}, \quad (2)$$

where f_{sat} and f_{bev} are embeddings of the same location from two platforms, and f_{bev}^{j} denotes the sample within the mini-batch. τ is a learnable temperature parameter. $s(\cdot,\cdot)$ denotes the cosine similarity. Similarly, the BEV-to-satellite similarity is $S_{bev2sat}$ and the contrastive loss $\mathcal{L}_C$ is:

$$\mathcal{L}_C = -\frac{1}{2}(log(S_{sat2bev}) + log(S_{bev2sat})). \quad (3)$$

In the **second stage**, we employ a two-layer MLP alongside the square-ring partition [58] to fuse two embeddings obtained from anchor-positive or anchor-negative samples Specifically, when BEV serves as the anchor, satellite and synthetic BEV serve as the positive sample and the negative sample, respectively. Similarly, when satellite acts as positive samples, BEV and synthetic satellite act as positive and negative samples respectively. Then we project the fused embeddings into the two-dimensional space using another MLP. Given inputs from paired samples, the matching loss $\mathcal{L}_M$ between them is calculated as:

$$\mathcal{L}_M = -(p_m log(\hat{p}_m) + (1 - p_m) log(1 - \hat{p}_m)), \quad (4)$$

where $\hat{p}_m$ is the estimated matching probability and p_m is a ground-truth binary label. If the two input data do not contain the synthetic data and are both from the same location, then $p_m = 1$; otherwise, $p_m = 0$. Specifically, for the BEVs, we calculate the matching loss two times. For the first calculation, we rank the similarity S and select three negative samples from the satellite-view images, ensuring that these samples do not belong to the same location simultaneously. For the second calculation, we similarly select another three negative samples from the synthetic BEVs, which are actually hard negatives from the same location. We apply similar operations to the satellite-view input. Finally, we accumulate and average matching losses across different combinations of inputs. In summary, the loss functions in our method include the instance loss $\mathcal{L}_I$, the contrastive loss $\mathcal{L}_C$, and the matching loss $\mathcal{L}_M$. Specifically, in the first stage of training, we optimize the encoder, classifier modules of our model, the temperature parameter τ with the instance loss $\mathcal{L}_I$ and the contrastive loss $\mathcal{L}_C$. Subsequently, we freeze the parameters fine-tuned in the first stage and train the MLPs from scratch in the second stage under the supervision of matching loss $\mathcal{L}_M$.

Discussion. What are the advantages of the synthetic negative samples? Inspired by successes in other fine-grained tasks [44, 72, 73], we encourage the model "see" more samples to prevent over-fitting as well as facilitate discriminative intra-platform feature learning. We are similar to GeNIe [20] in that both methods alter the image representation of the target object to generate hard negative samples. However, it is worth noting that there are two primary differences. (1) We have a larger modification space, and the negative sample pool is no longer constrained to a fixed size. Different from the GeNIe in changing limited categories for classification, we perturb the initial noise of the diffusion model, and it leads to diverse generations. Utilizing the diffusion model, we can theoretically generate an infinite number of images as negative samples, expanding the negative sample pool significantly. (2) We retain the semantic content of the anchor samples in our synthetic negative samples. This is because we employ identical captions from the original samples to synthesize the negative samples while only changing the initial noise. This slight modification ensures that our negative samples are appropriately challenging, and encourages the model to check the fine-grained discrepancies among semantical-similar samples.

5. Experiment

Implementation Details. Since the captions for the two views (drone and satellite) employ different wording to describe the same location, we synthesize samples based on the text and generate separate negative samples for each view. All input images are resized to 256×256. We train the first stage of the proposed model with the AdamW optimizer, with a batch size of 140, for 140 epochs, and a learning rate of $2e^{-5}$ and $2e^{-4}$ for the encoder and other modules in the first stage respectively. Then we freeze parameters in the first stage and train the second stage from scratch with a similar training configuration. During the test stage, we utilize the similarity scores from the first stage to select the top 32 samples from the gallery, and then re-rank these top 32 samples in the second stage. More details are provided in the supplementary materials.

Evaluation Metrics. Satellite-view data is in image format, while drone-view data is collected in video format. We can treat drone view data as images or video. In this paper, we adopt the video setting for the evaluation of competitive methods and our method. Specifically, we treat a drone video as an individual query or gallery by averaging the similarity scores of the images within the video in a late fusion manner. There is a similar averaging operation on similarity scores of the BEVs which is also in video format.

5.1. Comparisons with Competitive Methods

Quantitative Results. As shown in Tab. 2a, we compare the proposed method with other competitive methods on the UniV dataset. The performance of our method has surpassed that of other competitive methods [6, 8, 58, 59]. On the 45° subset, our method achieves gains of 0.30% Re-

Table 2. (a) Comparisons on the UniV for geo-localization between Drone (D) and Satellite (S) platforms. R@1 is recall at top1. AP (%) is average precision (high is good). (b) Comparisons in terms of an out-of-distribution testing on the SUES-200 (45° test set). Our method still yields the best results.

(a)

Method	$\theta = 45^\circ$				$\theta = 30^\circ$			
	D→S		S→D		D→S		S→D	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
LPN [58]	86.31	88.34	83.31	85.60	68.62	72.50	67.76	71.30
FSRA [6]	88.59	90.25	87.30	89.17	81.60	84.17	77.89	81.00
DWDR [59]	91.73	92.96	89.87	91.45	88.02	89.81	85.59	87.85
Sample4Geo [8]	96.29	96.75	95.29	95.99	83.02	86.00	80.45	82.68
Ours	**96.29**	**96.80**	**96.01**	**96.57**	**91.73**	**93.01**	**92.58**	**93.65**

(b)

Method	D→S		S→D	
	R@1	AP	R@1	AP
LPN	41.25	49.05	18.75	26.35
FSRA	48.75	54.64	32.50	40.09
DWDR	71.25	75.02	70.00	74.60
Sample4Geo	81.25	84.14	86.25	88.80
Ours	**89.74**	**91.50**	**91.25**	**92.53**

call and 0.58% AP for satellite → drone compared to the second-best method. On the 30° subset, all methods experience a performance drop. As shown in Fig. 6, we highlight some imperfect reconstructed regions by the Video2BEV transformation. The lower elevation of the drone flights raises more occlusions (see Tab. 1b), which also compromises our Video2BEV transformation. Compared to the second best method, our method is still robust, receiving improvements of 3.2% AP for drone → satellite and 5.8% AP for satellite → drone, respectively (see Tab. 2a). All methods are compared in the video setting, which means we temporally average the outputs of frames in a video from the drone view. For methods with officially released weights (Sample4Geo [8], DWDR [59]), we test these methods on the 45° test set directly and subsequently retrain and evaluate these methods on the 30° subsets. For methods without official weights (LPN [58], FSRA [6]), we retrain them on both the 45° and 30° subsets to ensure a fair comparison.

Out-of-Distribution (OOD) Scalability. We also evaluate the model trained on the UniV dataset (45°) on the unseen SUES-200 45° test set in an OOD manner. SUES-200 dataset contains dense frames collected in real-world environments, including real-world light, shadow transformations, and disturbances. We observe that our method shows strong OOD potential, surpassing the runner-up method by more than 7% AP for drone → satellite (see Tab. 2b). More results on robustness against weather and other real-world variants can be found in the **Suppl.**

Qualitative Results. We show qualitative results of the drone geo-localization on the UniV and SUES-200 datasets (see Fig. 7). In our method, drone-view videos are transformed to BEVs by the proposed Video2BEV transformation and we choose the representative sample from the BEV sequence for visualizations. For drone → satellite, we observe that the proposed method effectively retrieves reasonable locations with similar structural features, such as cross-shaped roofs and roofs equipped with solar panels. For satellite → drone, we find a similar result. Our method successfully retrieves true-matched results at the top of the candidate list among images with similar contents. We add more visualizations in the supplementary material.

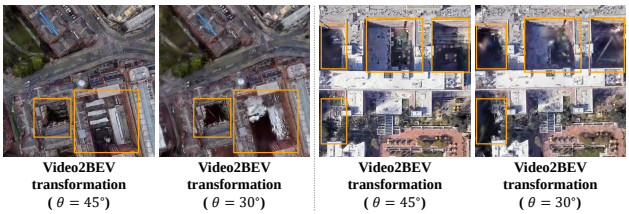


Figure 6. The transformed BEV comparison of videos with different evaluation θ. We highlight the challenging regions.

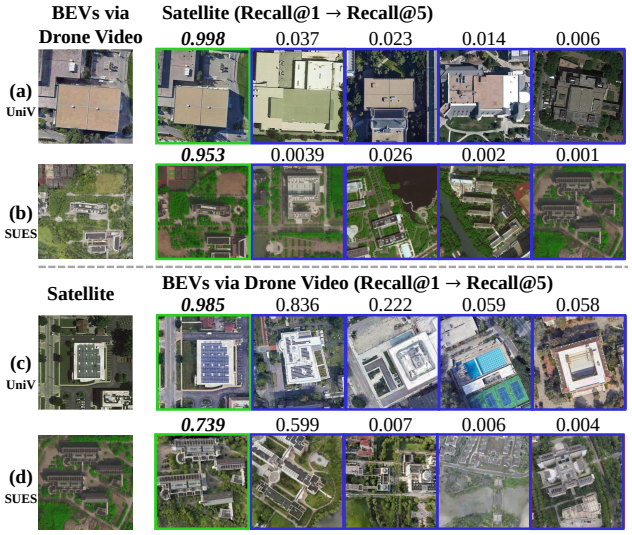


Figure 7. Qualitative results on the UniV (a, c) and SUES-200 (b,d) dataset for Drone → Satellite and Satellite → Drone. We depict the transformed output (BEVs) as the query or gallery. Given queries (left) from different platforms, matched galleries are in green box, and mismatched galleries are in blue box. The scores on the top are similarity scores estimated by the proposed method.

5.2. Ablation Study and Further Discussion

Effect of Primary Components. We conduct ablation studies on the UniV dataset (45° subset). We employ the first stage of our method as the baseline (**Baseline**), which consists of a shared backbone supervised with the instance loss and contrastive loss. The input data for the baseline are drone-view videos and satellite-view images. Then, we transform drone-view videos to BEVs via the proposed Video2BEV transformation and adopt BEVs as input for the baseline, denoted as **BEV**. Next, we introduce the second stage of our method to the baseline, which is supervised by the matching loss, denoting **Two Stage**. The negative samples for this architecture are from in-batch samples [24]. Finally, we incorporate the synthetic negative samples in

Table 3. Ablation studies on: (a) Video2BEV transformation, the second stage of our method, and synthetic negative samples. (b) Different training strategies. **Train Together**: we fine-tune the first stage based on the weights pre-trained on ImageNet [38], and train the second stage from scratch. **Fine-tune**: we load fine-tuned first-stage weights on UniV, and then train both the first stage and the second stage. **Freeze**: we load fine-tuned first-stage weights on UniV, then fix the first-stage weights and only train the second stage from scratch. Notably, the **Freeze** strategy yields the best results. (c) Re-ranking different top-k samples in the second stage of our method. Considering the balance between performance and testing time, we choose to re-rank top-32 samples. D and S denote Drone and Satellite, respectively.

(a)

Method	Video2BEV transformation	Second stage	Synthetic negatives	D→S		S→D	
				R@1	AP	R@1	AP
Baseline	✗	✗	✗	89.87	91.28	90.01	91.36
BEVs	✓	✗	✗	95.01	95.64	93.44	94.44
Two Stage	✓	✓	✗	95.86	96.48	95.01	95.78
Ours	✓	✓	✓	**96.29**	**96.80**	**96.01**	**96.57**

(b)

Strategy	Load fine-tuned first-stage weights	Train first stage	Train second stage	D→S		S→D	
				R@1	AP	R@1	AP
Train Together	✗	✓	✓	74.75	79.29	82.17	85.39
Fine-tune	✓	✓	✓	96.29	**96.83**	95.29	95.99
Freeze	✓	✗	✓	**96.29**	96.80	**96.01**	**96.57**

(c)

Top-K	D→S		S→D	
	R@1	AP	R@1	AP
8	96.01	96.52	95.58	96.10
16	96.01	96.51	95.72	96.25
32	96.29	96.80	**96.01**	96.57
64	96.29	96.81	**96.01**	**96.60**
128	**96.43**	96.98	**96.01**	**96.60**
256	**96.43**	**96.99**	**96.01**	**96.60**
512	**96.43**	**96.99**	**96.01**	**96.60**

Sec. 4.2 to train the second stage of our method and form the final version of our method, referred to as **Ours**. As shown in Tab. 3a, BEVs receive the largest performance improvement. We attribute this improvement to the reduction of the appearance gap between the drone-view images and the satellite-view images through the proposed Video2BEV transformation. Additionally, synthetic negative samples contribute to a substantial performance boost due to the enhanced quality of the negative samples for the second stage of Ours. The two-stage method (Two Stage) also receives improved performance, indicating that many false negative predictions are ranked within the range of the top 32. A fine-grained re-ranking can effectively rectify the matching results from the first stage of our method.

Effect of Training Strategies. We explore three different strategies for training. For the **Train Together** strategy, we load the matched weights pre-trained on the ImageNet dataset [38]. Then we fine-tune the first stage of the proposed model and train the second stage of the proposed model from scratch. The **Fine-tune** strategy entails loading fine-tuned weights of the first stage on the UniV dataset. After this, we fine-tune the first stage with a smaller learning rate while training the second stage from scratch. The **Freeze** strategy consists of loading fine-tuned weights of the first stage on UniV, then fixing all weights of the first stage, while training the second stage from scratch. The results of three training strategies are in Tab. 3b. The **Train Together** strategy yields the worst results. We attribute this to the difficulty of training both stages simultaneously, as the first stage of the proposed model is designed for coarse-grained retrieval, while the second stage of the proposed model focuses on fine-grained retrieval, relying on the output of the first stage. When both stages are trained together, the first stage fails to retrieve reliable candidates for the second stage, affecting the overall training process. The **Fine-tune** strategy achieves a significant performance boost, as the first stage is able to produce reliable embeddings for the second stage. Finally, we freeze the first stage after loading its corresponding weight. The **Freeze** strategy yields the best result, and we adopt this strategy.

Effect of Re-ranking Top-K Samples. During the test stage, we select top-k samples from the gallery, leveraging the similarity score from the first stage of our method and subsequently re-rank these samples by the second stage. We conduct hyper-parameter experiments with varying values of top-k, and select k $\in \{8, 16, 32, 64, 128, 256, 512\}$ (see Tab. 3c). Re-ranking the top-512 and top-256 samples yields the best performance and re-ranking the top-256, top-128, top-64, and top-32 samples results in a slight performance drop, respectively. Re-ranking the top-16 and top-8 samples leads to a further decline in performance. Considering the balance between the performance and the testing time, we re-rank the top 32 samples as default.

6. Conclusion

In this work, we propose to leverage videos to mitigate the impact of environmental constraints in drone visual geo-localization. We propose a new Video2BEV paradigm that transforms drone-view videos into Bird's Eye View (BEV) images by 3D gaussian splatting. This transformation effectively reduces the **inter-platform** viewpoint disparity between the drone view and the satellite view. Our Video2BEV paradigm also includes a diffusion-based module to generate negative samples, enhancing the **intra-platform** discriminative ability of the model. To support the video setting and validate the proposed framework, we introduce the UniV dataset, a new video-based drone geo-localization dataset. The dataset includes flight paths of the drone at 30° and 45° elevation angles and corresponding videos recorded at up to 10 frames per second. Extensive experiment validates that our Video2BEV paradigm outperforms other competitive approaches in both supervised setting on UniV and OOD testing on unseen SUES-200.

7. Acknowledgment

We acknowledge support from Guangdong Basic and Applied Basic Research Foundation 2025A1515012281, Nanjing Municipal Science and Technology Bureau 202401035, University of Macau MYRG-GRG2024-00077-FST-UMDF, University of Macau (Conference Grant – FST, CG2025-FST), National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (No.62461160308, U23B2010), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01161).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 3

[2] GR Bhat, MA Dudhedia, RA Panchal, YS Shirke, NR Angane, SR Khonde, SP Khedkar, JR Pansare, SS Bere, RM Wahul, et al. Autonomous drones and their influence on standardization of rules and regulations for operating–a brief overview. *Results in Control and Optimization*, 14:100401, 2024. 1

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 3

[4] Florian Chabot, Nicolas Granger, and Guillaume Lapouge. Gaussianbev: 3d gaussian representation meets perception models for bev segmentation. *arXiv*, 2024. 2, 4

[5] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *ECCV*, pages 213–231, 2024. 3

[6] Ming Dai, Jianhong Hu, Jiedong Zhuang, and Enhui Zheng. A transformer-based feature segmentation and region alignment method for uav-view geo-localization. *IEEE TCSVT*, 32(7):4376–4389, 2021. 1, 2, 3, 6, 7

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv*, 2023. 3

[8] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *ICCV*, pages 16847–16856, 2023. 1, 2, 3, 4, 6, 7

[9] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE TPAMI*, 44(11):7778–7796, 2021. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 5

[11] Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, and Mubarak Shah. Multiview aerial visual recognition (mavrec): Can multi-view improve aerial visual perception? In *CVPR*, pages 22678–22690, 2024. 2

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 3

[13] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE TPAMI*, 45(1):87–110, 2022. 3

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 5

[16] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv*, 2024. 4, 5

[17] Andy V Huynh, Lauren E Gillespie, Jael Lopez-Saucedo, Claire Tang, Rohan Sikand, and Moisés Expósito-Alonso. Contrastive ground-level image and remote sensing pre-training improves representation learning for natural world imagery. In *ECCV*, pages 173–190, 2024. 2

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139–1, 2023. 4

[19] Anna Konert and Piotr Kasprzyk. Very low level flight rules for manned and unmanned aircraft operations. *Journal of Intelligent & Robotic Systems*, 110(2):82, 2024. 1

[20] Soroush Abbasi Koohpayegani, Anuj Singh, KL Navaneet, Hadi Jamali-Rad, and Hamed Pirsiavash. Genie: Generative hard negative images through diffusion. *arXiv*, 2023. 6

[21] Parth Parag Kulkarni, Gaurav Kumar Nayak, and Mubarak Shah. Cityguessr: City-level video geo-localization on a global scale. In *ECCV*, pages 293–311, 2024. 3

[22] Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *CVPR*, pages 16719–16729, 2024. 3

[23] Haoyuan Li, Chang Xu, Wen Yang, Huai Yu, and Gui-Song Xia. Learning cross-view visual geo-localization without ground truth. *IEEE TGRS*, 2024. 3

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021. 5, 7

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 5

[26] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE TIP*, 31:3780–3792, 2022. 3

[27] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *CVPR*, pages 5007–5015, 2015. 3, 4

[28] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *CVPR*, pages 5624–5633, 2019. 3, 4

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[30] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 3

[31] Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. Congeo: Robust cross-view geo-localization across ground view variations. *ECCV*, pages 214–230, 2024. 1, 3, 4

[32] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 3

[33] Onisimo Mutanga and Lalit Kumar. Google earth engine applications, 2019. 4

[34] Simone Alberto Peirone, Francesca Pistilli, Antonio Alliegro, and Giuseppe Averta. A backpack full of skills: Egocentric video understanding with diverse task perspectives. In *CVPR*, pages 18275–18285, 2024. 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 5

[36] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *ICCV*, pages 470–479, 2019. 3

[37] Krishna Regmi and Mubarak Shah. Video geo-localization employing geo-temporal feature learning and gps trajectory smoothing. In *ICCV*, pages 12126–12135, 2021. 3

[38] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv*, 2021. 8

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 4, 5

[40] Xiaolong Shen, Zhedong Zheng, and Yi Yang. Stepnet: Spatial-temporal part-aware network for isolated sign language recognition. *IEEE TMM*, 20(7):1–19, 2024. 3

[41] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *NeurIPS*, 32, 2019. 2, 3

[42] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *CVPR*, pages 4064–4072, 2020. 3

[43] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *AAAI*, pages 11990–11997, 2020. 3

[44] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, pages 2107–2116, 2017. 6

[45] Jaewon Son, Jaehun Park, and Kwangsu Kim. Csta: Cnn-based spatiotemporal attention for video summarization. In *CVPR*, pages 18847–18856, 2024. 3

[46] Ze Song, Xudong Kang, Xiaohui Wei, Shutao Li, and Haibo Liu. Unified and real-time image geo-localization via fine-grained overlap estimation. *IEEE TIP*, 2024. 3

[47] Jian Sun, Hao Sun, Lin Lei, Kefeng Ji, and Gangyao Kuang. Tirsa: A three stage approach for uav-satellite cross-view geo-localization based on self-supervised feature enhancement. *IEEE TCSVT*, 2024. 3

[48] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv*, 2023. 3

[49] Xiaoyang Tian, Jie Shao, Deqiang Ouyang, and Heng Tao Shen. Uav-satellite view synthesis for cross-view geo-localization. *IEEE TCSVT*, 32(7):4804–4815, 2021. 3

[50] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *CVPR*, pages 3608–3616, 2017. 3, 4

[51] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 35:10078–10093, 2022. 3

[52] Thuy-Hang Tran and Dinh-Dung Nguyen. Management and regulation of drone operation in urban environment: A case study. *Social Sciences*, 11(10):474, 2022. 1

[53] Shimon Ullman. The interpretation of structure from motion. *Biological Sciences*, 203(1153):405–426, 1979. 4

[54] Andrea Vallone, Frederik Warburg, Hans Hansen, Søren Hauberg, and Javier Civera. Danish airs and grounds: A dataset for aerial-to-street-level place recognition and localization. *IEEE RAL*, 7(4):9207–9214, 2022. 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3

[56] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *ECCV*, pages 494–509, 2016. 3, 4

[57] Shruti Vyas, Chen Chen, and Mubarak Shah. Gama: Cross-view video geo-localization. In *ECCV*, pages 440–456, 2022. 3

[58] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE TCSVT*, 32(2):867–879, 2021. 1, 2, 3, 5, 6, 7

[59] Tingyu Wang, Zhedong Zheng, Zunjie Zhu, Yuhan Gao, Yi Yang, and Chenggang Yan. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization. *arXiv*, 2022. 1, 2, 3, 6, 7

[60] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. Multiple-environment self-adaptive network for aerial-view geo-localization. *PR*, 152: 110363, 2024. 3

[61] Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *NeurIPS*, 36, 2024. 2, 3

[62] Yuntao Wang, Jinpu Zhang, Ruonan Wei, Wenbo Gao, and Yuehuan Wang. Mfrgn: Multi-scale feature representation generalization network for ground-to-aerial geo-localization. In *ACM MM*, pages 2574–2583, 2024. 3

[63] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Image and object geo-localization. *IJCV*, 132(4): 1350–1392, 2024. 2

[64] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *ICCV*, pages 3961–3969, 2015. 3, 4

[65] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *NeurIPS*, 34:29009–29020, 2021. 3

[66] Jiaqiang Yang, Danyang Qin, Huapeng Tang, Sili Tao, Haoze Bie, and Lin Ma. Dinov2-based uav visual self-localization in low-altitude urban environments. *IEEE RAL*, 2025. 3

[67] Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-bev co-retrieval network. In *ECCV*, pages 74–90, 2024. 3

[68] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*, pages 25994–26009, 2022. 5

[69] Zelong Zeng, Zheng Wang, Fan Yang, and Shin'ichi Satoh. Geo-localization via ground-to-satellite cross-view image retrieval. *IEEE TMM*, 25:2176–2188, 2022. 3

[70] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *AAAI*, pages 3480–3488, 2023. 3

[71] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: Toward generic cross-view geolocalization via geometric disentanglement. *IEEE TPAMI*, 2024. 2, 3, 4

[72] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017. 6

[73] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 6

[74] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *ACM MM*, pages 1395–1403, 2020. 1, 3, 4

[75] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *IEEE TMM*, 16 (2):1–23, 2020. 5

[76] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE TPAMI*, 44(11):7380–7399, 2021. 2

[77] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE TCSVT*, 33(9):4825–4839, 2023. 3, 4

[78] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *CVPR*, pages 3640–3649, 2021. 3, 4

[79] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *CVPR*, pages 1162–1171, 2022. 3