remote sensing

MDPI

Article

Cross-View Geo-Localization via 3D Gaussian Splatting-Based Novel View Synthesis

Xiaokun Ding [1,2], Xuanyu Zhang [1], Shangzhen Song [2], Bo Li [1], Le Hui [1] and Yuchao Dai [1,*]

[1] School of Electronics and Information & Shaanxi Key Laboratory of Information Acquisition and Processing, Northwestern Polytechnical University, Xi'an 710072, China; xuanyuzhang@mail.nwpu.edu.cn (X.Z.)
[2] Xi'an Flight Automatic Control Research Institute, Xi'an 710076, China
* Correspondence: daiyuchao@nwpu.edu.cn

Highlights

What are the main findings?

- We propose a pipeline designed to enhance cross-view geo-localization (CVGL) by integrating novel view synthesis. The core of our framework reduces the cross-view feature discrepancy through the generation of perspective-aware overhead images, leading to superior geo-localization accuracy.
- A novel camera pose generation method is specifically designed for autonomous driving scenarios to address the challenge of missing vertical view pose.

What are the implications of the main findings?

- The proposed method establishes a continuous feature transition between street-level and satellite imagery, thereby enhancing the model's capability in cross-view geo-localization tasks.
- By integrating 3D Gaussian Splatting (3DGS)-based novel view synthesis into deep learning frameworks for CVGL, our approach enables the autonomous generation of corresponding bird's-eye-view images directly from street-view inputs.

Abstract

Cross-view geo-localization allows an agent to determine its own position by retrieving the same scene from images taken from dramatically different perspectives. However, image matching and retrieval face significant challenges due to substantial viewpoint differences, unknown orientations, and considerable geometric distribution disparities between cross-view images. To this end, we propose a cross-view geo-localization framework based on novel view synthesis that generates pseudo aerial-view images from given street-view scenes to reduce the view discrepancies, thereby improving the performance of cross-view geo-localization. Specifically, we first employ 3D Gaussian splatting to generate new aerial images from the street-view image sequence, where COLMAP is used to obtain initial camera poses and sparse point clouds. To identify optimal matching viewpoints from reconstructed 3D scenes, we design an effective camera pose estimation strategy. By increasing the tilt angle between the photographic axis and the horizontal plane, the geometric consistency between the newly generated aerial images and the real ones can be improved. After that, the DINOv2 is employed to design a simple yet efficient mixed feature enhancement module, followed by the InfoNCE loss for cross-view geo-localization. Experimental results on the KITTI dataset demonstrate that our approach can significantly improve cross-view matching accuracy under large viewpoint disparities and achieve state-of-the-art localization performance.

check for updates

Academic Editor: Lefei Zhang

Received: 14 September 2025
Revised: 27 October 2025
Accepted: 4 November 2025
Published: 8 November 2025

Citation: Ding, X.; Zhang, X.; Song, S.; Li, B.; Hui, L.; Dai, Y. Cross-View Geo-Localization via 3D Gaussian Splatting-Based Novel View Synthesis. *Remote Sens.* **2025**, *17*, 3673. https://doi.org/10.3390/rs17223673

Keywords: cross-view geo-localization; novel view synthesis; 3D gaussian splatting; contrastive learning

1. Introduction

Vision-based localization is a fundamental technology in many intelligent systems, including autonomous driving [1], augmented reality [2], and mobile robotics [3]. However, traditional localization approaches that rely on ground-level image databases suffer from several inherent limitations. Firstly, ground images often provide limited coverage, making it difficult to meet the requirements of large-scale environments. Secondly, such datasets depend heavily on costly manual GPS annotation [4,5], which struggle to handle cross-view or multi-scale variations. In addition, ground-level images are sensitive to environmental change, including illumination, weather, and season, resulting in poor robustness. More critically, these methods are effective only when operating within similar view points, whose imaging perspectives are not significantly different. In some extreme matching applications between ground-level (street) views and aerial-level (satellite or UAV) views, conventional approaches struggle to establish reliable correspondences. To overcome this problem, cross-view geo-localization has recently become a promising research direction [6]. This task leverages satellite images as reference, which inherently provides wide spatial coverage with precise GPS labels. By establishing correspondences between ground-level and aerial-level images, it becomes possible to overcome the challenges of viewpoint variations, thereby enabling a robust localization even in GPS-denied environments. Nevertheless, cross-view geo-localization remains highly challenging due to the significant domain gap between aerial and ground perspectives. Early studies primarily relied on handcrafted features such as SIFT [7], SURF [8], and ORB [9] descriptors. These approaches attempted to exploit invariances in local gradients or texture patterns to build correspondences between images. Such handcrafted features are simple but inherently limited, as they are unable to capture semantic-level similarities under large viewpoint and scale variations, leading to significant performance degradation.

With the rapid development of deep learning, researchers have shifted their focus to learning deep feature representations [10–12]. CNN-based methods became dominant owing to their ability to provide robust representations invariant to photometric distortions. Workman et al. [6] first explored CNN-based features for cross-view matching based on the assumption that high-level semantic features from pre-trained networks encode geographical information. Subsequent studies incorporated more advanced backbones such as VGG [13], ResNet [14], and DenseNet [15–17], significantly improving retrieval performance. In addition to architectural improvements, metric learning played a pivotal role in forming discriminative embeddings through losses such as soft-margin triplet loss [18] and global descriptor networks like NetVLAD. Orientation-aware strategies were also developed, including coordinate embedding [19] and geometric transformation techniques [20] to reduce perspective distortions. More recently, Transformer-based architectures have gained attention for their ability to model long-range dependencies and global context, with several studies demonstrating their strong potential in capturing complex cross-view relationships. These approaches have shown particular effectiveness in handling large perspective changes and complex urban scenarios through self-attention mechanisms and adaptive feature learning.

Despite considerable progress, significant challenges also persist. The most critical issue stems from the severe perspective discrepancy between ground and aerial images: ground-level photos typically offer horizontal, often occluded views, while aerial images

provide vertical, unobstructed overviews. This discrepancy is further exacerbated when using standard non-panoramic cameras in practical settings, where limited field of view results in substantial information loss, making consistent geometric correspondence difficult to establish. Another major challenge arises from the pronounced domain shift in imaging conditions, including variations in illumination, scale, and resolution. Ground images are often affected by dynamic lighting changes, while aerial views are generally captured under uniform natural light but at drastically different scales and typically with coarser spatial resolution. These intrinsic differences hinder the learning of invariant feature. Furthermore, semantic ambiguity poses a significant obstacle, particularly in areas with high visual self-similarity such as highways, farmlands, or repetitive urban structures. While such regions appear homogeneous from an aerial perspective, they exhibit distinct characteristics at ground level, leading to mismatches in feature space. These challenges highlight the need for innovative frameworks capable of bridging the large domain gap through both geometric reasoning and discriminative representation learning.

In recent years, generative models such as Generative Adversarial Networks (GANs) [21] and Neural Radiance Fields (NeRFs) [22] have opened new pathways for 3D scene reconstruction and localization through view synthesis. These methods aim to reconstruct realistic 3D scenes from multiple 2D images captured from different viewpoints, offering promising strategies for mitigating the domain gap in cross-view matching tasks. By learning implicit representations of geometry and appearance, they enhance feature consistency across highly divergent perspectives. Recent advances in 3D Gaussian Splatting (3DGS) [23] have demonstrated remarkable efficiency in high-fidelity novel view synthesis, achieving real-time rendering performance(often exceeding 100 frames per second) while preserving geometric detail. In contrast to NeRF's implicit volumetric representation, 3DGS explicitly models a scene using millions of Gaussian ellipsoids, which enables direct manipulation (e.g., moving, deleting, or modifying elements), greatly improving editability. Although 3DGS typically requires more storage due to a larger number of parameters, its computational efficiency during rendering leads to better overall memory utility compared to NeRF-based methods. Furthermore, 3DGS shows superior adaptability to dynamic scenes. While native NeRF is generally confined to static settings and requires non-trivial extensions to model motion, 3DGS can natively handle dynamics through mechanisms such as deformation fields. In terms of initialization, NeRF usually demands hundreds of input views for stable convergence, whereas 3DGS can start from a sparse point cloud generated via Structure-from-Motion (SfM), reducing data requirements and broadening applicability. These properties make 3DGS particularly suitable for applications requiring real-time interaction and high render efficiency, such as virtual reality, augmented reality, and interactive scene editing.

In this paper, we propose a novel cross-view geo-localization framework based on 3D Gaussian splatting, which synthesizes novel images of large tilt angles from street-view inputs to effectively bridge the domain gap between ground and aerial views. Specifically, for initializing the 3D Gaussians, we employ COLMAP [24] to estimate accurate camera poses and sparse point clouds from street-level imagery. Furthermore, we design a dedicated camera pose estimation strategy that determines optimal aerial viewpoints by progressively increasing the tilt angle of the photographic axis relative to the horizontal plane. This approach enhances the geometric consistency and realism of the synthesized aerial images. After synthesizing the novel aerial views, we process them through a mixed feature enhancement module that leverages both DINOv2 [25] and a feature-mixer network to extract discriminative and robust representations. These features are then used to perform cross-view matching. The entire framework is trained end-to-end using the InfoNCE loss, which facilitates effective learning of viewpoint-invariant features. Experimental

results demonstrate that compared to purely Transformer-based methods such as TransGeo and other advanced network backbones, the novel view images synthesized by 3DGS in our method significantly enhance retrieval accuracy by effectively reducing the domain gap. Furthermore, the mixed feature enhancement network based on DINOv2 employed in our framework exhibits stronger feature retrieval capabilities, further improving retrieval performance. Extensive experiments on the KITTI dataset [26] demonstrate that our method effectively overcoming large perspective disparities and achieving state-of-the-art retrieval performance.

In summary, our contributions are as follows:

- We introduce a novel cross-view geo-localization framework based on 3D Gaussian splatting, which synthesizes highly realistic aerial-view images from ground-level inputs. This approach explicitly mitigates severe perspective and domain gaps between the two view images by generating geometrically consistent intermediate viewpoints.
- We design a dedicated camera pose estimation strategy that progressively optimizes virtual aerial viewpoints by increasing the tilt angle of the camera axis. This method ensures high-fidelity view synthesis within 3D Gaussian-reconstructed scenes. Furthermore, we integrate DINOv2 as a robust feature extraction backbone to capture more discriminative representations, enhancing the performance of cross-view matching.
- Experiments demonstrate that our method significantly improves cross-view matching and localization accuracy, particularly under large perspective changes and challenging urban scenarios

2. Related Works

2.1. Cross-View Geo-Localization

Early studies in cross-view geo-localization [27–29] primarily relied on handcrafted feature descriptors such as self-similarity patterns and color histograms. Although intuitive, these manually designed features exhibited limited discriminative power due to their sensitivity to illumination changes, scale variations, and significant viewpoint shifts, thereby restricting their practical applicability. Driven by the progress in deep learning, apart from their first exploration of CNN-based feature extraction for cross-view matching, subsequent work of Workman et al. [30] fine-tuned networks using contrastive losses to minimize feature distances between cross-view image pairs. They also established the CVUSA dataset, which has become one of the most widely used datasets in this field. Inspired by advances in face recognition, Lin et al. [31] employed a siamese architecture optimized with contrastive loss [32–35], while Zhai et al. [36] integrated NetVLAD modules [37] to enhance robustness against viewpoint variations.

Another significant research direction focuses on metric learning, which aims to devise specialized objective functions that promote discriminative feature embedding. Vo et al. [18] introduced a soft-margin triplet loss as a standard training objective, improving generalization through better geometric adaptation. Hu et al. [38] further embedded NetVLAD layers into the backbone network to generate highly compact global descriptors. To address the problem of slow convergence, they proposed a weighted soft-margin ranking loss that adaptively scales distances between positive and negative pairs, thereby accelerating training and boosting retrieval precision. Despite these advances, a common limitation among these methods is their over-reliance on global feature matching, often overlooking finer-grained contextual information.

Further investigations have addressed the fundamental challenge of cross-view domain gap. Liu et al. [19] explicitly incorporated orientation awareness by embedding coordinate information into the feature learning process, significantly improving spatial discriminability. Shi et al. [20] proposed a polar transformation technique to align the

spatial layout of remote sensing images with street-view perspectives. While effective under ideal conditions, this geometric prior is sensitive to misalignment in image centers and may introduce harmful distortions that degrade localization accuracy.

In recent years, Vision Transformer (ViT)-based frameworks have gained prominence in cross-view geo-localization by leveraging their superior capability in capturing long-range dependencies and global contextual relationships compared to CNNs. These approaches have demonstrated state-of-the-art performance in matching ground and aerial images under significant viewpoint changes. He et al. [39] introduced a multi-view scene matching framework based on a dual-attention Vision Transformer, which enhances global feature modeling and strengthens contextual correlations between adjacent regions. To address sample imbalance between ground and aerial images, a contrastive loss function was incorporated to improve learning efficiency and feature alignment. Pillai et al. [40] proposed a GeoAdapter module capable of aggregating image-level representations and adapting them for video-sequence inputs. They also designed a TransRetriever architecture to resolve temporal inconsistencies in trajectory data by predicting per-frame GPS coordinates, thereby supporting robust video-based cross-view localization. Zhu et al. [41] developed TransGeo, a pure ViT-based framework that eliminates conventional pre-processing steps such as polar transformation and data augmentation. The model employs adaptive sharpness-aware minimization (ASAM) [42] to optimize the sharpness of the loss landscape, effectively mitigating overfitting and improving generalization. Furthermore, TransGeo incorporates an attention-guided non-uniform cropping strategy that selectively removes occluded regions in satellite imagery—which contribute minimally to street-view matching—while increasing resolution in semantically salient areas.

2.2. Novel View Synthesis

Novel view synthesis (NVS), the task of generating photorealistic images of a scene from arbitrary viewpoints, has become valuable in applications such as cross-view geo-localization [43,44]. Driven by neural rendering [45], two paradigms have become particularly dominant: NeRF and 3D Gaussian Splatting (3DGS). This section reviews seminal and representative works from both categories that form the foundation of modern real-time, high-fidelity view synthesis.

NeRF pioneered a new approach by representing a static scene as a continuous implicit function encoded by a multi-layer perceptron (MLP). This function maps 5D coordinates (3D location and 2D viewing direction) to volume density and view-dependent radiance. Images are rendered by querying this MLP along camera rays and integrating colors and densities using classical volume rendering. While the original NeRF achieved state-of-the-art quality, its slow training and rendering speeds motivated extensive subsequent research. To address aliasing and improve detail rendering at various scales, Barron et al. introduced Mip-NeRF [46], which models the volume of a conical frustum rather than an infinitesimal ray. Concurrently, significant efforts targeted computational bottlenecks: Plenoxels [47] replaced the MLP with an explicit voxel grid parametrized by spherical harmonics coefficients, drastically reducing training time. Building on this, Instant-NGP [48] introduced multi-resolution hash encodings, enabling high-quality NeRF training in seconds to minutes. D-NeRF [49] incorporated a deformation network and latent appearance codes to model dynamic scenes from a single canonical representation. These advancements collectively established NeRF as a powerful and versatile, albeit computationally intensive, method for high-fidelity view synthesis.

While NeRF-based methods excel in quality, their reliance on dense sampling of an implicit function often hinders real-time rendering. 3DGS emerged as a transformative approach. It employs an explicit scene representation composed of millions of 3D

Gaussians that are rendered in real-time using a differentiable tile-based rasterizer. By combining the explicit nature of point-based rendering with a volumetric interpretation and a highly optimized GPU pipeline, 3DGS achieves state-of-the-art visual quality at real-time speeds. The interpretable nature of Gaussian primitives has facilitated several significant extensions. Four-dimensional Gaussian Splatting [50] models temporal evolution using compact decompositions into temporal basis functions for dynamic scenes; relightable three-dimensional Gaussians [51] decompose appearance into material properties for realistic relighting under novel illumination. Furthermore, there are other improvements of 3DGS including high-quality surface extraction, few-view synthesis, and memory-efficient deployment. In summary, 3DGS has undergone rapid development and is gaining significant momentum in the field of 3D reconstruction.

In this work, we propose a novel method that addresses the domain gap in cross-view geo-localization by leveraging 3DGS. Street-view scenes are reconstructed with high fidelity, enabling the rapid synthesis of pseudo images from larger tilt angles. These synthesized views serve as crucial data augmentation, providing our deep retrieval network with a richer, more spatially aware dataset. By training on this augmented dataset, the model learns features that effectively bridge the visual and geometric gap between the two view pairs, thereby significantly enhancing accuracy and robustness in cross-view retrieval—a capability beyond what conventional datasets alone can provide.

3. Method

As illustrated in Figure 1, the paper presents the overall architecture of our proposed pipeline based on 3D Gaussian splatting. We first employ 3DGS to synthesize corresponding pseudo aerial-view images from a sequence of street-view images as input, where a central contribution is an effective camera pose generation strategy that actively identifies optimal aerial viewpoints to maximize geometric and semantic alignment with the original street-view scenes. Subsequently, we introduce a mixed feature enhancement module designed to extract highly discriminative features from both street-view and synthesized aerial-view images. This module integrates multi-scale contextual cues to enhance representation learning, thereby improving robustness against cross-domain discrepancies. For network training, a supervised contrastive learning framework is adopted using the InfoNCE loss, which effectively leverages all available negative samples within batches to promote enhanced feature separation and clustering across views. This strategy significantly improves the model's capability to accurately match street-view queries with their corresponding geo-referenced aerial images under challenging conditions.

3.1. Preliminaries on 3D Gaussian Splatting

Unlike methods based on NeRF that rely on implicit representations, 3DGS models a scene as a collection of explicit, point-based 3D Gaussians. This approach offers a compelling trade-off between rendering quality and computational efficiency. Each 3D Gaussian is defined by a set of trainable parameters that are optimized to accurately represent the scene geometry and appearance.

Each individual 3D Gaussian is parameterized by the following attributes:

Position (μ): The mean vector $\boldsymbol{\mu}$ specifies the centroid of the Gaussian in 3D world coordinates. These positions are initialized directly from a sparse point cloud, which is typically reconstructed using a Structure-from-Motion (SfM) algorithm like COLMAP.

Covariance Matrix (Σ): The covariance matrix $\sum$ defines the shape, size, and orientation of the Gaussian ellipsoid. To ensure $\sum$ is a positive semi-definite matrix, it is parameterized by a scaling matrix $\boldsymbol{S}$ and a rotation matrix $\boldsymbol{R}$, which correspond to the ellip-

soid's axes lengths and orientation, respectively. This decomposition ensures the validity of the covariance matrix during optimization. The relationship is expressed as

$$\boldsymbol{\Sigma} = \boldsymbol{RSS}^{\mathsf{T}}\boldsymbol{R}^{\mathsf{T}}, \tag{1}$$

where $\boldsymbol{S} = diag(\boldsymbol{s}_x, \boldsymbol{s}_y, \boldsymbol{s}_z)$ is a diagonal matrix of scaling factors. The initial scaling parameters are adaptively determined based on the local density of the SfM point cloud, ensuring that denser regions are initialized with smaller Gaussian radii to preserve fine details.

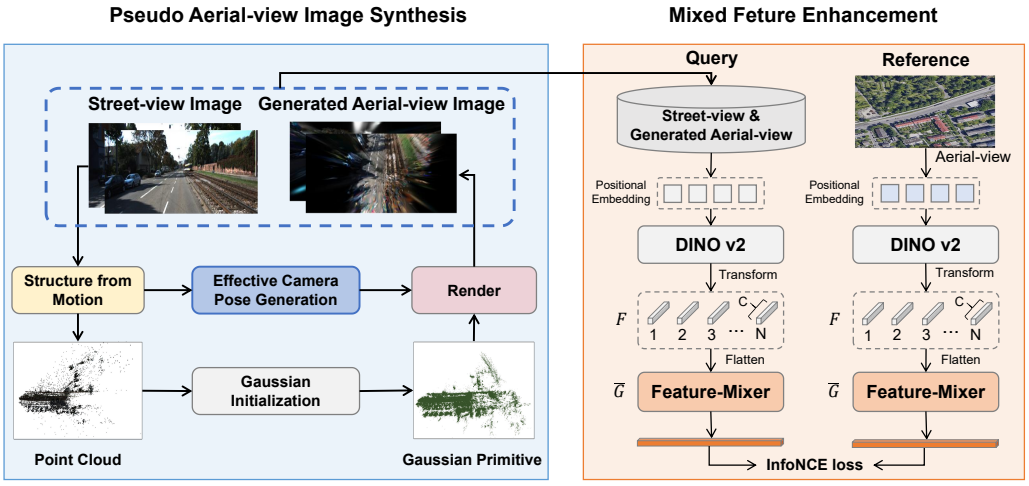

Figure 1. The architecture of the proposed 3D Gaussian splatting-based cross-view geo-localization framework. We first propose a pseudo aerial-view image synthesis module, which leverages the 3D Gaussian splatting combined with an effective camera pose generation strategy to render new-view images for data augmentation. Then, we propose a mixed feature enhancement module to obtain discriminative features for retrieval.

Color (c): The view-dependent color of each Gaussian is represented using a set of Spherical Harmonics (SH) coefficients. This representation allows for the efficient encoding of directional lighting and reflections. The SH coefficients are initialized by averaging the multi-view colors of the points from which the Gaussians are initialized. Specifically, the initial coefficients are computed as

$$\text{SH coefficients} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{I}_i(\theta, \phi), \tag{2}$$

where $\boldsymbol{I}_i(\theta, \phi)$ denotes the color intensity at spherical coordinates (θ, ϕ) for the i-th of the N input views.

Opacity (α): A scalar value ranging from 0 to 1, opacity determines the transparency of each Gaussian. It is initialized to 0.5 and subsequently optimized through a gradient-based approach to control the blending of Gaussians during rendering.

To render a novel view, the 3D Gaussians must be projected onto a 2D image plane. This projection is a non-linear process, but 3DGS leverages a local affine approximation to make it computationally efficient. This is achieved by performing a second-order Taylor expansion of the perspective projection transformation centered at the centroid of Gaussians. The resulting projection of the 3D Gaussian covariance matrix $\boldsymbol{\Sigma}$ into the 2D screen space, denoted as $\boldsymbol{\Sigma}'$, is computed as

$$\boldsymbol{\Sigma}' = \boldsymbol{JW\Sigma W}^{\mathsf{T}}\boldsymbol{J}^{\mathsf{T}}, \tag{3}$$

where $\boldsymbol{W}$ is the view transformation matrix and $\boldsymbol{J}$ is the Jacobian matrix representing the local linear approximation of the projection. This formulation ensures that the projected 3D Gaussians remain as 2D ellipses, which can be efficiently rasterized.

The final rendering process is performed using a tile-based rasterization pipeline. The 2D screen space is partitioned into pixel tiles to manage computational complexity. A filtering step, including frustum culling and bounding box tests, is first applied to retain only the Gaussians that fall within the current view frustum and can influence a given tile. Within each tile, the Gaussians are depth-sorted from front-to-back and blended using an alpha compositing formula to compute the final pixel color:

$$C = \sum_i \boldsymbol{c}_i \boldsymbol{\alpha}_i \prod_{j=1}^{i-1} (1 - \boldsymbol{\alpha}_j). \tag{4}$$

where $\boldsymbol{c}_i$ and $\boldsymbol{\alpha}_i$ represent the color and opacity of the i-th Gaussian, respectively. This rasterization-based approach bypasses the need for repeated, per-ray MLP inferences characteristic of NeRF, which significantly accelerates the rendering process and enables real-time frame rates.

3.2. Pseudo Aerial-View Image Synthesis

To effectively mitigate the perspective gap between street-view and aerial-view images, we design a synthesis strategy of pseudo aerial-view images directly from ground-level inputs. 3DGS combines explicit 3D representation with differentiable rendering. Its principle involves optimizing millions of parameterized 3D Gaussian ellipsoids to achieve high-quality real-time view synthesis. Capitalizing on these capabilities, we introduce an aerial-view synthesis module based on 3DGS to generate highly realistic pseudo-aerial images from street-view sequences. To the best of our knowledge, this work pioneers the use of 3D Gaussian Splatting (3DGS) to synthesize a pseudo aerial image from a given street-view image, effectively bridging the disparity between these two perspectives. The overall architecture of the proposed system is depicted in Figure 1.

Specifically, our approach utilizes a pre-trained 3DGS model to synthesizes multi-perspective images with larger tilt angles using a trained model. By processing sequences of street-view images, the model produces a variety of synthetic aerial perspectives that collectively form a enriched cross-view dataset, thereby enhancing feature alignment and improving match accuracy. We first employ COLMAP to estimate corresponding camera poses and reconstruct a sparse point cloud. This point cloud subsequently serves as the initial set of 3D Gaussian primitives, each defined by properties such as position, anisotropic covariance, opacity, and view-dependent color represented via spherical harmonics. These primitives undergo iterative optimization alongside the input images using gradient descent, minimizing a reconstruction loss that compares rendered against actual views. For novel view rendering, each Gaussian is projected into 2D screen space through a differentiable splatting process, followed by alpha-blending of overlapping points based on depth ordering. The entire pipeline supports rendering from arbitrary viewpoints, allowing flexible generation of pseudo-aerial images that exhibit high geometric and photometric consistency with the original street-level scene.

However, in practice, acquiring ideal 360° video streams to extract sufficient continuous images for training is often infeasible. This limitation makes it difficult for 3DGS to directly synthesize the required images. To overcome this problem, we introduce a two-step method to compute an optimal camera pose for a given tilt angle: (1) determining the 3D coordinates of the camera's optical center and (2) computing the corresponding rotation matrix.

The process of estimating the camera's optical center position involves finding both the average point cloud center and the average plane normal vector. The strategy for finding the average point cloud center leverages the prior knowledge that the captured altitude of street-view images remains relatively consistent. We hypothesize that the optimal camera viewpoint should position its optical center collinear with the centroid of the average point cloud, thereby maximizing scene coverage and information acquisition. The average point cloud center is calculated as follows:

$$\bar{\boldsymbol{t}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{t}_i, \tag{5}$$

where $\boldsymbol{t}_i = (x, y, z)^\top$ refers to the 3D coordinates of the i-th point of all n points.

After computing the point cloud's average center, we proceed to estimate the camera pose. The normal vector of a plane is determined using three key points: the average center of the point cloud together with the optical centers from any two images. By iterating this procedure across multiple image pairs, we obtain a set of candidate normal vectors. Principal Component Analysis (PCA) is then applied to filter these candidates, yielding a robust average normal vector that guides the orientation adjustment of the camera's optical center. The process is illustrated in Figure 2, and corresponding formulation is written as

$$\mathcal{N} = \frac{1}{n-1}\sum_{i=2}^{n}[(\boldsymbol{P}_i - \bar{\boldsymbol{t}}) \times (\boldsymbol{P}_{i-1} - \bar{\boldsymbol{t}})], \tag{6}$$

where $\boldsymbol{P}_i$ denotes the optical center coordinates of the i-th image, n is the total number of images captured by the platform, and $\times$ indicates the cross product of two vectors. It is important to note that reversing the order of the vectors in the cross product yields a normal vector in the opposite direction. To ensure consistency, our method computes each pairwise product only once while retaining a consistent vector order throughout the process. The computed average plane normal vector $\mathcal{N}$ plays a critical role in aligning the camera's optical axis. By incorporating this geometric prior, we optimally adjust the orientation of the camera's optical center for subsequent rendering steps. This adjusted pose, combined with the 3D Gaussian representations and original camera pose data, enables the synthesis of high-quality 2D images from strategically chosen aerial viewpoints. Ultimately, this pipeline allows us to generate highly realistic pseudo aerial-view images directly from the input street-view sequences, effectively bridging the perspective gap between ground and aerial imagery.

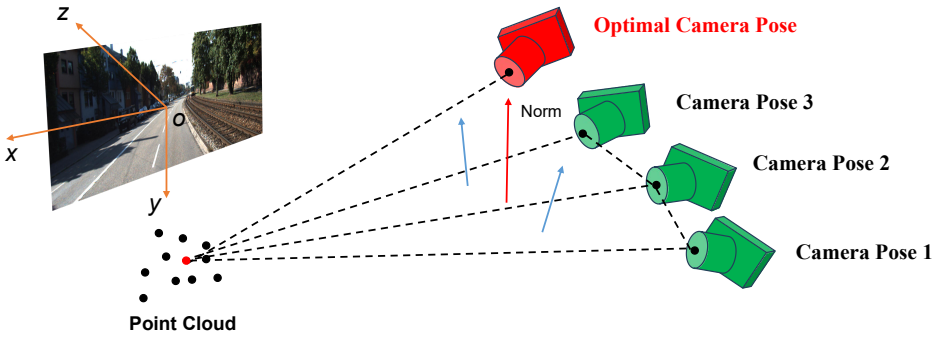

Figure 2. The process of solving the average plane normal vector and the optimal camera pose.

3.3. Mixed Features Enhancement

In recent years, foundational vision models have significantly advanced the field of computer vision by employing deep architectures such as CNNs and Transformers, scaled to hundreds of millions of parameters. Trained on large and diverse datasets, these models

exhibit superior representational power and generalization capability. Among these, DINOv2 [25] a self-supervised visual model, learns general-purpose visual representations directly from unlabeled images, overcoming limitations of supervised pre-training. According to its advantages, we leverage DINOv2 as a feature extraction backbone to construct a Mixed Feature Enhancement module for learning discriminative representations for cross-view retrieval. Note that consistent with many other studies, we employed the pre-trained weights of DINOv2 without further fine-tuning. As illustrated in Figure 1, the module first processes both query (street-view or pseudo aerial-view) and reference (aerial) images using DINOv2 to extract visual features. A feature-mixing mechanism is then applied to capture global contextual relationships through cascaded transformations.

Specifically, we combine original street-view and generated pseudo aerial-images into a unified query set, with original aerial imagery serving as the reference. The goal is to identify optimal matches where the query and reference images are accurately aligned. The process begins by dividing an input image into patches and projecting them into patch-level embeddings. These are fed into DINOv2 to produce pre-trained features, taken from the last ViT block while excluding the classification head.

Let the pretrained feature map be donated as $\boldsymbol{F} \in \mathbb{R}^{N \times C}$, where N is the number of channels and C is the feature dimension. This can be interpreted as a set of one-dimensional representations, denoted as

$$\boldsymbol{F} = \{\boldsymbol{X}^i\}, i = \{1, \cdots, N\}. \tag{7}$$

The feature-mixer consists of L successive MLP blocks, which is illustrated in Figure 3. Each block refines the features by incorporating global interactions through a residual transformation:

$$\boldsymbol{X}^i \leftarrow \boldsymbol{W}_2\Big(\sigma\Big(\boldsymbol{W}_1\boldsymbol{X}^i\Big)\Big) + \boldsymbol{X}^i, i = \{1, \cdots, N\}, \tag{8}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are learnable weights of fully connected layer. σ denotes the ReLU activation function. It is desired that the feature-mixer can leverage the capacity of fully connected layers for holistic feature aggregation. After processing through all L feature-mixer layers, the final output feature $\boldsymbol{G}$ is given by

$$\boldsymbol{G} = FM_L(FM_{L-1}(\ldots FM_1(\boldsymbol{F}))). \tag{9}$$

Note that $\boldsymbol{G}$ has the same size $N \times C$ as $\boldsymbol{F}$. To further reduce the dimension, we adopt a depth-wise projection that maps G from $\mathbb{R}^{N \times C}$ to $\mathbb{R}^{N \times D}$ such as

$$\hat{\boldsymbol{G}} = \boldsymbol{W}_D(Transpose(\boldsymbol{G})), \tag{10}$$

where $\boldsymbol{W}_D$ is the weight of the fully connected layer. Similarly, a row-wise projection that maps $\hat{\boldsymbol{G}}$ from $\mathbb{R}^{N \times D}$ to $\mathbb{R}^{n \times D}$ is applied, such as

$$\bar{\boldsymbol{G}} = \boldsymbol{W}_n\big(Transpose\big(\hat{\boldsymbol{G}}\big)\big). \tag{11}$$

Finally, the feature is L_2-normalized to produce the global descriptor for retrieval.

To optimize our model, we adopt the InfoNCE [52] losswe, following the practice [53]. The loss function is defined as

$$\mathcal{L}_{InfoNCE} = -\log \frac{exp(q \cdot r_+/\tau)}{\sum_{i=0}^{N_R} exp(q \cdot r_i/\tau)}, \tag{12}$$

where q denotes the query image, which is the street-view image in the method. r_+ and $\{r_i\}_{i=1}^{N_R}$ indicate the reference images. r_+ is the positive sample, while r_i is the negative

sample. Note that for each query q, there is exactly on positive sample r_+. The InfoNCE loss measures similarity via dot products in the latent space, causing the objective to decrease when the query closely aligns with its positive counterpart and increase when it is similar to negative samples.

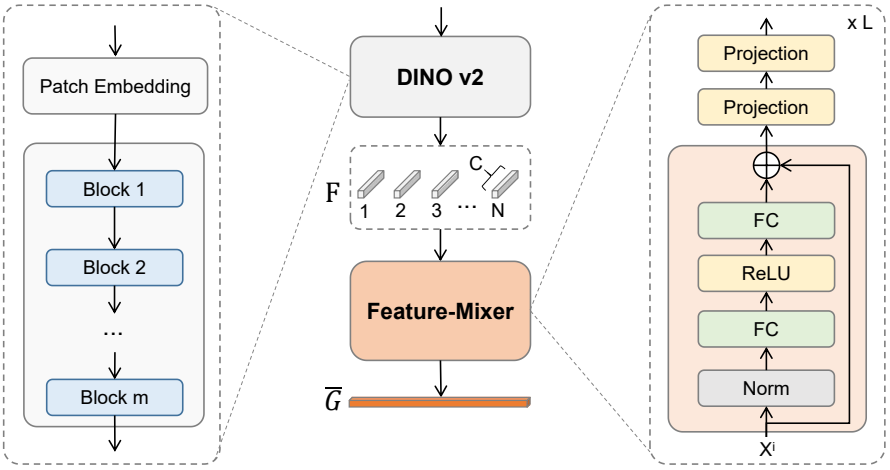

Figure 3. The detailed architecture of the mixed feature enhancement module.

4. Results

4.1. Dataset and Setting

Dataset. The proposed framework is rigorously evaluated using the KITTI dataset [26] and the Oxford Robot Car dataset [54] KITTI is a benchmark widely recognized in the field of autonomous driving and computer vision. The dataset comprises a diverse collection of image sequences captured by a moving vehicle, encompassing a variety of urban and suburban environments under different lighting and weather conditions. For the purpose of our study, the dataset is partitioned into distinct training and testing subsets to ensure a robust and unbiased evaluation. Specifically, the training set is meticulously curated by selecting specific trajectories from KITTI, adhering to the inherent serialization of the data and the requirements of feature-based localization methods like COLMAP [24]. To further enhance the complexity and generalizability of the evaluation, we adopt the extended KITTI dataset constructed by Shi [55]. This augmented dataset incorporates corresponding satellite imagery, thereby enabling a comprehensive cross-view analysis. The dataset is organized into two distinct test sets, namely Test-1 and Test-2. Notably, Test-2 is specifically designed to assess the model's robustness to domain shift, as its image sequences are deliberately chosen from scenes with significantly different visual characteristics compared to the training set.

Except for the KITTI dataset, we additionally conducted experiments on other datasets. Because the sparse nature of the panoramic image sequences in CVUSA, VIGOR, and TorontoCity dataset fails to provide a good initialization for 3DGS in large-scale scenes, which leads to degraded reconstruction quality. The Oxford Robot Car dataset is originally designed for autonomous driving research, which shares a similar data format with KITTI. It contains a total of 23,854 valid ground-to-satellite image pairs, divided into 17,067 pairs for training, 1698 for validation, and 5089 for testing. The street-view images cover a variety of illumination and weather conditions—including sunny, overcast, and cloudy scenes—across both summer and winter seasons.From this perspective, we apply the dataset to further verify our method.

The satellite images are sourced from Google Maps [56], providing a per-pixel ground resolution of 0.20 m. To facilitate effective training and inference, a large geographical area covering the vehicle trajectories is first identified. This region is then uniformly

subdivided into a grid of overlapping satellite image blocks, each with a spatial resolution of 1280 × 1280 pixels. This block-based approach ensures that the model can handle large-scale geographical areas efficiently while maintaining sufficient contextual information for effective localization.

Evaluation Metrics. The performance of our proposed method is quantitatively assessed using the widely adopted recall at top k (Rk) metric, a standard practice in the cross-view geo-localization literature [57]. This metric quantifies the retrieval accuracy by measuring the proportion of query images for which the correct corresponding reference image is found within the top k retrieved candidates. Specifically, for each query image, we compute its cosine similarity with all reference images in the embedded feature space. The top k nearest neighbors are then retrieved. A retrieval is considered successful if the ground truth reference image is present within this set of top k candidates. This metric provides a clear and intuitive measure of the model's ability to localize a street-view image within a large-scale satellite map.

Experimental Settings. All experiments were conducted on a high-performance computing platform equipped with an NVIDIA GeForce RTX 3080 Ti GPU and an Intel Xeon CPU operating at 2.40 GHz. Our framework is implemented in PyTorch 1.12, a widely used deep learning framework, enabling efficient training and deployment. The input resolutions for the satellite and street-view images were resized to 640 × 640 and 512 × 128 pixels, respectively, balancing computational efficiency with information preservation. The training process was configured with a batch size of 16, an initial learning rate of 1e-4, and a total of 100 epochs. The AdamW optimizer [58] was employed for its superior performance in weight decay regularization. The majority of the hyperparameters were set to align with established practices in the field to ensure a fair comparison with prior works. The final embedding dimension for both views was set to 1000. This is a deliberate choice, as it results in a significantly more compact and efficient representation compared to many conventional CNN-based methods, which often rely on high-dimensional feature vectors. For 3D Gaussian initialization, regarding the choice of initial scale based on the local density of the SfM point cloud, this strategy was adopted to adaptively determine the initial size of each Gaussian primitive according to the spatial distribution of the reconstructed 3D points. In dense regions, a smaller initial scale helps preserve finer details, while in sparser areas, a larger scale offers better coverage and stability in optimization. This density-aware initialization leads to more stable training and higher-quality reconstruction compared to using a uniform initial scale. For the COLMAP settings, we follow prior works on large-scale outdoor reconstruction. Specifically, the maximum number of features is set to 8192, "Sequential" is used as the matching strategy, and the minimum number of matches is set to 15.

In the experiment, we also tested the retrieval results with varying quantities of synthesized viewpoint images used as training data, specifically 50, 100, 150, and 200 novel images. The results indicated that using 100 synthesized images yielded better retrieval performance compared to 50. When the quantity was increased to 150 and 200, no further significant improvement in retrieval results was observed. Instead, the computational cost increased. Therefore, considering computational efficiency, 100 images were selected as the optimal quantity.

4.2. Performance Comparison

To rigorously validate the efficacy of our proposed framework, we conducted a comprehensive performance evaluation against several mainstream methods on our designated test datasets. The quantitative results, summarized in Table 1, unequivocally demonstrate the superior performance of our approach.

Table 1. Comparison with different methods on single image based localization. The best results are highlighted in bold.

Method	Test-1				Test-2			
	R@1	**R@5**	**R@10**	**R@1%**	**R@1**	**R@5**	**R@10**	**R@1%**
CVM-NET [38]	6.43	20.74	32.47	84.07	1.01	4.33	7.52	32.88
CVFT [20]	1.78	7.20	14.40	73.55	0.20	1.29	3.03	16.86
SAFA [59]	4.89	15.77	23.29	87.75	1.62	4.73	7.40	30.13
DSM [60]	13.18	41.16	58.67	97.17	5.38	18.12	28.63	75.70
Zhu et al. [15]	5.26	17.79	28.22	88.44	0.73	3.28	5.66	27.86
Toker et al. [61]	2.79	7.72	11.69	58.92	2.39	5.50	8.90	27.05
CVLNet [55]	17.71	44.56	62.15	98.38	9.38	24.06	34.45	85.00
TransGeo [41]	80.65	97.24	97.31	95.48	17.82	34.08	45.50	90.10
Ours	**82.90**	**98.38**	**98.43**	**98.46**	**19.20**	**38.04**	**48.90**	**91.38**

Quantitative Analysis. The results in Table 1 highlight the critical role of our proposed methodology in enhancing cross-view geo-localization performance. The performance improvement observed across all metrics is directly attributable to the introduction of synthesized high-tilt images generated via 3DGS. This novel approach effectively bridges the inherent domain gap between the query street-view images and the reference satellite-view images. The results on the Oxford Robot Car dataset are shown in Table 2. It is shown that on the second dataset, compared to the best-performing method TransGeo among the comparison methods, our approach still demonstrates superior performance. In addition, our work introduces a paradigm-shifting insight: leveraging generative models to synthesize cross-view imagery. This approach illuminates a versatile pathway for a broader spectrum of cross-view geo-localization challenges. The core methodology, which translates one data modality into the view of another, can be directly adapted to bridge other modality gaps. Therefore, it offers a foundational insight for the broader cross-view geo-localization field.

Table 2. Performance of the proposed method on the Oxford Robot Car dataset. The best results are highlighted in bold.

Method	**R@1**	**R@5**	**R@10**	**R@1%**
TransGeo [41]	70.14	87.63	92.91	94.33
Ours	**71.62**	**89.03**	**95.41**	**96.10**

By generating these intermediate, geometrically consistent views, our method transforms the complex retrieval task into a more manageable feature matching problem. The synthesized images provide a richer, multi-perspective representation of the scene, which significantly improves the similarity measurement between the query and reference domains. This strategic geometric alignment simplifies the feature learning process, leading to a substantial boost in retrieval accuracy and overall localization performance. The proposed method consistently surpasses all baseline algorithms, achieving the highest recall rates and demonstrating its superior robustness and effectiveness. While a Transformer-based architecture is employed for robust feature extraction, the primary driver for the observed performance gains is the innovative use of 3DGS to create a more favorable feature space for cross-view matching.

The synergistic interplay between 3DGS-driven augmentation, DINOv2 features, and contrastive learning forms the cornerstone of our approach, and its interpretation is key to understanding the observed performance gains. The synergy is multiplicative: 3DGS creates the geometric bridge, DINOv2 provides the semantic stability to cross it, and contrastive learning trains the model to traverse it effectively. This integrated approach

moves beyond mere data augmentation, providing an end-to-end strategy for closing the cross-view domain gap.

Qualitative Visualization. To further investigate the efficacy of our proposed query synthesis method, we present a qualitative analysis of the visual results under two typical yet challenging scenarios: a suburban scene with dense foliage and an urban environment with complex building structures. These examples are intended to provide an intuitive understanding of how our approach successfully bridges the significant domain gap between ground and aerial views.

Robustness to Foliage Occlusion. Figure 4 illustrates a representative case from a suburban area, where the street-level view is heavily occluded by trees. In such scenarios, traditional feature matching methods often fail because the discriminative ground-level features (e.g., building facades, road markings) are largely invisible in the corresponding top-down aerial image. As shown in the figure, our model synthesizes a set of pseudo aerial-views from novel pitch angles. These synthesized images effectively learn to "see through" the foliage, hallucinating the underlying geometric layout of the road and the approximate footprint of the building. The synthesized views at 45° and 60° are particularly crucial, as they create an intermediate representation that shares contextual information with both the ground-level perspective (building sides) and the aerial perspective (rooftops and layout). This ability to infer and render the essential spatial structure despite significant natural occlusion demonstrates the robustness and generalization capability of our method in cluttered real-world environments.

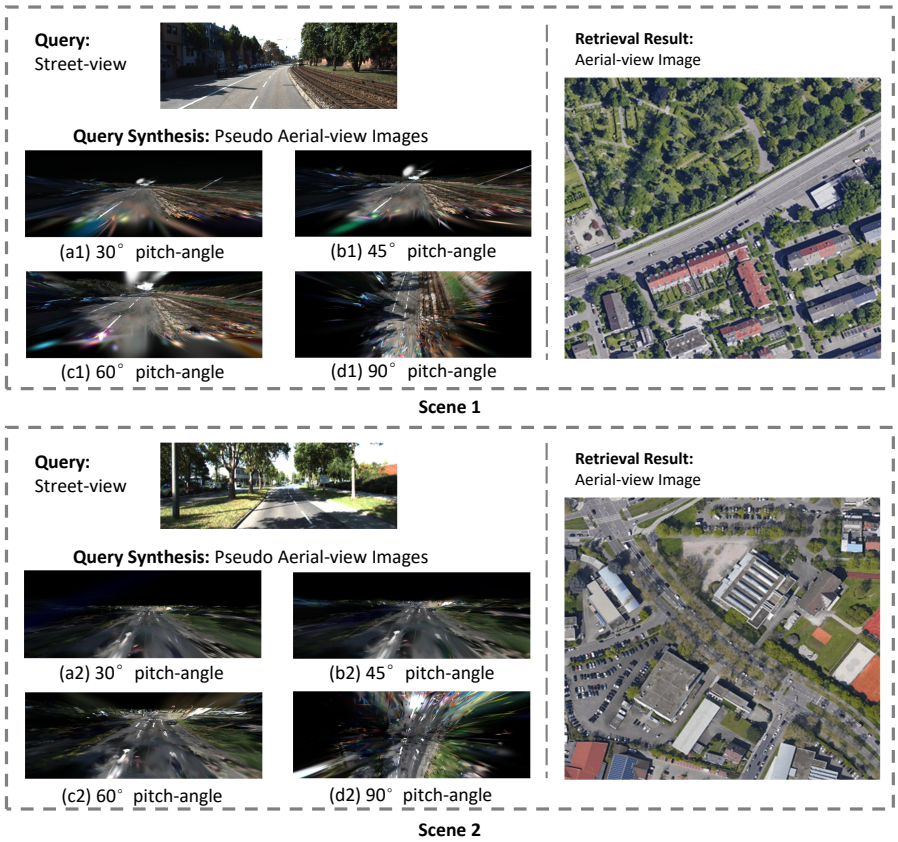

Figure 4. Visualization results for cross-view localization in challenging suburban scenes characterized by dense foliage. For a given street-view query, our method synthesizes pseudo aerial-views at varying pitch angles for matching. The images presented showcase: the input street-view query, the synthesized views at pitch angles of (**a1**,**a2**) 30°, (**b1**,**b2**) 45°, (**c1**,**c2**) 60°, and (**d1**,**d2**) 90°, and the corresponding top-retrieved true aerial image from the database.

Handling of Extreme Viewpoint Disparity. Figure 5 showcases a dense urban scene, which presents a different challenge: severe perspective distortion and complex geometric relationships between buildings. The direct matching between a ground-level image capturing vertical building facades and a nadir (90°) aerial image capturing horizontal rooftops is inherently ill-posed. Our approach mitigates this by progressively generating views that bridge this geometric transformation. The sequence of synthesized images from (a) to (d) clearly shows a smooth transition from an oblique perspective to a top-down view. This process correctly models the spatial arrangement and relative positioning of the surrounding buildings, transforming the street-view perspective into an aerial-view representation that is structurally consistent with the ground truth. The success in this complex urban setting highlights that our method does not merely rely on appearance cues but effectively learns and reasons about the underlying 3D geometry of the scene to perform accurate localization.

In summary, these qualitative results validate that our query synthesis module is the key to overcoming the challenges of cross-view localization. By generating geometrically plausible intermediate views, our method significantly enhances matching accuracy in diverse and complex real-world scenarios.

Computational costs. We report the computational cost of the proposed method on the KITTI dataset. During training, the model consumes 11.8GB of memory and takes about 4 days to complete. During testing, the model consumes 2.18GB of memory and takes about 5 days.

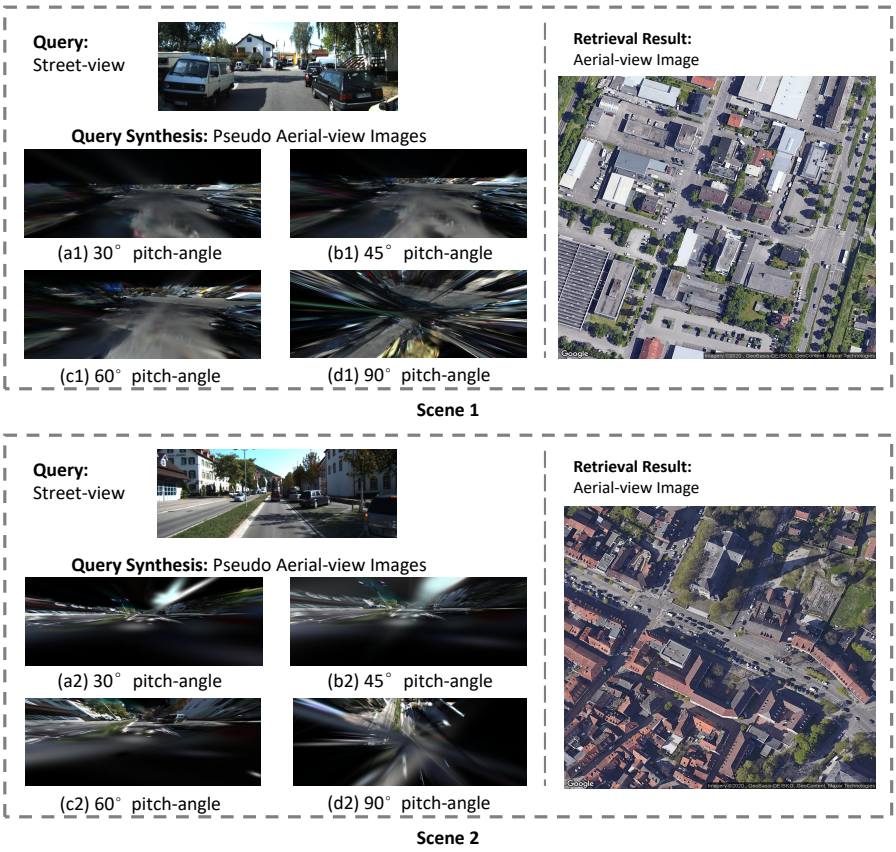

Figure 5. Visualization results for cross-view localization in challenging urban environments with dense building structures. For a given street-view query, our method synthesizes pseudo aerial-views at varying pitch angles for matching. The figure illustrates: the input street-view query, the synthesized views at pitch angles of (**a1**,**a2**) 30°, (**b1**,**b2**) 45°, (**c1**,**c2**) 60°, and (**d1**,**d2**) 90°, and the corresponding top-retrieved true aerial image from the database.

4.3. Ablation Study

To systematically validate the contribution of each key component within our proposed framework, we conducted a series of comprehensive ablation experiments. These studies were designed to quantify the performance impact and confirm the generalization capability of our method over a strong baseline. The results are meticulously analyzed below.

Impact of pseudo aerial-view image. A core component of our approach is the synthesis of pseudo-aerial-view images to enrich the training data. To assess its efficacy, we performed an ablation study on the Test-1 and Test-2 dataset. The results, as detailed in Table 3, clearly demonstrate that incorporating the pseudo-aerial-view synthesis module significantly improves the performance of our baseline model. While the generated pseudo-views may not possess the photorealistic quality of genuine aerial imagery, their primary value lies in their ability to effectively mitigate the substantial domain gap between street-view and satellite-view images. By providing geometrically aligned, albeit non-photorealistic, representations, the network is empowered to learn a richer, more robust cross-view scene representation. This data-level augmentation inherently reduces the difficulty of the cross-view localization task and simultaneously enhances the model's generalization capability to unseen scenes. The experimental evidence confirms that this module is a crucial contributor to the overall performance gains.

Table 3. Ablation study of the proposed components. The best results are highlighted in bold.

Method	R@1	R@5	R@10	R@1%
	Test-1			
Baseline	80.80	96.51	97.40	95.83
Baseline + Pseudo Aerial-view	81.80	97.50	97.01	97.98
Baseline + Pseudo Aerial-view + Mixed feature	**82.90**	**98.38**	**98.43**	**98.46**
	Test-2			
Baseline	17.80	34.11	45.37	89.98
Baseline + Pseudo Aerial-view	18.73	36.92	47.14	90.77
Baseline + Pseudo Aerial-view + Mixed feature	**19.20**	**38.04**	**48.90**	**91.38**

Effect of mixed feature enhancement. We further investigated the effectiveness of our proposed Mixed Feature Enhancement (MFE) module. The ablation results, also presented in Table 3, illustrate a notable performance boost when this module is integrated. The underlying rationale for the MFE module is to refine the generic features learned from a pre-trained model like DINOv2. While DINOv2 features are powerful and generalize well to a variety of tasks due to their training on diverse datasets, they are not inherently optimized for the specific nuances of cross-view geo-localization. Our feature-mixer operation is specifically designed to fine-tune these generic features, allowing them to capture the subtle but critical details required for accurate cross-view matching. The experimental findings validate that this operation effectively enhances the discriminative power of the feature embeddings, thereby leading to improved performance in the retrieval task. The consistent performance uplift across the board confirms the MFE module's role as a vital component for maximizing localization accuracy.

Effect of synthesized tilted views. In our process of synthesizing novel viewpoint images based on 3DGS, we indeed synthesized new images with tilt angles of 30°, 45°, 60°, and 90°, respectively. The results of ablation studies are listed in Table 4. It can be observed that the tilt angles of 60° and 90° achieve comparable results. The selection of specific tilt angles such as 60°/90° is primarily motivated by their role in establishing a smoother transitional path in the feature space between street-level and aerial perspectives.

These intermediate viewpoints facilitate better feature alignment, effectively reducing the domain gap and enhancing the model's cross-view geo-localization accuracy. Moreover, by intentionally constraining the number of synthesized transitional views, we are able to not only validate the robustness of this smooth-interpolation strategy but also maintain high efficiency and low computational overhead throughout the novel view synthesis process.

Table 4. Ablation study of different tilt angles on the KITTI Test-1 subset. The best results are highlighted in bold.

Method	R@1	R@5	R@10	R@1%
Baseline + Pseudo Aerial-view (tilt angle 30°)	81.05	96.50	96.56	97.11
Baseline + Pseudo Aerial-view (tilt angle 45°)	81.65	97.44	97.03	97.82
Baseline + Pseudo Aerial-view (tilt angle 60°)	81.61	97.00	**97.10**	97.53
Baseline + Pseudo Aerial-view (default tilt angle 90°)	**81.80**	**97.50**	97.01	**97.98**

5. Discussion

Although our method produces good results in some complex scenarios, it still struggles with certain challenges such as occlusion, illumination variation, and high urban density. As illustrated in the Figure 6, our method currently struggles with challenging street scenes characterized by significant lighting variations and complex clutter, which can lead to retrieval failures. Regarding occlusion, we acknowledge that severe occlusion can pose challenges for 3D reconstruction and novel view synthesis based on 3DGS. We have considered some factors in our method to alleviate these issues. Specifically, by incorporating DINOv2, which provides powerful and robust visual features, and combining it with 3DGS, our approach is engineered to mitigate the adverse effects posed by such challenging conditions.

Figure 6. Visualization of failure cases in our method.

6. Future Work

Our current approach has limitations in generalizing to diverse cross-view scenarios, particularly for panoramic-to-perspective benchmarks like CVUSA and VIGOR. Indeed, the dense, sequential nature of perspective images in datasets such as KITTI and Oxford RobotCar enables high-quality 3D reconstruction with 3DGS. In contrast, panoramic benchmarks like CVUSA and VIGOR are composed of sparsely captured images with limited overlap, which challenges the current 3DGS-based pipeline that relies on dense views for robust geometry modeling. This structural difference currently restricts the direct applicability of our method to such panoramic-to-perspective settings. we regard overcoming this domain

gap as a vital research direction. In the future, we plan to explore techniques tailored to sparse or non-sequential imagery, such as cross-view generative models that bypass explicit 3D reconstruction, or geometry-aware methods capable of learning from limited overlapping views.

7. Conclusions

In this paper, we proposed a novel 3D Gaussian splatting-based cross-view geo-localization framework to resolve the difficulty of geo-localization caused by low feature similarity between street-view and aerial image. We first synthesized novel aerial perspectives from street-view sequences via 3D Gaussian splatting, utilizing COLMAP-derived initial poses and sparse point clouds. A key contribution is our camera pose estimation strategy, which optimizes matching viewpoints by increasing the tilt angle relative to the horizontal plane, thereby improving geometric consistency between the synthesized and real aerial imagery. Following this, we employed DINOv2 in a straightforward yet efficient mixed feature enhancement module, optimized using InfoNCE loss. Experiments on the KITTI dataset demonstrate that our method delivers substantial improvements in matching accuracy despite significant viewpoint differences and achieves advanced retrieval results.

Author Contributions: Conceptualization, X.D.; Methodology, X.D.; Software, X.D. and X.Z.; Validation, X.D., X.Z., S.S. and L.H.; Formal analysis, X.D. and S.S.; Investigation, S.S. and L.H.; Resources, Y.D.; Data curation, Y.D.; Writing—original draft, X.D.; Writing—review & editing, B.L., L.H. and Y.D.; Supervision, B.L. and Y.D.; Project administration, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Author Shangzhen Song was employed by the company Xi'an Flight Automatic Control Research Institute. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
2. Chen, H.; Hou, L.; Wu, S.; Zhang, G.; Zou, Y.; Moon, S.; Bhuiyan, M. Augmented reality, deep learning and vision-language query system for construction worker safety. *Autom. Constr.* **2024**, *157*, 105158. [CrossRef]
3. Rubio, F.; Valero, F.; Llopis-Albert, C. A review of mobile robots: Concepts, methods, theoretical framework, and applications. *Int. J. Adv. Robot. Syst.* **2019**, *16*, 1729881419839596. [CrossRef]
4. Lin, T.Y.; Belongie, S.; Hays, J. Cross-view image geolocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 891–898.
5. Tian, Y.; Chen, C.; Shah, M. Cross-view image matching for geo-localization in urban environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3608–3616.
6. Workman, S.; Jacobs, N. On the location dependence of convolutional neural network features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 70–78.
7. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
8. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded up robust features. In Proceedings of the European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006; pp. 404–417.
9. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.

10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Lake Tahoe, CA, USA, 3–6 December 2012; Volume 25.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
12. Zhai, X.; Kolesnikov, A.; Houlsby, N.; Beyer, L. Scaling vision Transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12104–12113.
13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Zhu, S.; Yang, T.; Chen, C. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3640–3649.
16. Cai, S.; Guo, Y.; Khan, S.; Hu, J.; Wen, G. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8391–8400.
17. Regmi, K.; Borji, A. Cross-view image synthesis using conditional GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3501–3510.
18. Vo, N.N.; Hays, J. Localizing and orienting street views using overhead imagery. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 494–509.
19. Liu, L.; Li, H. Lending orientation to neural networks for cross-view geo-localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5624–5633.
20. Shi, Y.; Yu, X.; Liu, L.; Zhang, T.; Li, H. Optimal feature transport for cross-view image geo-localization. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 11990–11997.
21. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014.
22. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Virtual, 2020; pp. 405–421.
23. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **2023**, *42*, 139–152 . [CrossRef]
24. Schönberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
25. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
26. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
27. Castaldo, F.; Zamir, A.; Angst, R.; Palmieri, F.; Savarese, S. Semantic cross-view matching. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Santiago, Chile, 7–13 December 2015; pp. 9–17.
28. Senlet, T.; Elgammal, A. A framework for global vehicle localization using stereo images and satellite and road maps. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Barcelona, Spain, 6–13 November 2011; IEEE: New York, NY, USA, 2011; pp. 2034–2041.
29. Bansal, M.; Sawhney, H.S.; Cheng, H.; Daniilidis, K. Geo-localization of street views with aerial image databases. In Proceedings of the ACM International Conference on Multimedia (ACM MM), Scottsdale, Arizona, 28 November–1 December 2011; pp. 1125–1128.
30. Workman, S.; Souvenir, R.; Jacobs, N. Wide-area image geolocalization with aerial reference imagery. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3961–3969.
31. Lin, T.Y.; Cui, Y.; Belongie, S.; Hays, J. Learning deep representations for ground-to-aerial geolocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5007–5015.
32. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; pp. 1735–1742.
33. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 994–1003.

34. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
35. Xu, C.; Hui, L.; Xie, J.; Yang, J. Weakly Supervised Object Localization with Progressive Activation Diffusion. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 15194–15206. [CrossRef] [PubMed]
36. Zhai, M.; Bessinger, Z.; Workman, S.; Jacobs, N. Predicting ground-level scene layout from aerial imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 867–875.
37. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
38. Hu, S.; Feng, M.; Nguyen, R.M.; Lee, G.H. CVM-NET: Cross-view matching network for image-based ground-to-aerial geo-localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7258–7267.
39. He, Q.; Xu, A.; Zhang, Y.; Ye, Z.; Zhou, W.; Xi, R.; Lin, Q. A contrastive learning based multiview scene matching method for UAV view geo-localization. *Remote Sens.* **2024**, *16*, 3039. [CrossRef]
40. Pillai, M.S.; Rizve, M.N.; Shah, M. GAReT: Cross-view video geolocalization with adapters and auto-regressive transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 466–483.
41. Zhu, S.; Shah, M.; Chen, C. TransGeo: Transformer is all you need for cross-view image geo-localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1162–1171.
42. Kwon, J.; Kim, J.; Park, H.; Choi, I.K. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 5905–5914.
43. Cui, Z.; Zhou, P.; Wang, X.; Zhang, Z.; Li, Y.; Li, H.; Zhang, Y. A novel geo-localization method for UAV and satellite images using cross-view consistent attention. *Remote Sens.* **2023**, *15*, 4667. [CrossRef]
44. Ding, L.; Zhou, J.; Meng, L.; Long, Z. A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization. *Remote Sens.* **2020**, *13*, 47. [CrossRef]
45. Chen, G.; Wang, W. A survey on 3D gaussian splatting. *arXiv* **2024**, arXiv:2401.03890. [CrossRef]
46. Barron, J.T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P.P. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 5855–5864.
47. Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; Kanazawa, A. Plenoxels: Radiance fields without neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5501–5510.
48. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* **2022**, *41*, 1–15. [CrossRef]
49. Pumarola, A.; Corona, E.; Pons-Moll, G.; Moreno-Noguer, F. D-NeRF: Neural radiance fields for dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Montreal, QC, Canada, 11–17 October 2021; pp. 10318–10327.
50. Wu, G.; Yi, T.; Fang, J.; Wang, L.; Zhang, X.; Wang, W.; Wang, Q.; Zha, C.; Tai, Y.L.; Tang, C.Z. 4D Gaussian splatting for real-time dynamic scene rendering. *arXiv* **2023**, arXiv:2310.08528. [CrossRef]
51. Zollmann, S.; Zafeiridis, P.; Agapito, L.; Pont-Tuset, J.; Ranftl, R. Relightable 3D Gaussians: Real-time Point Cloud Relighting with BRDF Decomposition and Ray Tracing. *arXiv* **2024**, arXiv:2311.17922.
52. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
53. Deuser, F.; Habel, K.; Oswald, N. Sample4Geo: Hard negative sampling for cross-view geo-localisation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 16847–16856.
54. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1000 km: The Oxford RobotCar dataset. *Int. J. Robot. Res.* **2017**, *36*, 3–15. [CrossRef]
55. Shi, Y.; Yu, X.; Wang, S.; Li, H. CVLNet: Cross-view semantic correspondence learning for video-based camera localization. In Proceedings of the Asian Conference on Computer Vision (ACCV), Macao, China, 4–8 December 2022; pp. 123–141.
56. Mehta, H.; Kanani, P.; Lande, P. Google maps. *Int. J. Comput. Appl.* **2019**, *178*, 41–46. [CrossRef]
57. Shi, Y.; Yu, X.; Liu, L.; Campbell, D.; Koniusz, P.; Li, H. Accurate 3-DoF camera geo-localization via ground-to-satellite image matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2682–2697. [CrossRef] [PubMed]
58. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
59. Shi, Y.; Liu, L.; Yu, X.; Li, H. Spatial-aware feature aggregation for image based cross-view geo-localization. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.

60. Shi, Y.; Yu, X.; Campbell, D.; Li, H. Where am i looking at? Joint location and orientation estimation by cross-view matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4064–4072.
61. Toker, A.; Zhou, Q.; Maximov, M.; Leal-Taixé, L. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 6488–6497.