



## Research Article

## A guided approach for cross-view geolocalization estimation with land cover semantic segmentation

Nathan A.Z. Xavier<sup>a,b,c,\*</sup>, Elcio H. Shiguemori<sup>b,d,e</sup>, Marcos R.O.A. Maximo<sup>b</sup>, Mubarak Shah<sup>c</sup><sup>a</sup> Technical College of UFMG, Federal University of Minas Gerais, Belo Horizonte 31270-901, Brazil<sup>b</sup> Aeronautics Institute of Technology, São José dos Campos 12228-900, Brazil<sup>c</sup> Center for Research in Computer Vision, University of Central Florida, Orlando 32816, USA<sup>d</sup> Institute for Advanced Studies, São José dos Campos 12228-001, Brazil<sup>e</sup> National Institute for Space Research, São José dos Campos 12227-010, Brazil

## ARTICLE INFO

## Article history:

Received 29 September 2024

Revised 9 December 2024

Accepted 15 December 2024

Available online 11 January 2025

## Keywords:

Cross-view geolocalization

Semantic segmentation

Satellite and ground image fusion

Simultaneous localization and mapping

(SLAM)

## ABSTRACT

Geolocalization is a crucial process that leverages environmental information and contextual data to accurately identify a position. In particular, cross-view geolocalization utilizes images from various perspectives, such as satellite and ground-level images, which are relevant for applications like robotics navigation and autonomous navigation. In this research, we propose a methodology that integrates cross-view geolocalization estimation with a land cover semantic segmentation map. Our solution demonstrates comparable performance to state-of-the-art methods, exhibiting enhanced stability and consistency regardless of the street view location or the dataset used. Additionally, our method generates a focused discrete probability distribution that acts as a heatmap. This heatmap effectively filters out incorrect and unlikely regions, enhancing the reliability of our estimations. Code is available at <https://github.com/nathanxavier/CVSegGuide>.

© 2025 The Author(s). Published by Elsevier B.V. on behalf of Shandong University. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Cross-view geolocalization is an emerging challenge in outdoor guidance that involves aligning images captured from different platforms but depicting the same geographic location. With the growing reliance on digital maps and real-time geolocation, this problem has gained increasing attention. The primary real-time geospatial localization systems are global navigation satellite systems (GNSS), including GPS (United States), GLONASS (Russia), Galileo (European Union), BeiDou (China), NavIC (India), and QZSS (Japan) [1]. However, GNSS is frequently subject to failures due to jamming, denial, and other signal interruptions [2–9].

In contrast, visual navigation utilizing image retrieval methods offers a promising solution to GNSS shortcomings. Satellite images, which are consistently available, can serve as reliable references for outdoor location estimation [10,11]. Techniques such as visual odometry and landmark recognition, leveraging extensive satellite datasets, are becoming increasingly popular for geolocation tasks [12–15]. With technological advancements, the integration of satellite and drone imagery has become more prominent. Drones, providing bird's-eye view (BEV) perspectives through oblique photos, present unique challenges for cross-view

geolocalization due to their distinct perpendicular viewpoints compared to ground-based images [16–20].

A significant area of research focuses on matching ground images with aerial or overview images, considering factors like viewpoint variation, lighting, obstructions, and seasonal changes [21–23]. The subsequent geolocalization process typically involves techniques such as polar transformations, convolutional neural networks (CNNs), or transformers, which learn to match street view images to their corresponding locations in aerial views [24–27].

Several datasets facilitate cross-view geolocalization research, including CVUSA [28], CVACT [21], and VIGOR [29]. CVUSA provides images from 20 cities in the United States, including major urban centers such as New York, San Francisco, and Los Angeles, and serves as a benchmark with GPS coordinates for ground-view images. CVACT offers high-resolution images from Google Maps and Google Street View in Canberra, Australia, with panoramic ground images geolocalized by GPS. The VIGOR dataset stands out for a consistent GPS interval between panorama samples, maintaining around 30 m distance between images. Other datasets emphasize specific aspects like orientation and depth [30,31].

Solutions proposed for cross-view geolocalization employ methods such as polar transformation CNNs [22,27,28,32–36], transformers [24,37–40], and other deep learning architectures [41–45]. These approaches have garnered

\* Corresponding author.

E-mail address: [nathanxavier@ufmg.br](mailto:nathanxavier@ufmg.br) (N.A.Z. Xavier).

significant academic interest and are considered viable alternatives to simultaneous localization and mapping (SLAM) for outdoor guidance [46].

Recognizing that topological mapping can reduce the complexity of cross-view geolocalization, researchers have investigated several advanced techniques based on semantic segmentation maps. These include multi-agent segmentation methods [47], the integration of aerial and street image segmentation [48], visual odometry approaches [49], and the combination of aerial segmentation with ground depth mapping [20].

In this paper, we propose leveraging an aerial semantic segmentation map as a guiding tool for more accurate location estimation. This approach aims to filter out transient and irrelevant details from ground-view segmented images. By focusing on aerial features such as buildings and roads, along with their spatial relationships, the aerial semantic map offers crucial context for cross-view geolocalization. In contrast, street-view semantic segmentation often emphasizes moving objects like cars and pedestrians, rather than the more stable environmental features [50]. Our method estimates a discrete probability distribution (DPD), similar to a heatmap, for the location of a street-view image on an overhead satellite photo, using only the information from the aerial segmented map. This technique enhances the performance of fusion systems, including Kalman filters [25,51] and particle filtering [49,52], by providing a stackable solution for estimating outdoor location in applications such as SLAM [53,54] and autonomous navigation. We train our approach on the Brooklyn and Queens dataset [55] and validate it on the VIGOR dataset [29], both of which include satellite images, 360° street-view images. The proposed architecture is designed to be flexible and applicable to various cross-view datasets containing satellite and ground-view images without the need for retraining, as only the training dataset including aerial semantic maps. To our knowledge, detailed studies on cross-view geolocalization guided by semantic segmentation maps are still limited.

The main contributions of this paper include a novel cross-view geolocalization framework that estimates a discrete probability distribution for image locations. We also present an innovative architecture that leverages semantic segmentation maps from satellite images to enhance location estimation, while introducing new transformer backbones to refine and improve the outdoor location estimation pipeline. This adaptability makes the proposed method suitable for diverse geographic regions and use cases, from urban to suburban areas, without the need for dataset-specific tuning.

The organization of this article starts by describing related works in Section 2. The problem statement, the dataset used, and the proposed methodology are presented in Section 4. Section 5 presents the experiments and their results. Finally, Section 6 provides the conclusion of this paper.

## 2. Related works

In this section, we present multiple developed studies applied to semantic segmentation and cross-view geolocalization estimation, since these are the main contributions of this research.

### 2.1. Semantic segmentation map

Semantic segmentation involves learning to cluster similar features together while distinguishing between mismatched pairs. This process is often measured using similarity metrics such as Euclidean distance [35]. In this context, the segmentation map

aims to identify and classify a discrete set of objects in satellite images [32,51,56–58], effectively simplifying the aerial view into a manageable homography map [35,59,60].

However, satellite imagery-based mapping faces several limitations, such as the challenge of segmenting non-visible or occluded regions [58]. Incorporating ground-level images can enhance the understanding of urban areas and improve the accuracy of the mapped regions [61]. Advances in aerial mapping have been achieved through research focused on ground-level fusion techniques [55,62].

Segmenting large areas presents its own challenges, as it requires high resolution to accurately delineate objects. The diverse shapes and sizes of objects further complicate aerial segmentation [35,54]. Solutions like Mask2Former [63] and UNetFormer [38] often struggle to produce reliable and continuous semantic maps. In contrast, the Segment Anything Model (SAM) [64,65] excels at defining precise mask boundaries but tends to generate numerous masked objects [54]. The FeatUp [66], a recent and robust segmentation tool, utilizes joint bilateral upsampling (JBU) [67] to enhance image features, thereby improving spatial resolution and overall semantic segmentation performance.

In this study, the semantic segmentation mask serves as an auxiliary task, designed to guide and enhance the performance of our cross-view geolocalization estimation method.

### 2.2. Cross-view geolocalization

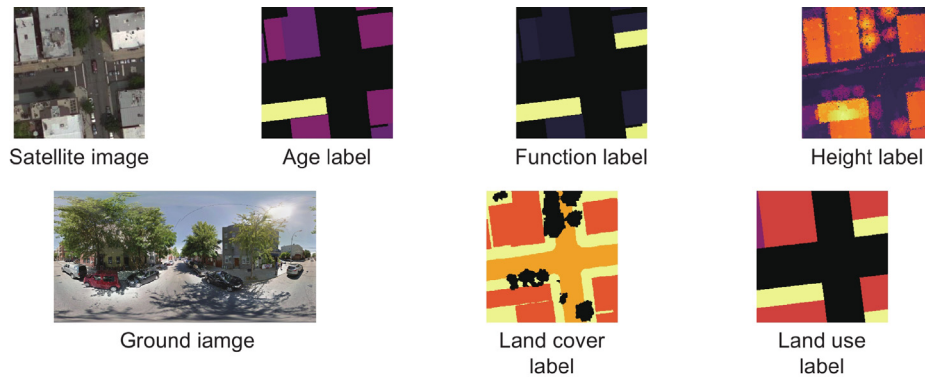
Cross-view geolocalization is a rapidly advancing field with numerous applications across various industries, including automotive [39,52,60], aerospace [68,69], and robotics [70], among others [45,71]. The two predominant techniques for estimating global positions are CNNs and transformers.

CNNs operate by extracting features from both aerial and ground-view images and learning the relationships between these feature spaces [27,28,32,34]. A notable advancement in this area is spatial-aware feature aggregation (SAFA) [33], a revolutionary solution that introduced polar transformations to simplify the alignment between different views. Subsequent research continued to use polar transformations, building on the SAFA approach [22,33,36]. Additionally, generative adversarial networks (GANs) have been applied to enhance feature extraction through data augmentation and synthetic images, providing a more generalized and flexible solution [41–44].

Transformers, in contrast, offer an alternative that bypasses the need for polar transformations and data augmentation. TransGeo [24] demonstrated that transformers typically require lower computational resources, including GPU usage, memory, and inference time, compared to CNN-based approaches [24]. This innovation has led to developments utilizing transformers exclusively [25,26,48,57] or in combination with CNNs [37,39,40], aiming to harness the strengths of both architectures within a unified neural network model.

Cross-view geolocalization also leverages the fundamental principle of image retrieval, which involves querying based on image similarity [27]. Recent solutions, such as GeoDTR [72] and the Feature Recombination Module (FRM) [73], explore advanced feature representations to reduce ambiguities and enhance the spatial alignment between aerial and ground images. In contrast, Sample4Geo [74] focuses on improving the architecture and pipeline, optimizing the overall process for more robust cross-view geolocalization.

Datasets such as CVUSA [28], CVACTION [21], VIGOR [29], and the Brooklyn and Queens [28], among others [71], contain ground images from various arbitrary locations within a given area. These datasets present challenges for models using polar



**Fig. 1.** Dataset overview showcasing a sample of aerial and ground-level images, along with their corresponding aerial labels: age, function, height, land cover, and land use.

transformations, as the ground images are often not spatially aligned [22,24,27].

Combining segmentation with cross-view geolocation has proven to be an effective strategy for addressing a range of technical and scientific challenges. Semantic segmentation enhances neural network architectures by reducing complexity, enabling parallel processing, and facilitating real-time operation [53,54,60]. By leveraging semantic segmentation maps, it is possible to filter out dynamic elements such as cars and pedestrians, as well as transient or seasonal features like foliage and color changes [35,53]. Pseudo-segmentation methods have been developed to improve the correlation between satellite imagery and bird's-eye view (BEV) images [57, 75]. Additionally, pseudo-labeled pose estimation has been employed to enhance cross-view geolocation predictions and other multi-model distributions [25,27,76]. Sequential ground-view images and segmented satellite photos have also been integrated within visual odometry frameworks to improve pose estimation [49]. More recently, the CVLocationTrans model [77] introduced a fine-grained cross-view approach that combines self-attention and cross-attention layers. In this model, features are initially extracted by a ResNet50, then combined to establish correspondences, followed by classification and regression headers to predict locations.

To our knowledge, there is limited research on applying true semantic segmentation to satellite images specifically for cross-view geolocation using aerial semantic segmentation as ground truth. Our approach introduces a novel method where semantic segmentation maps are utilized to guide location estimation, leveraging datasets with street-view samples that are not aligned with satellite imagery. This method is designed to be adaptable and applicable to any cross-view image dataset, with semantic segmentation performed internally within the model. Additionally, we evaluate the performance of this approach both as a stackable solution and as an end-to-end system.

### 3. Brooklyn and queens dataset

The main purpose of the Brooklyn and Queens dataset was to estimate three challenging labels (building age, building function, and land use) from images taken from the two major boroughs of New York City, the neighborhoods Brooklyn and Queens [55]. The same dataset was extensively used for classification and segmentation [78,79]. The dataset was also extended with two new labels (height and land cover) [62]. Fig. 1 presents one sample of the dataset.

The dataset contains non-overlapping satellite images with, approximately, 30 cm resolution and panoramas of ground-level

**Table 1**  
Number of image samples per neighborhood.

Neighborhood	Overview images	Ground-level images
Brooklyn	43,605	139,327
Queens	10,044	38,603

images obtained from Google Street View. The whole dataset is made available with a large number of images, as presented in Table 1.

We can see that the number of street-view images is larger than the satellite images, which indicates multiple panorama images in the same region. Fig. 2 presents the rate of ground view samples per aerial image, showing that more than 20% of the aerial samples dataset has no street sample. Once the main goal for this study is the correct estimation of the location based on cross-view images, we consider only the aerial set with, at least, one ground-level panorama image available.

All satellite image sizes are  $256 \times 256 \times 3$ , while the ground-view image sizes are  $1664 \times 3328 \times 3$ , composed of a three colors channel. The label masks share the same resolution as the corresponding overview photo but have a single channel that indicates the class based on the type of label. Since the aim is to show the improvement of the cross-view geolocation aided by the semantic segmentation map, we opted to use only the land cover mask. The labels are classified as tree canopy, grass, bare soil, water, buildings, roads, railroads, and other impervious [62].

Fig. 3 shows the percentage of time each class was present or absent on the interested set of aerial images. As observed, some classes are only available on a few satellite imagery, such as water, bare soil, and railroads, which makes the segmentation even more challenging [80].

A second analysis was made on the average pixel extension of each class when the occurrence is observed in the image, presented in Fig. 4. Comparing the percentage of times the classes appear in Fig. 3 and the average pixel extension in Fig. 4, we decided to exclude the grass class because of its low average extension, once the segmentation works as a guidance in this research.

Knowing that satellite images can be taken in different seasons, with different lighting, and aiming for a solution that avoids most transient objects [35], in addition to the previous analysis presented about the segmentation maps, we opted to use the classes (i) tree canopy; (ii) buildings; (iii) roads, and; (iv) other impervious, for development of the solution.

### 4. Problem statement and methodology

This section presents the problem and the steps taken to structure the solution. The method is divided into two main

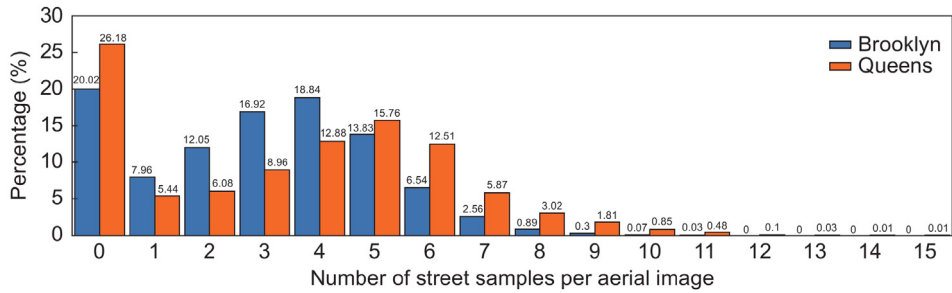


Fig. 2. Number of street-view images per overview image.

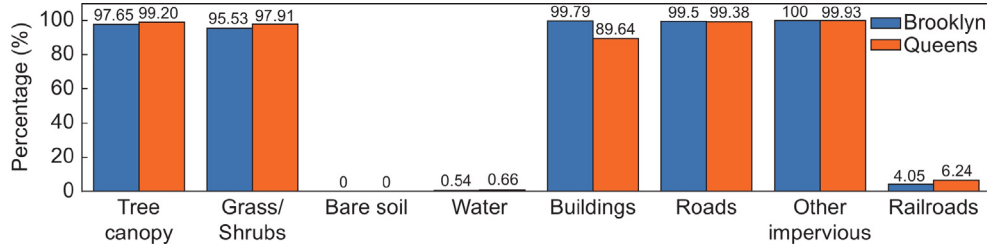


Fig. 3. Percentage of binary occurrence of each class in the land cover label image.

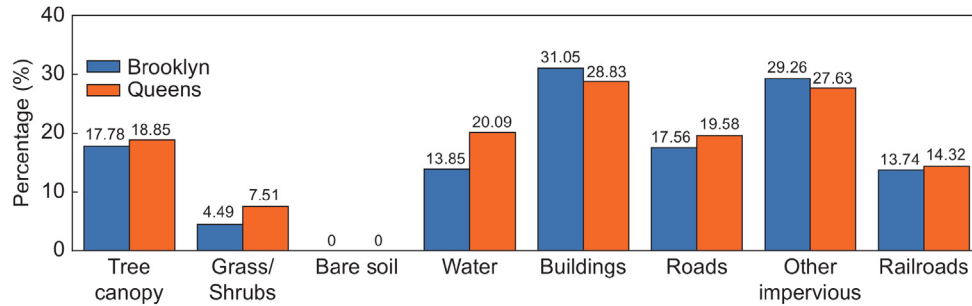


Fig. 4. Average pixel extent of each class when present in the land cover label image.

branches, the first focusing on the segmentation map and the second on the geolocalization of the ground-level image.

#### 4.1. Problem statement

Given a street-view image  $I_g \in \mathbb{R}^{H \times W \times C}$  and a tagged aerial image  $I_a \in \mathbb{R}^{L \times L \times C}$ , where the satellite imagery covers the ground view image location, our objective is to estimate the position  $\hat{\rho} \in \mathbb{R}^2$  of the camera that captured  $I_g$ . The images have a resolution of  $(H, W)$  or  $(L, L)$ , and  $C$  is the number of channels available.

The proposed solution involves a fusion system that outputs the discrete probability distribution  $\mathbf{y}_g \in \mathbb{R}^{L \times L}$ . This distribution represents the probability of  $I_g$  being taken at each pixel in  $I_a$  effectively enabling us to localize the street-view image within the aerial context.

#### 4.2. Implementation details

Fig. 5 presents an overview of the proposed method to solve the problem stated. The first branch of the architecture calculates the semantic segmentation map of  $I_a$  by a segmentation filter block from the transformer embedding features. We evaluate two different approaches for the transformer block. The first one uses a multi-scale transformer (MST) [57,81], which extracts features in different patch sizes. The second transformer uses a pre-trained FeatUp model [66], designed specifically for the semantic

segmentation of images. The second branch receives  $I_g$ , extracting its embedding features by the same transformer model. By calculating the cosine similarity between these features and those from  $I_a$ , and concatenating with the semantic segmentation map and the aerial embedding features, the architecture computes the DPD, indicating the likelihood of  $I_g$  be taken at each pixel of  $I_a$ .

As observed, the model can be used as a semantic segmentation of satellite imagery such as calculating the discrete probability distribution of a ground-view image taken in the same aerial region imaged.

##### 4.2.1. Overview image

The input overhead image used in the first branch of the model has a size of  $256 \times 256 \times 3$ . The same size is observed on the training land cover map. Similarly, this configuration is aimed at the output of the proposed methodology.

##### 4.2.2. Ground-view image

The ground-view image is based on an equirectangular projection, creating a panorama image. Intending to reduce the distortion obtained on the top and bottom of the image, only for our methods, we crop around 50% of the image. After cropping, the image is resized to  $128 \times 512$ , maintaining its rate between height and width. The resulting image is a panorama image of size  $128 \times 512 \times 3$  which is used as input for the proposed solution [62]. Fig. 6 shows an example of this pre-processing. The result focuses on environmental information, with less sky and underfoot ground data.

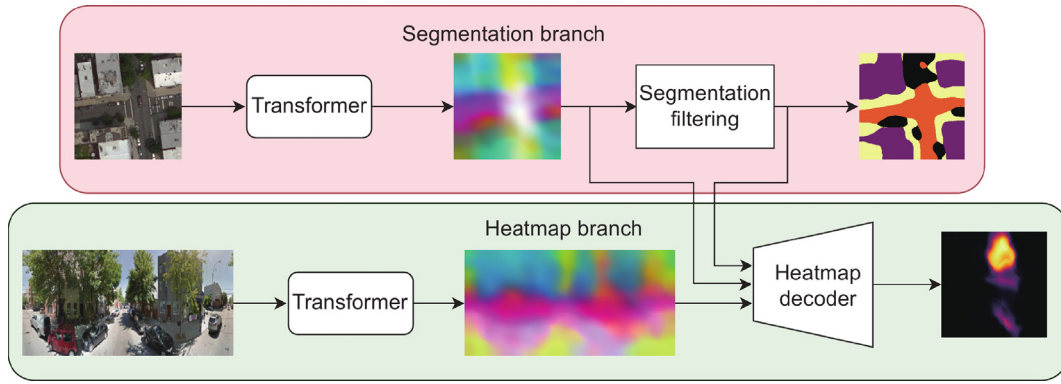


Fig. 5. Methodology overview where the first branch is identified as the segmentation branch and the second, heatmap branch. We present the embedding features after the transformer blocks that feed the segmentation filtering and the heatmap decoder.



Fig. 6. Process for removing the distorted top and bottom panoramic areas of the ground view image.

#### 4.2.3. Sightline map

The sightline map is obtained by combining the ground truth of the aerial land cover and the correct position where the ground view was taken, generating a 2D aerial field of view map to calculate the positional aerial discrete distribution.

The sightline map represents a Gaussian distribution centered on the street view's real position. The land cover reduces the probability created by the Gaussian distribution. Based on the field of view, we use the building class from the land cover as a barrier, blocking the distribution from spreading behind the building. Other classes do not impact the sightline map since it is impossible to identify its relevance without analyzing each street image separately. Fig. 7 shows the sightline map creation for a single street and aerial photo.

#### 4.2.4. Vision transformer for geolocalization

The vision transformer (ViT) [82] was first adopted as an image-extracting features method applying a standard Transformer architecture. ViT introduced components such as patch embeddings, position embeddings, and multi-head attention.

The patch embedding represents the conversion of an input image  $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$  into a sequence of flattened patches  $\mathbf{x}_p \in \mathbb{R}^{N(P \times P \times C)}$ , where  $(P, P)$  is the resolution of the image patch, and  $N = H \cdot W / P^2$  is the resulting number of patches [82]. The  $N$  patches are fed into a trainable linear projection layer, generating  $N$  tokens with  $D$  feature dimensions.

Similar to BERT [83], a class token is appended to the  $N$  tokens to integrate classification information from the image representation. Position embeddings can also be added to the patch embeddings to maintain the positional information. This patch vector goes to the transformer encoder which alternates from multiple layers of multi-headed self-attention (MSA), multilayer perceptron (MLP), and layer normalization (LN), generating class embedding feature  $\mathbf{y}_{class} \in \mathbb{R}^D$  and patch embedding features  $\mathbf{y}_p \in \mathbb{R}^{N(P \times P \times D)}$  [82].

**Multi-Scale Transformer (MST):** The MST is based on the Crossformer transformer [84], a variation of the ViT. The architecture extracts the embedding from different scales in

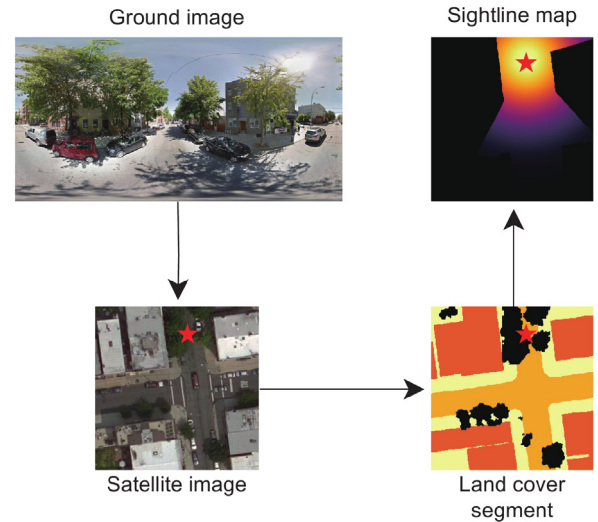
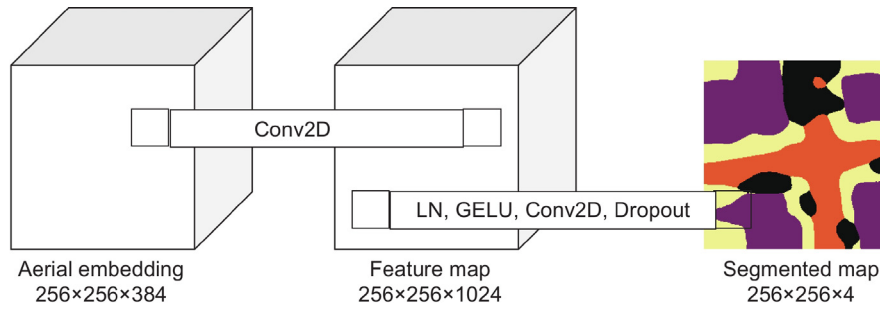


Fig. 7. Creation of the sightline map involving (1) identification of the location of the ground-level image from the satellite imagery, marked by the red star; (2) extraction of the land cover semantic segmentation map from the aerial image, and; (3) generation of the sightline map based on this segmentation map and the location of the street image.

multiple stages of patching embedding and fuses the information with the previous block with a fusion module. The method also recaptures some global-local features from sequential attention blocks at each stage's end.

The MST's output is similar to ViT but available in different scales. In this regard, we have access to a list of patch embedding features  $\mathbf{y}_p = [\mathbf{y}_{p_1}, \mathbf{y}_{p_2}, \dots, \mathbf{y}_{p_n}]$  from each  $n$ th stage of the MST.

Since we aim to obtain high-resolution features, we upscale the images in a similar manner to what is suggested in the methods SegFormer [85] and TransUNet [86], merging and upsampling the similarly scaled information interleaved by convolutional block until we obtain high-resolution feature embedding.



**Fig. 8.** Segmentation filtering overview showing the expansion into feature maps and decoding into the predicted land cover segmented map.

**FeatUp:** The FeatUp, differently from previous approaches, has as its main objective to obtain a high-resolution featured output, prosecuting consistency between lots of low-resolution feature maps [66]. It considers the joint bilateral upsampling (JBU) [67] as an upsampling strategy. The study shows its performance for segmentation and depth prediction.

To upsample down-sized feature embeddings, the FeatUp proposition considers multiple transformations to the input image, such as pads, scales, and flips [66]. All these transformations are important for identifying small differences between features during the training stage of the upsampler model.

This method is comparable to neural radiance fields (NeRF) [87] since it forces implicit representation and renders fine details. The results showed that the downsampled features and the transformed original images are comparable, therefore the model can reconstruct a good high-resolution feature map.

Therefore, considering a generic input image  $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ , the output of the FeatUp  $\mathbf{y} \in \mathbb{R}^{H \times W \times D}$  preserves the size while still containing the same feature dimension of the downsized patch embedding features  $\mathbf{y}_p$  from ViT.

#### 4.2.5. Semantic segmentation filtering

The semantic segmentation filtering is developed based on the output from the transformer model. Since the transformers' output is the same size as the input, the segmentation model consists of a feed-forward network (FFN) that maps the features directly to class prediction added to a linear layer [88]. This proposition makes a pixel-wise prediction of each class in parallel by transforming the embedding features  $\mathbf{y} \in \mathbb{R}^{H \times W \times D}$  of the generic image into a segmented mask  $\mathbf{y}_s \in \mathbb{R}^{H \times W \times S}$ , where  $S$  describes the number of classes desired.

Fig. 8 details the information flow proposed for the segmentation filter. The high-resolution embedding feeds the segmentation filter which expands the embedding with a size of  $256 \times 256 \times 384$  to a feature map of size  $256 \times 256 \times 1024$  and then applies a feed-forward network. The FFN is composed of a convolutional layer, a layer normalization, and a Gaussian error linear unit (GELU) as function activation [89]. The last layer predicts the four classes, similar to the target label from the land cover segmented map. For a detailed description of the neural network, refer to the Appendix.

In the inference process, the predicted semantic segmentation can serve as an output mask for datasets where labels are unavailable. This approach enhances the model's ability to generalize and improves performance, especially in scenarios with limited labeled data.

#### 4.2.6. Heatmap prediction

For the heatmap prediction, we compute the discrete cross-view geolocalization distribution that consistently estimates the camera position that took the ground-panorama image over

the aerial region. This process consider using the embedding calculated from the transformers approaches for the aerial and street photos, with sizes of  $256 \times 256 \times 384$  and  $128 \times 512 \times 384$ , respectively. The same FFN is applied to the aerial embedding to estimate the aerial feature map, while the street embedding class is extracted using an adaptive average pool, resulting in a vector with 1024 environmental features.

A cosine similarity is applied between the aerial image embedding features and the ground view class embedding feature for the heatmap prediction. Once the class embedding carries the whole image representation, this operation captures the environment's appearance of the street view image. It correlates these features pixel-wisely with the satellite photo [90]. As a novel approach, the aerial segmented map is concatenated with the cosine similarity map resulting in an array of size  $256 \times 256 \times 5$ . Finally, a convolutional block predicts the cross-view geolocalization discrete probability distribution  $\mathbf{y}_g \in \mathbb{R}^{H \times W}$ . An overview of the heatmap decoder is presented in Fig. 9. For a thorough explanation of the neural network, see the Appendix.

The heatmap generated by the cosine similarity operation provides positional encodings, which are used in various cross-view geolocalization solutions [25,60,90]. The segmented mask produced by the semantic segmentation filter adds additional context to the cosine similarity, helping the model to focus on relevant features from the aerial image and avoid improbable regions by emphasizing recognizable land cover patterns.

#### 4.2.7. Loss functions

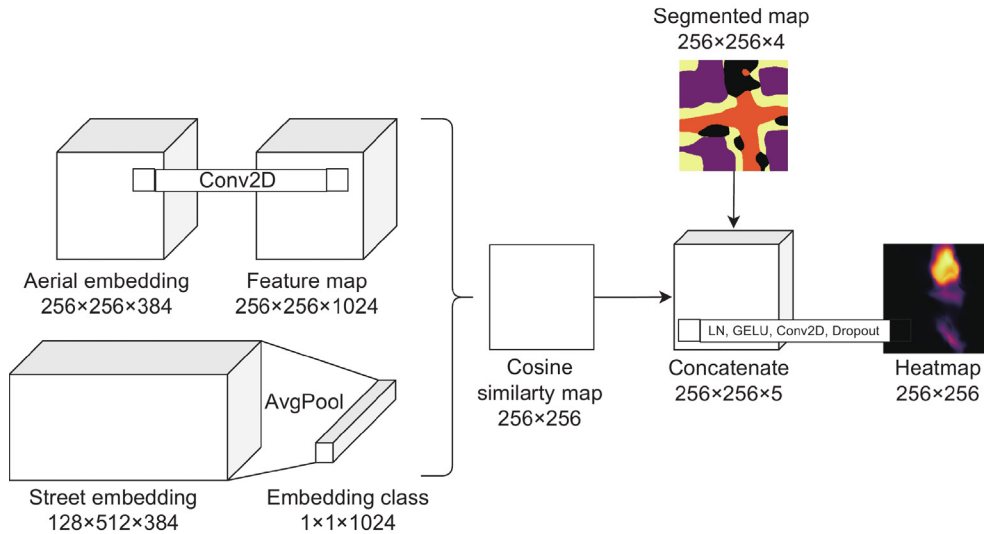
The loss function consists of two parts, based on available information on the dataset and each branch of the proposed solution.

We apply cross-entropy (CE) and Dice losses by looking at the aerial branch as a segmentation problem. The CE loss compares the ground-truth class of each pixel with the predicted segmentation map [91–93]. The Dice loss evaluates the consistency of each segment traced as convex regions [91,93].

The loss for the first branch is given by

$$\text{First Branch Loss} = \alpha \left( \frac{1}{N} \sum \underbrace{\left( - \sum \mathbf{y}_s \log(\hat{\mathbf{y}}_s) \right)}_{\text{CE Loss}} \right) + \beta \left( \frac{1}{N} \sum \underbrace{\left( 1 - 2 \sum \sum \sum \frac{(\hat{\mathbf{y}}_s \cdot \mathbf{y}_s) + \epsilon}{(\hat{\mathbf{y}}_s + \mathbf{y}_s) + \epsilon} \right)}_{\text{Dice Loss}} \right) \quad (1)$$

where  $\hat{\mathbf{y}}_s$  represents the predicted segmented mask with resolution  $(H, W)$  and  $S$  classes, and  $\mathbf{y}_s$  is the ground truth land cover segmentation map. A small constant  $\epsilon$  is added to prevent



**Fig. 9.** Heatmap decoder overview extracting the CS from the aerial and ground-level feature maps and concatenating with the predicted land cover segmentation map.

division by zero when there are no positive pixels in either the predicted or the true map. The contributions from the cross-entropy loss and the Dice loss are weighted by the factors  $\alpha$  and  $\beta$ , respectively.

Two losses are also developed for the second branch. First, a binary cross-entropy (BCE) compares the probability of the single existing class, with the BCE loss being a similar method of the CE but applied to multiple classes [71,93]. The last loss applied is the mean squared error (MSE) between the real street image position and the heatmap maximum probability location [71].

The loss for the Second Branch is defined as

Second Branch Loss

$$\begin{aligned}
 &= \gamma \left( \underbrace{-\left( \mathbf{y}_g \log \left( \frac{1}{1 + e^{\hat{\mathbf{y}}_g}} \right) + (1 - \mathbf{y}_g) \log \left( 1 - \frac{1}{1 + e^{\hat{\mathbf{y}}_g}} \right) \right)}_{\text{BCE Loss}} \right) \\
 &+ \delta \left( \underbrace{\frac{1}{N} \sum_{i=1}^N (\mathbf{y}_g - \hat{\mathbf{y}}_g)^2}_{\text{MSE Loss}} \right) \quad (2)
 \end{aligned}$$

where  $\mathbf{y}_g$  is ground truth for the DPD, and  $\hat{\mathbf{y}}_g$  is the predicted heatmap representing the model's probability distribution in the aerial image. The BCE and MSE losses are weighted by the factors  $\gamma$  and  $\delta$ , respectively.

The global loss function combines all the above criteria described, having as objective its minimization, defined as:

$$\text{Global Loss} = \alpha \cdot \text{CE Loss} + \beta \cdot \text{Dice Loss} + \gamma \cdot \text{BCE Loss} + \delta \cdot \text{MSE Loss} \quad (3)$$

The weights, determined via iterative refinement, yield the following values:  $\alpha = 1$ ,  $\beta = 1$ ,  $\gamma = 5$ , and  $\delta = 10$ .

#### 4.2.8. Experimental setup

The method proposed is implemented in PyTorch [94]. The FeatUp operation is demonstrated to serve as transfer learning for semantic segmentation [66,95]. In contrast, the MST is used intending to extract environmental features in different scales. For training the methodology, we considered the Brooklyn set of the Brooklyn and Queens dataset [55,62], while the Queens set

**Table 2**  
Dataset image sizes.

Image	Brooklyn and Queens	VIGOR
Overhead	256 × 256	640 × 640
Ground	3,328 × 1,664	1,664 × 832 or 2,048 × 1,024

was separated for validation. We also use the VIGOR dataset [29] as a third stage for testing the methods' performance and generalization.

We use the Adam optimizer [96] with a learning rate of  $1 \times 10^{-5}$  and a batch size of 5. The embedding dimension is 384, the feature dimension is 1024, and the number of classes for the semantic segmentation map is 4. The data processing used a computer containing an AMD Ryzen 9 and a NVIDIA GP102GL with 24 GB RAM.

## 5. Experiments and results

The experiments are conducted on the Brooklyn and Queens dataset [55,62] and the VIGOR dataset [29]. The model is trained on the Brooklyn set and evaluated on the Queens set. We also use the VIGOR cross-test dataset to verify the generalization of the model. A comparison is made with the models CCVPE [27] and CVLocationTrans [77], trained in the Brooklyn and Queens dataset. Both models are considered state-of-the-art for cross-view geolocation performance. We use the codes released by the authors for model implementation.

Both datasets contain street view images and aerial geo-tagged images. The image sizes are described in Table 2. The VIGOR dataset provides overhead images with higher resolutions compared to the Brooklyn and Queens datasets. However, the street-view images in VIGOR are of smaller resolution.

### 5.1. Evaluation metrics

The positional accuracy is reported using the baseline method, which includes the mean and median distance errors (in meters) between the ground truth and the predicted maximum probability location [76,77]. We categorize the performance into two groups: "positive" and "semi-positive", based on the alignment of the street view position with the aerial image. This categorization mirrors the approach used in the VIGOR

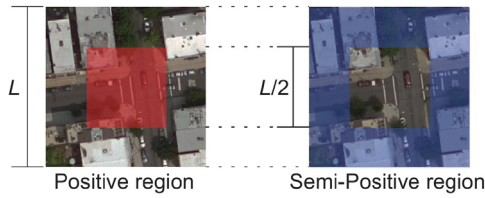


Fig. 10. Positive and semi-positive regions.

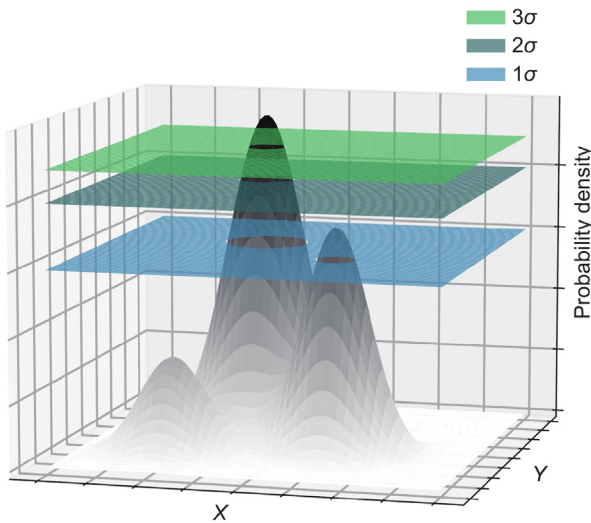


Fig. 11. Top-ranked sections in a random probability distribution.

dataset [29]. A ground view is classified as semi-positive if the corresponding aerial image includes only a portion of the scene. Basically, a street view position is deemed positive if it was taken within the central region of size  $L/2 \times L/2$ ; otherwise, it is classified as semi-positive. Fig. 10 presents an example of both categories.

Additionally, we present the positional accuracy as the percentage of correct estimations within circumferences of radii 1 m, 5 m, and 10 m [24,27].

For fusion systems, we propose evaluating the top-ranked regions of the predicted discrete distribution. This evaluation is based on the top-ranked probabilities, removing all values lower than 68.27% ( $1\sigma$ ), 95.45% ( $2\sigma$ ), and 99.73% ( $3\sigma$ ) quantiles of the entire distribution, as shown in Fig. 11. This assessment focuses on the most reliable data point, which is crucial for applications involving SLAM, such as particle filtering, Kalman filter, and other techniques, as observed in autonomous navigation [7,27,97].

Finally, as a supplementary contribution, we calculate the mean intersection over union (mIoU), which represents the pixel accuracy classification. Accurate predictions of the semantic segmentation map, while a secondary objective, are important for the overall performance of the proposed method. This also assesses the performance of label predictions on datasets where semantic segmentation ground truth labels are unavailable.

## 5.2. Training stage

The training stage involves evaluating two distinct conceptual models to enhance the semantic segmentation map for cross-view geolocation, as outlined below:

- **Guided:** This is the primary architecture, which predicts land cover and utilizes its predictions as guidance for the second branch;

- **GT-Guide:** This model bypasses land cover semantic segmentation map prediction. Instead, it uses the land cover ground truth as guidance to improve the DPD prediction, and;
- **W/o-Guide:** The last model has no guidance, predicting the cross-view geolocation only using aerial and ground view images to predict the heatmap.

In addition, we compare the performances with the CCVPE [27] and CVLocationTrans [77] models, which do not use semantic segmentation maps in their methods. We decided not to use the FeatUp backbone without guidance, as its focus on segmentation could skew the performance comparison.

## 5.3. Performance evaluation

We evaluate the models' performance using images of Queens borough from the Brooklyn and Queens dataset and the cross-test set from the VIGOR dataset. The performance comparisons are described in the following sections.

### 5.3.1. Maximum probability location comparison

We begin by evaluating the performance of the predicted maximum probability location, derived from the estimated heatmap. This metric serves as a key indicator of the end-to-end application's effectiveness in cross-view geolocation estimation.

A summary of the experimental results comparison is presented in Table 3. We evaluate the performance of GT-Guide solely on the Brooklyn and Queens dataset, as the VIGOR dataset lacks a ground truth semantic segmentation map. Every model receives the same images, following the height and width expected from the model.

The performance of the proposed method with the FeatUp backbone demonstrates an alignment with the CCVPE model in end-to-end location estimation, demonstrating a modest improvement in the positive and semi-positive regions. Leveraging the ground-truth semantic segmentation map for guidance yields a significant reduction in localization error – approximately 1 meter – compared to the fully trained model. In contrast, when trained without land cover guidance, the MST model demonstrated poorer performance, particularly in the positive region, underscoring the critical role of land cover information in improving localization accuracy. Notably, when considering the entire image, the MST W/o-Guide model achieved the best performance in location estimation. This suggests a crossover point, which is further explored in the following analysis, where performance differences are influenced by the size of the positive region.

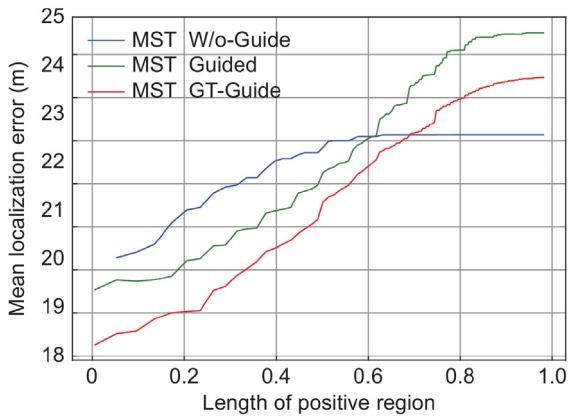
When tested on the VIGOR dataset, the CCVPE model was unable to predict locations in the positive region, with all predictions falling within the semi-positive region. In this dataset, the CVLocationTrans model outperformed the validation data in the positive and semi-positive regions. Overall, the FeatUp backbone demonstrated consistent performance across both datasets, being comparable to the CCVPE model for images from the Queens borough and to the CVLocationTrans model in the VIGOR dataset. Regarding the MST models, similar trends were observed when applied to the VIGOR dataset. The guided version showed improved performance in the positive region, with only a slight improvement, while the unguided version performed better when considering the entire image.

### 5.3.2. MST crossover point

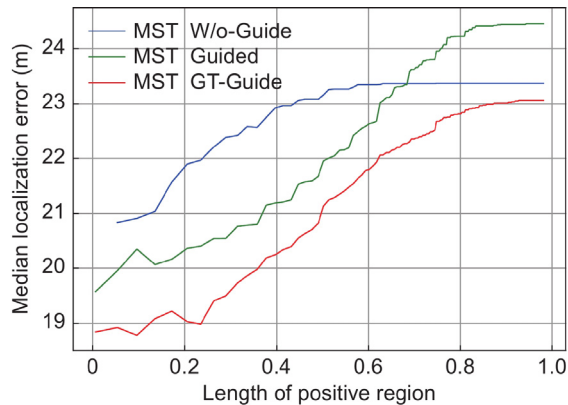
This section analyzes the crossover performance observed in the MST results, as highlighted in Table 3. As seen, the

**Table 3**  
Localization error on Queens borough and VIGOR cross-test set in meters. The best result is in bold, and the second best is underlined.

Scheme	Method	Queens set				VIGOR (Cross test set)			
		Positive		Pos. +Semi-Pos.		Positive		Pos. +Semi-Pos.	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
CCVPE	-	<b>18.60</b>	<b>17.37</b>	29.98	28.42	X	X	41.44	43.34
CVLocationTrans	-	25.89	25.34	35.92	36.37	31.13	29.64	29.62	29.23
Ours (MST)	W/o-Guide	22.85	23.16	<b>23.14</b>	<b>23.37</b>	28.03	28.64	<u>28.14</u>	<u>28.75</u>
	Guided	22.26	21.96	<u>25.51</u>	<u>24.46</u>	<u>27.99</u>	<u>28.05</u>	32.59	31.61
	GT-Guide	21.59	21.13	24.49	23.07	-	-	-	-
Ours (FeatUp)	Guided	<u>20.67</u>	<u>19.70</u>	27.97	26.14	<b>23.49</b>	<b>21.77</b>	<b>28.21</b>	<b>25.28</b>
	GT-Guide	22.44	19.35	28.43	26.69	-	-	-	-



**Fig. 12.** Mean location error evolution based on the positive region length for Queens set.



**Fig. 13.** Median location error evolution based on the positive region length for Queens set.

guided method performs better in the positive region, while the unguided method outperforms when the semi-positive region is included.

Figs. 12 and 13 illustrate the relationship between the length of the positive region (on the x-axis) and the mean and median location errors (on the y-axis), respectively, for all the MST methods.

As observed, the crossover points for the mean localization error in the unguided method occur at around 60% for the MST Guided method and at 70% for the MST GT-Guide method, respectively. For the median error, the MST method using ground truth land cover segmentations consistently outperforms the other techniques. Meanwhile, the MST Guided and W/o-Guide methods exhibit the second lowest error when the positive region covers approximately 70% of the image area. For all guided

**Table 4**  
Radius-based localization accuracy in percentage. The best and second-best performances are indicated in bold and underlined, respectively.

Scheme	Method	Queens set			VIGOR (Cross test set)		
		1 m	5 m	10 m	1 m	5 m	10 m
CCVPE	-	<b>0.21</b>	<b>4.96</b>	<u>13.88</u>	0.05	0.94	3.64
CVLocationTrans	-	0.13	1.70	5.99	<b>0.18</b>	<b>3.60</b>	<u>12.09</u>
Ours (MST)	W/o-Guide	0.13	2.47	9.69	0.03	1.08	5.74
	Guided	0.11	3.09	11.72	0.10	2.13	7.93
	GT-Guide	0.14	3.50	12.87	-	-	-
Ours (FeatUp)	Guided	<u>0.18</u>	<u>4.60</u>	<b>14.21</b>	<b>0.50</b>	<b>8.33</b>	<b>20.72</b>
	GT-Guide	0.20	4.71	14.45	-	-	-

techniques, as the positive region length increases, the errors also increase. In contrast, the error for the MST W/o-Guide method remains stable once the positive region exceeds 60% of the image area.

This suggests that using land cover guidance is always beneficial for localizing the central region, with performance improving as the semantic segmentation maps approach the ground truth land cover segmentations.

### 5.3.3. Radius-based location accuracy

We also present the accuracy of the ground-truth location within a fixed radius centered on the maximum probability location in Table 4. The accuracy is calculated considering constant radiuses of 1, 5, and 10 m.

The accuracy achieved with a constant radius demonstrates the effective performance of the CCVPE model on the Queens borough images, closely aligning with the predicted maximum probability location. Although the overall accuracy is relatively low, this pattern is particularly evident at a 1-meter radius. As the radius increases, accuracy improves, reaching up to 20% at a 10-meter distance. Notably, our method using the FeatUp backbone is comparable to CCVPE for the validation data, presenting a close accuracy for all radii. Despite the crossover observed earlier, the MST accuracy remains unaffected, with the MST W/o-Guide method showing low accuracy, while the guided method demonstrates a modest improvement.

For the VIGOR dataset, the performance of the FeatUp backbone is significantly more accurate than that of CVLocation, despite the similar location errors observed in the previous analysis.

### 5.3.4. Top-ranked regions comparison

The heatmap solution provides a flexible approach for sensor fusion systems by visualizing probabilistic regions and assessing the reliability of each area. The estimated regions are expected to be closed and concave, capturing complex, variable boundaries and fluctuating probabilities. This analysis is relevant when computing environmental features and evaluating the reliability of cross-view geolocation estimations. Table 5 presents the top-ranked regions of the probabilities estimated where  $3\sigma$

**Table 5**

Top-ranked region accuracy in percentage. The best score is in bold, and the second-best is underlined.

Scheme	Method	Queens set			VIGOR (Cross test set)		
		$3\sigma$	$2\sigma$	$1\sigma$	$3\sigma$	$2\sigma$	$1\sigma$
CCVPE	–	<b>44.97</b>	<b>94.48</b>	<b>100.00</b>	15.23	67.15	74.99
Ours (MST)	W/o-Guide	20.20	60.39	94.25	18.26	56.34	91.45
	Guided	24.67	78.34	99.46	<u>18.93</u>	<u>68.64</u>	<u>98.48</u>
	GT-Guide	17.54	78.58	99.38	–	–	–
Ours (FeatUp)	Guided	<u>38.24</u>	<u>87.00</u>	<u>99.93</u>	<b>29.84</b>	<b>80.26</b>	<b>99.76</b>
	GT-Guide	39.43	88.87	99.96	–	–	–

denotes the accuracy of areas of higher reliability and  $1\sigma$  indicates the accuracy of larger areas that also include lower probabilities. The CVLocationTrans performance is not presented in this table, since this method focuses on the end-to-end solution given in the maximum probability location.

The results exhibit a larger accuracy of the CCVPE at the top-ranked probabilities estimated regions in accordance with previous analysis for the Brooklyn and Queens dataset. This indicates more ground-view positions are located inside the boundaries regions defined. Additionally, the accuracy between the two best-performing methods varies by less than 10% in the validation data. Combining the analyses from Tables 4 and 5, we observe the CCVPE performance suggests a larger or sparse DPD compared to the other proposed models, even for the  $3\sigma$  parameter, which would imply a radius of over 10 meters from the estimated maximum probability location if the regions were more concise.

The performance of the CCVPE model on the VIGOR dataset was compromised due to its inability to estimate locations in the positive region. In contrast, the proposed method demonstrated consistent accuracy compared to the Queens set.

### 5.3.5. Qualitative comparison

This section qualitatively examines the performance differences among the various methods. Figs. 14 and 15 display samples of the position prediction of each model trained in the Queens and VIGOR datasets, respectively. The qualitative comparison includes ground and satellite images, the maximum probability location for each studied technique, and the top-ranked regions predicted by the CCVPE and the proposed methods. This analysis helps evaluate insights into the cross-view geolocalization estimation and the DPD consistency. In this section, we focus on the MST Guided and FeatUp Guided methods from our proposed solutions.

In the evaluation of the Queens borough, we observe that all techniques are able to predict the ground view image locations using the cross-view methodology. Generally, they tend to predict locations along or near the streets, since most images were captured outdoors. As observed, the predicted DPD varies across approaches. The CCVPE method can present a  $1\sigma$  region that encompasses the entire image. As anticipated, the regions are larger, contributing to higher top-ranked accuracies. In contrast, the proposed method using MST creates regions that contour buildings and emphasize streets and pathways. The DPD contours may present some discontinuity relative to the semantic segmentation map. Finally, the proposed method with the FeatUp backbone shows similar performance, focusing on routes and pathways. It also contours building, but less aggressively compared to the MST approach.

Evaluating the performances in the VIGOR dataset enhances our understanding of the model's generalization capacities across diverse datasets, consecutively, the estimation performance in

**Table 6**

Mean intersection over union (mIoU) expressed as a percentage for pixel-wise classification. Best score is in bold.

Scheme	Tree Canopy	Buildings	Roads	Other Impervious
Ours (MST)	<b>56.86</b>	<b>67.94</b>	<b>70.54</b>	<b>52.90</b>
Ours (FeatUp)	51.98	65.60	63.41	47.45

different cities, regions, and environments. In this analysis, the CCVPE's  $1\sigma$  top-rank region keeps covering the entire image, while the  $2\sigma$  top-rank region displays sparse peaks along the image borders. Notably, these sparse peaks were rarely observed in the Queens set.

Although the proposed methods continue to avoid constructions, they demonstrate reduced performance, indicating a decline in semantic segmentation capacity on the VIGOR dataset. This trend is primarily evident in the  $1\sigma$  top-rank region, whereas other top-ranked regions remain closely aligned with the streets. The MST model exhibits a significant decline in performance, as its  $1\sigma$  and  $2\sigma$  top-rank regions are no longer as concise as previously observed. In contrast, the proposed method using the FeatUp backbone is less affected by the change in the dataset.

Lastly, in both evaluations, the top-ranked regions presented distinct characteristics across approaches. The DPD regions of FeatUp are predominantly rounded, while the regions from MST and CCVPE tend to be more squared, with the latter having a softer appearance.

### 5.3.6. Semantic segmentation comparison

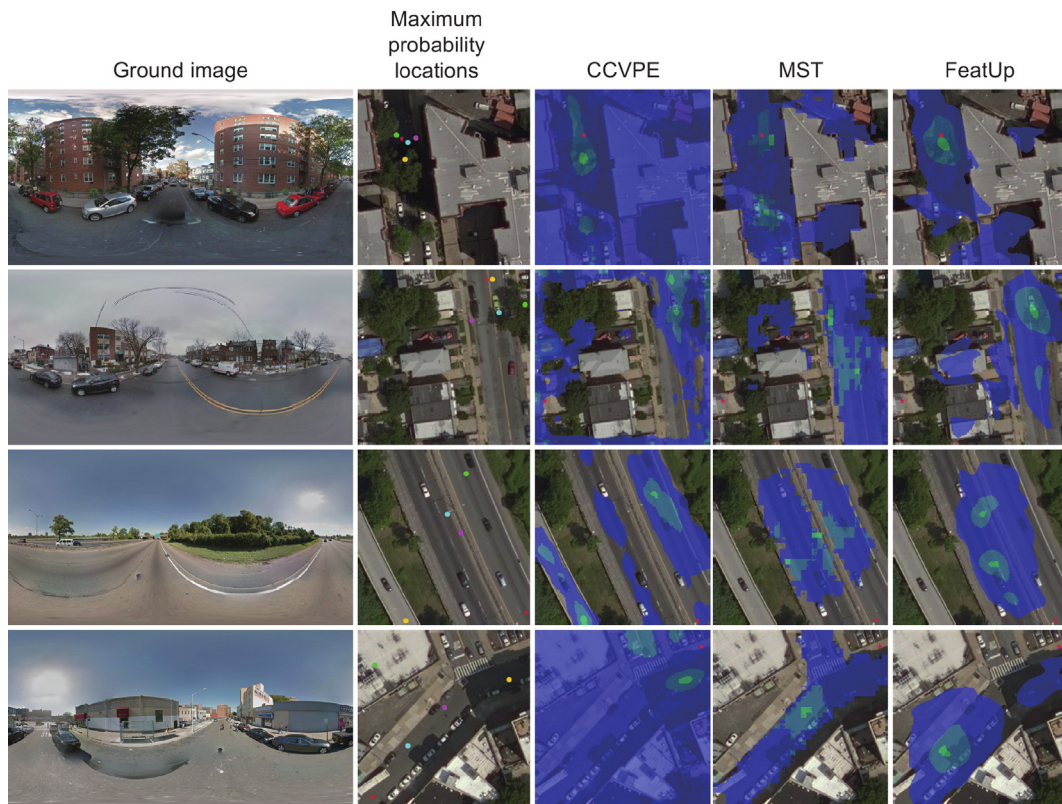
The final analysis we present focuses on the aerial semantic segmentation performance estimated from the first branch of the proposed model. We expect the predicted classes from the aerial image to align with the ground truth land cover labels. The accuracy of the segmented map is crucial for the proposed aided cross-view geolocalization although it is a secondary objective. Table 6 displays the semantic segmentation mIoU by class.

The classification of roads and buildings achieved the best performance, as these are the most important classes for guidance and sightline comparison. Overall, the MST model demonstrated a higher mIoU performance than the FeatUp backbone, even in light of its location performance.

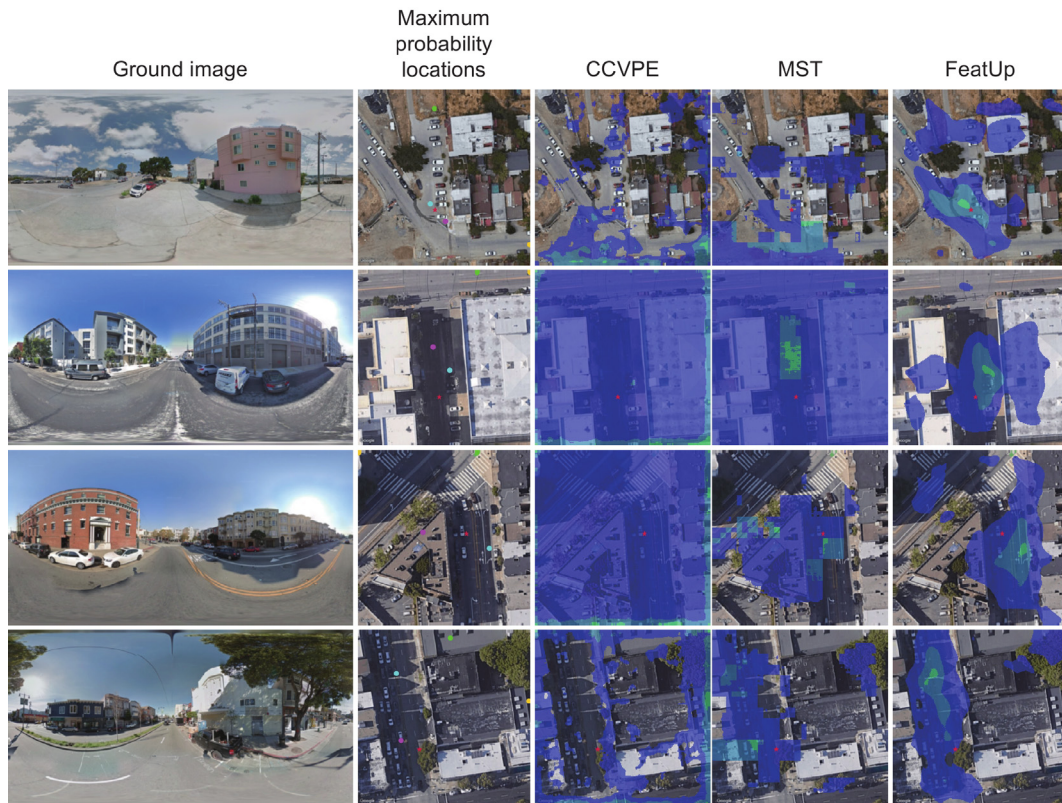
We use the same qualitative samples to provide a visual performance comparison of the semantic segmentation maps predicted by the proposed methods. In Fig. 16 we show the performance in Queens borough images, while Fig. 17 illustrates the performance on the VIGOR dataset.

The overall performance for semantic segmentation is similar between the proposed methods. The segmented maps exhibit the same characteristics of rounded and squared formats observed in the top-ranked regions. However, the segmentation map from the MST model shows noisy classifications along the edges of objects or between different classes. In contrast, the FeatUp model does not exhibit this issue, though it presents more discontinuous class boundaries. We highlight some incoherent classifications, particularly in the tree canopy class, which are evident when analyzing the aerial images and land cover segmentation.

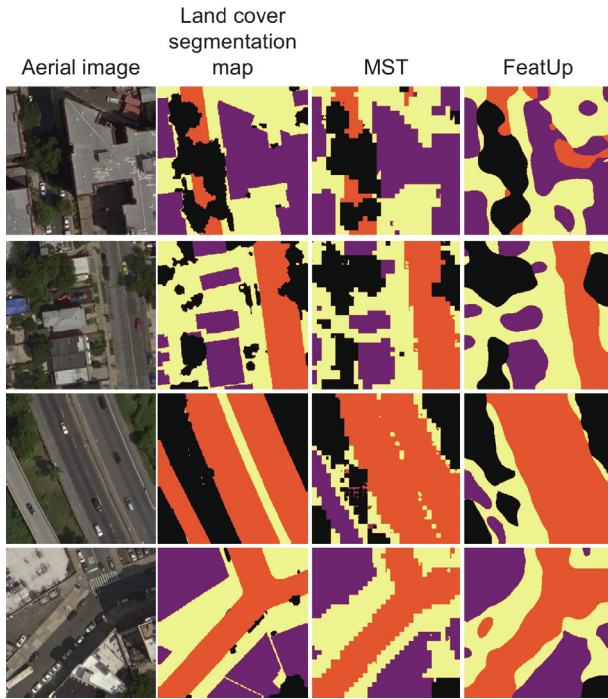
As observed in Fig. 17, the semantic segmentation performance of the proposed method is lower compared to the previous dataset. The MST model struggles significantly with classifying each class from the satellite imagery. Although the FeatUp backbone demonstrates better performance, it still falls short overall. Notably, while the building and road classes are generally well positioned, they lack continuity in the segmentation results



**Fig. 14.** Qualitative comparison in Queen's set. Each row presents, from left to right: (1) the ground image whose position needs to be estimated; (2) the satellite image tagged with the maximum probability locations estimated from the CCVPE (gold), CVLocationTrans (green), MST (pink), and FeatUp (cyan) and the ground truth (red), and the (3) CCVPE; (4) MST, and; (5) FeatUp top-ranked regions as  $1\sigma$  (blue),  $2\sigma$  (dark cyan), and  $3\sigma$  (green).



**Fig. 15.** Qualitative comparison in VIGOR cross-test set. Each row presents, from left to right: (1) the ground image whose position needs to be estimated; (2) the satellite image tagged with the maximum probability locations estimated from the CCVPE (gold), CVLocationTrans (green), MST (pink), and FeatUp (cyan) and the ground truth (red), and the (3) CCVPE; (4) MST, and; (5) FeatUp top-ranked regions as  $1\sigma$  (blue),  $2\sigma$  (dark cyan), and  $3\sigma$  (green).



**Fig. 16.** Qualitative comparison in Queens set. Each row presents, from left to right: (1) the satellite image; (2) the ground truth semantic segmentation map; (3) MST estimation, and; (4) FeatUp estimation. Each color in the semantic segmentation map represents one class, where tree canopy (black), buildings (purple), roads (orange), and other impervious (yellow).

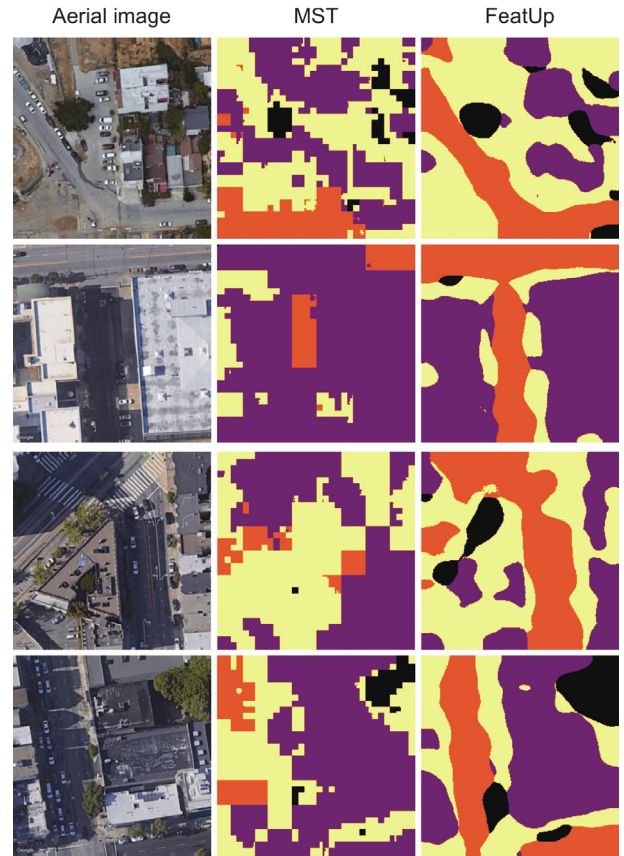
## 6. Conclusion

In this paper, we explored cross-view geolocation estimation using ground view and satellite imagery. We proposed a method that utilizes a land cover semantic segmentation map derived solely from the overview photo, primarily identifying tree canopies, buildings, and roads. This predicted overhead semantic segmentation map serves as guidance for cross-view geolocation estimation, enhancing location accuracy by filtering out transient and movable objects while emphasizing environmental features.

We utilized the Brooklyn and Queens [55] and the VIGOR [29] datasets for this task. The Brooklyn and Queens dataset provided the training and validation data, while the VIGOR was used in the test stage. Both datasets include satellite images and 360°street-view images are available, however, only the Brooklyn and Queens dataset has the aerial segmented maps. Intending to create a more flexible solution, the proposed models also predict the semantic segmentation map, making it more comparable and applicable to any cross-view dataset for location estimation. Additionally, by combining aerial and land cover images, we extract the sightline map, which more consistently describes the ground view location and perspective.

The results indicate that the proposed methods perform comparably to the state-of-the-art approaches for cross-view geolocation estimation in most conducted analyses. They demonstrate consistent localization performance in both positive and semi-positive regions and when switching datasets. Furthermore, the proposed methods yield top-ranked DPD regions confined to streets and pathways.

Most evaluations focused on analyzing the performance of land cover segmentation, showing increased accuracy with the guided methods, while revealing the maximum performance



**Fig. 17.** Qualitative comparison in VIGOR dataset. Each row presents, from left to right: (1) the satellite image; (2) the MST semantic segmentation map estimated, and; (3) FeatUp estimation. Each color in the semantic segmentation map represents one class, where tree canopy (black), buildings (purple), roads (orange), and other impervious (yellow).

**Table A.1**

FeatUp semantic segmentation filtering architecture.

Layer	Type	Input shape	Output shape	Param #
1	Conv2d	[None,384,256,256]	[None,1024,256,256]	394,240
2	LayerNorm	[None,1024,256,256]	[None,1024,256,256]	134,217,728
3	GELU	[None,1024,256,256]	[None,1024,256,256]	0
4	Conv2d	[None,1024,256,256]	[None,4,256,256]	4,100
5	Dropout	[None,4,256,256]	[None,4,256,256]	0
6	Sigmoid	[None,4,256,256]	[None,4,256,256]	0

**Table A.2**

FeatUp heatmap prediction architecture.

Layer	Type	Input shape	Output shape	Param #
1	Conv2d	[None,384,256,256]	[None,1024,256,256]	394,240
2	Conv2d	[None,384,128,512]	[None,1024,128,512]	394,240
3	AdaptiveAvgPool2d	[None,1024,128,512]	[None,1024,1,1]	0
4	CosineSimilarity	Layers 1 and 3	[None,256,256]	0
5	Conv2d	[None,1,256,256]	[None,10,256,256]	60
6	LayerNorm	[None,10,256,256]	[None,10,256,256]	1,310,720
7	GELU	[None,10,256,256]	[None,10,256,256]	0
8	Conv2d	[None,10,256,256]	[None,1,256,256]	11
9	Dropout	[None,1,256,256]	[None,1,256,256]	0
10	Sigmoid	[None,1,256,256]	[None,1,256,256]	0

when using the ground truth semantic segmentation map as guidance. This highlights the significance of the proposed methodology in employing the semantic segmentation map for cross-view geolocation.

Finally, the proposed methods properly classify the image objects, where the predicted aerial semantic segmentation map

**Table A.3**  
MST semantic segmentation filtering architecture.

Layer	Type	Input shape	Output shape	Param #
1	Conv2d	[None,786,8,8]	[None,1024,8,8]	787,456
2	LayerNorm	[None,1024,8,8]	[None,1024,8,8]	131,072
3	GELU	[None,1024,8,8]	[None,1024,8,8]	0
4	Conv2d	[None,1024,8,8]	[None,384,8,8]	393,600
5	Dropout	[None,384,8,8]	[None,384,8,8]	0
6	Upsample	[None,384,8,8]	[None,384,16,16]	0
7	Concat	[None,384,16,16] +Layer 6	[None,768,16,16]	0
8	Conv2d	[None,768,16,16]	[None,512,16,16]	393,728
9	LayerNorm	[None,512,16,16]	[None,512,16,16]	262,144
10	GELU	[None,512,16,16]	[None,512,16,16]	0
11	Conv2d	[None,512,16,16]	[None,192,16,16]	98,496
12	Dropout	[None,192,16,16]	[None,192,16,16]	0
13	Upsample	[None,192,16,16]	[None,192,32,32]	0
14	Concat	[None,192,32,32] +Layer 13	[None,384,32,32]	0
15	Conv2d	[None,384,32,32]	[None,256,32,32]	98,560
16	LayerNorm	[None,256,32,32]	[None,256,32,32]	524,288
17	GELU	[None,256,32,32]	[None,256,32,32]	0
18	Conv2d	[None,256,32,32]	[None,96,32,32]	24,672
19	Dropout	[None,96,32,32]	[None,96,32,32]	0
20	Upsample	[None,96,32,32]	[None,96,64,64]	0
21	Conv2d	[None,96,64,64]	[None,128,64,64]	12,416
22	LayerNorm	[None,128,64,64]	[None,128,64,64]	1,048,576
23	GELU	[None,128,64,64]	[None,128,64,64]	0
24	Conv2d	[None,128,64,64]	[None,96,64,64]	12,384
25	Dropout	[None,96,64,64]	[None,96,64,64]	0
26	Upsample	[None,96,64,64]	[None,96,128,128]	0
27	Conv2d	[None,96,128,128]	[None,128,128,128]	12,416
28	LayerNorm	[None,128,128,128]	[None,128,128,128]	4,194,304
29	GELU	[None,128,128,128]	[None,128,128,128]	0
30	Conv2d	[None,128,128,128]	[None,96,128,128]	12,384
31	Dropout	[None,96,128,128]	[None,96,128,128]	0
32	Upsample	[None,96,128,128]	[None,96,256,256]	0
33	Conv2d	[None,96,128,128]	[None,128,256,256]	12,416
34	LayerNorm	[None,128,256,256]	[None,128,256,256]	16,777,216
35	GELU	[None,128,256,256]	[None,128,256,256]	0
36	Conv2d	[None,128,256,256]	[None,4,256,256]	516
37	Dropout	[None,4,256,256]	[None,4,256,256]	0
38	Sigmoid-36	[None,4,256,256]	[None,4,256,256]	0

presented small noises in the edges and different edge formats. The performance of the semantic segmentation map in the VIGOR dataset shows a slight decrease, with some discontinuities and misclassifications. Conversely, the estimated DPD demonstrates consistent performance for the maximum probability location and the top-ranked regions.

In summary, this study contributes to the ongoing advancement of cross-view geolocalization estimation technologies, proposing a novel methodology for estimating and integrating land cover semantic segmentation in real-world applications. We also introduce a new pipeline for location estimation that utilizes a sightline map derived from the aerial land cover map, combined with a heatmap to predict locations effectively.

While this study has demonstrated significant advancements in cross-view geolocalization estimation, there are still some perspectives for research continuity, including (a) integration with additional data sources, such as light detection and ranging (LiDAR) and inertial sensors; (b) incorporation of aerial oblique images; (c) exploration of seasonal conditions and more diverse environments; (d) analysis of real-time performance, and; (e) development of an ablation study on training hyperparameters.

**CRedit authorship contribution statement**

**Nathan A.Z. Xavier:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Elcio H. Shiguemori:** Writing – review & editing, Supervision. **Marcos R.O.A. Maximo:** Writing – review &

**Table A.4**  
MST heatmap prediction architecture.

Layer	Type	Input shape	Output shape	Param #
1	Conv2d	[None,786,8,8]	[None,1024,8,8]	787,456
2	Conv2d	[None,786]	[None,1024,1,1]	787,456
3	CosineSimilarity	Layers 1 and 2	[None,8,8]	0
4	Conv2d	[None,1,8,8]	[None,10,8,8]	20
5	LayerNorm	[None,10,8,8]	[None,10,8,8]	1,280
6	GELU	[None,10,8,8]	[None,10,8,8]	0
7	Conv2d	[None,10,8,8]	[None,1,8,8]	11
8	Dropout	[None,1,8,8]	[None,1,8,8]	0
9	Upsample	[None,1,8,8]	[None,1,16,16]	0
10	Conv2d	[None,384,16,16]	[None,1024,16,16]	394,240
11	Conv2d	[None,384,8,32]	[None,1024,8,32]	394,240
12	AdaptiveAvgPool2d	[None,1024,8,32]	[None,1024,1,1]	0
13	CosineSimilarity	Layers 10 and 12	[None,16,16]	0
14	Concat	Layers 9 and 13	[None,2,16,16]	0
15	Conv2d	[None,2,16,16]	[None,10,16,16]	30
16	LayerNorm	[None,10,16,16]	[None,10,16,16]	5,120
17	GELU	[None,10,16,16]	[None,10,16,16]	0
18	Conv2d	[None,10,16,16]	[None,1,16,16]	11
19	Dropout	[None,1,16,16]	[None,1,16,16]	0
20	Upsample	[None,1,16,16]	[None,1,32,32]	0
21	Conv2d	[None,192,32,32]	[None,1024,32,32]	197,632
22	Conv2d	[None,192,16,64]	[None,1024,16,64]	197,632
23	AdaptiveAvgPool2d	[None,1024,16,64]	[None,1024,1,1]	0
24	CosineSimilarity	Layers 21 and 23	[None,32,32]	0
25	Concat	Layers 20 and 24	[None,2,32,32]	0
26	Conv2d	[None,2,32,32]	[None,10,32,32]	30
27	LayerNorm	[None,10,32,32]	[None,10,32,32]	20,480
28	GELU	[None,10,32,32]	[None,10,32,32]	0
29	Conv2d	[None,10,32,32]	[None,1,32,32]	11
30	Dropout	[None,1,32,32]	[None,1,32,32]	0
31	Upsample	[None,1,32,32]	[None,1,64,64]	0
32	Conv2d	[None,1,64,64]	[None,10,64,64]	20
33	LayerNorm	[None,10,64,64]	[None,10,64,64]	81,920
34	GELU	[None,10,64,64]	[None,10,64,64]	0
35	Conv2d	[None,10,64,64]	[None,1,64,64]	11
36	Dropout	[None,1,64,64]	[None,1,64,64]	0
37	Upsample	[None,1,64,64]	[None,1,128,128]	0
38	Conv2d	[None,1,128,128]	[None,10,128,128]	20
39	LayerNorm	[None,10,128,128]	[None,10,128,128]	327,680
40	GELU	[None,10,128,128]	[None,10,128,128]	0
41	Conv2d	[None,10,128,128]	[None,1,128,128]	11
42	Dropout	[None,1,128,128]	[None,1,128,128]	0
43	Upsample	[None,1,128,128]	[None,1,256,256]	0
44	Conv2d	[None,1,128,128]	[None,10,256,256]	60
45	LayerNorm	[None,10,256,256]	[None,10,256,256]	1,310,720
46	GELU	[None,10,256,256]	[None,10,256,256]	0
47	Conv2d	[None,10,256,256]	[None,1,256,256]	11
48	Dropout	[None,1,256,256]	[None,1,256,256]	0
49	Sigmoid	[None,1,256,256]	[None,1,256,256]	0

editing, Supervision. **Mubarak Shah:** Writing – review & editing, Supervision, Resources, Project administration.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgments**

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) (88887.929508/2023-00 and 88887.937224/2024-00).

Marcos Maximo is partially funded by the National Research Council of Brazil (CNPq) (307525/2022-8).

This research was developed within the IDeepS which is supported by the Laboratório Nacional de Computação Científica (LNCC/MCTI, Brazil) via resources of the SDumont supercomputer (<http://sdumont.lncc.br>).

## Appendix. Supplementary material

This appendix contains supplemental material that provides additional details on the neural network architecture described in Section 4.

Tables A.1 and A.2 present the architecture for the segmentation and heatmap prediction blocks, respectively, within the FeatUp backbone model. Similarly, Tables A.3 and A.4 provide the corresponding architectural details for the MST model.

Each of these tables includes information on the layer type, input and output shapes, and the number of learnable parameters.

## References

- [1] C. Hegarty, E. Chatre, Evolution of the Global Navigation Satellite System (GNSS), *Proc. IEEE* 96 (12) (2008) 1902–1917, <http://dx.doi.org/10.1109/JPROC.2008.2006090>.
- [2] S. Cobb, D. Lawrence, J. Christie, T. Walter, Y. Chao, D. Powell, B. Parkinson, Observed GPS signal continuity interruptions, in: *Proceedings of Ion GPS, vol. 8, ION, INSTITUTE OF NAVIGATION, California, EUA, 1995*, pp. 793–795.
- [3] E.L. Afraimovich, O.S. Lesyuta, I.I. Ushakov, Magnetospheric disturbances, and the GPS operation, 2000, <http://dx.doi.org/10.48550/ARXIV.PHYSICS/0009027>, arXiv, Online.
- [4] Y. Xia, M. Song, J. Zhang, C. Hu, An autonomously navigation system for forestry quadrotor within GPS-denied below-canopy environment, in: 2018 IEEE CSAA Guidance, Navigation and Control Conference, CGNCC, IEEE, Xiamen, China, 2018, pp. 1–6, <http://dx.doi.org/10.1109/GNCC42960.2018.9019136>.
- [5] A.C.B. Chiella, B.O.S. Teixeira, G.A.S. Pereira, State estimation for aerial vehicles in forest environments, in: 2019 International Conference on Unmanned Aircraft Systems, ICUAS, IEEE, Atlanta, EUA, 2019, pp. 890–898, <http://dx.doi.org/10.1109/ICUAS.2019.8797822>.
- [6] M.S. Allauddin, G.S. Kiran, G.R. Kiran, G. Srinivas, G.U.R. Mouli, P.V. Prasad, Development of a surveillance system for forest fire detection and monitoring using drones, in: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2019, pp. 9361–9363, <http://dx.doi.org/10.1109/IGARSS.2019.8900436>.
- [7] V.A. Torres, B.R. Jaimes, E.S. Ribeiro, M.T. Braga, E.H. Shiguemori, H.F. Velho, L.C. Torres, A.P. Braga, Combined weightless neural network FPGA architecture for deforestation surveillance and visual navigation of UAVs, *Eng. Appl. Artif. Intell.* 87 (2020) 103227, <http://dx.doi.org/10.1016/j.engappai.2019.08.021>.
- [8] D.R. Alves de Almeida, E. Broadbent, A.M. Almeyda Zambrano, M.P. Ferreira, P.H. Santin Brancalion, Fusion of lidar and hyperspectral data from drones for ecological questions: The gatereye atlantic forest restoration case study, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, Online, 2021, pp. 714–715, <http://dx.doi.org/10.1109/IGARSS47720.2021.9554023>.
- [9] Y. Xu, Y. Wei, D. Wang, K. Jiang, H. Deng, Multi-UAV path planning in GPS and communication denial environment, *Sensors* 23 (6) (2023) 2997, <http://dx.doi.org/10.3390/S23062997>.
- [10] J. Zeil, Visual navigation: properties, acquisition and use of views, *J. Comp. Physiol. A* 209 (4) (2022) 499–514, <http://dx.doi.org/10.1007/S00359-022-01599-2>.
- [11] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, S. Levine, ViNT: A foundation model for visual navigation, 2023, <http://dx.doi.org/10.48550/ARXIV.2306.14846>, arXiv.
- [12] G. Dimas, D.E. Diamantis, P. Kalozoumis, D.K. Iakovidis, Uncertainty-aware visual perception system for outdoor navigation of the visually challenged, *Sensors* 20 (8) (2020) 2385, <http://dx.doi.org/10.3390/S20082385>.
- [13] A. Sivakumar, S. Modi, M. Gasparino, C. Ellis, A. Baquero Velasquez, G. Chowdhary, S. Gupta, Learned visual navigation for under-canopy agricultural robots, in: *Robotics: Science and Systems XVII*, in: RSS2021, Robotics: Science and Systems Foundation, 2021, <http://dx.doi.org/10.15607/RSS.2021.XVII.019>.
- [14] B. Fahima, N. Abdelkrim, Multispectral visual odometry using SVSF for mobile robot localization, *Unmanned Syst.* 10 (03) (2021) 273–288, <http://dx.doi.org/10.1142/S2301385022500157>.
- [15] J. Truong, A. Zitkovich, S. Chernova, D. Batra, T. Zhang, J. Tan, W. Yu, IndoorSim-to-OutdoorReal: Learning to navigate outdoors without any outdoor experience, *IEEE Robot. Autom. Lett.* 9 (5) (2024) 4798–4805, <http://dx.doi.org/10.1109/LRA.2024.3385611>.
- [16] M. Voodarla, S. Shrivastava, S. Manglani, A. Vora, S. Agarwal, P. Chakravarty, S-BEV: Semantic birds-eye view representation for weather and lighting invariant 3-DoF localization, 2021, <http://dx.doi.org/10.48550/ARXIV.2101.09569>, arXiv.
- [17] C.-J. Chen, Y.-Y. Huang, Y.-S. Li, Y.-C. Chen, C.-Y. Chang, Y.-M. Huang, Identification of fruit tree pests with deep learning on embedded drone to achieve accurate pesticide spraying, *IEEE Access* 9 (2021) 21986–21997, <http://dx.doi.org/10.1109/ACCESS.2021.3056082>.
- [18] A.B. Camiletto, A. Bochicchio, A. Liniger, D. Dai, A. Gawel, U-BEV: Height-aware bird's-eye-view segmentation and neural map-based relocalization, 2023, <http://dx.doi.org/10.48550/ARXIV.2310.13766>, arXiv.
- [19] J. Luo, Q. Ye, UAV large oblique image geo-localization using satellite images in the dense buildings area, *ISPRS Ann. Photogramm., Remote. Sens. Spatial Inf. Sci.* X-1/W1-2023 (2023) 1065–1072, <http://dx.doi.org/10.5194/ISPRS-ANNALS-X-1-W1-2023-1065-2023>.
- [20] J. Ye, Q. Luo, J. Yu, H. Zhong, Z. Zheng, C. He, W. Li, SG-BEV: Satellite-guided BEV fusion for cross-view semantic segmentation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2024, pp. 27748–27757, <http://dx.doi.org/10.1109/CVPR52733.2024.02621>.
- [21] L. Liu, H. Li, Lending orientation to neural networks for cross-view geolocalization, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2019, pp. 5617–5626, <http://dx.doi.org/10.1109/CVPR.2019.00577>.
- [22] S. Zhu, T. Yang, C. Chen, Revisiting street-to-aerial view image geolocalization and orientation estimation, in: 2021 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2021, pp. 756–765, <http://dx.doi.org/10.1109/WACV48630.2021.00080>.
- [23] F. Ge, Y. Zhang, Y. Liu, G. Wang, S. Coleman, D. Kerr, L. Wang, Multibranch joint representation learning based on information fusion strategy for cross-view geo-localization, *IEEE Trans. Geosci. Remote Sens.* 62 (2024) 1–16, <http://dx.doi.org/10.1109/TGRS.2024.3378453>.
- [24] S. Zhu, M. Shah, C. Chen, TransGeo: Transformer is all you need for cross-view image geo-localization, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2022, pp. 1152–1161, <http://dx.doi.org/10.1109/CVPR52688.2022.00123>.
- [25] F. Fervers, S. Bullinger, C. Bodensteiner, M. Arens, R. Stiefelhagen, Uncertainty-aware vision-based metric cross-view geolocalization, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2023, pp. 21621–21631, <http://dx.doi.org/10.1109/CVPR52729.2023.02071>.
- [26] F. Fervers, S. Bullinger, C. Bodensteiner, M. Arens, R. Stiefelhagen, C-BEV: Contrastive bird's eye view training for cross-view image retrieval and 3-DoF pose estimation, 2023, <http://dx.doi.org/10.48550/ARXIV.2312.08060>, arXiv.
- [27] Z. Xia, O. Booi, J.F.P. Kooij, Convolutional cross-view pose estimation, *IEEE Trans. Pattern Anal. Mach. Intell.* 46 (5) (2024) 3813–3831, <http://dx.doi.org/10.1109/TPAMI.2023.3346924>.
- [28] S. Workman, R. Souvenir, N. Jacobs, Wide-area image geolocalization with aerial reference imagery, in: 2015 IEEE International Conference on Computer Vision, ICCV, IEEE, 2015, <http://dx.doi.org/10.1109/ICCV.2015.451>.
- [29] S. Zhu, T. Yang, C. Chen, VIGOR: Cross-view image geo-localization beyond one-to-one retrieval, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2021, <http://dx.doi.org/10.1109/CVPR46437.2021.00364>.
- [30] N.N. Vo, J. Hays, Localizing and orienting street views using overhead imagery, in: *Lecture Notes in Computer Science*, Springer International Publishing, 2016, pp. 494–509, [http://dx.doi.org/10.1007/978-3-319-46448-0\\_30](http://dx.doi.org/10.1007/978-3-319-46448-0_30).
- [31] Y. Tian, C. Chen, M. Shah, Cross-view image matching for geo-localization in urban environments, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2017, pp. 1998–2006, <http://dx.doi.org/10.1109/CVPR.2017.216>.
- [32] R. Cao, J. Zhu, W. Tu, Q. Li, J. Cao, B. Liu, Q. Zhang, G. Qiu, Integrating aerial and street view images for urban land use classification, *Remote Sens.* 10 (10) (2018) 1553, <http://dx.doi.org/10.3390/RS10101553>.
- [33] Y. Shi, X. Yu, L. Liu, T. Zhang, H. Li, Optimal feature transport for cross-view image geo-localization, *Proc. AAAI Conf. Artif. Intell.* 34 (07) (2020) 11990–11997, <http://dx.doi.org/10.1609/AAAI.V34I07.6875>.
- [34] Y. Shi, X. Yu, D. Campbell, H. Li, Where am I looking at? Joint location and orientation estimation by cross-view matching, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2020, pp. 4063–4071, <http://dx.doi.org/10.1109/CVPR42600.2020.00412>.
- [35] Y. Zhu, B. Sun, X. Lu, S. Jia, Geographic semantic network for cross-view image geo-localization, *IEEE Trans. Geosci. Remote Sens.* 60 (2022) 1–15, <http://dx.doi.org/10.1109/TGRS.2021.3121337>.
- [36] Y. Shi, X. Yu, L. Liu, D. Campbell, P. Koniusz, H. Li, Accurate 3-DoF camera geo-localization via ground-to-satellite image matching, *IEEE Trans. Pattern Anal. Mach. Intell.* (2022) 1–16, <http://dx.doi.org/10.1109/TPAMI.2022.3189702>.
- [37] T. Wang, S. Fan, D. Liu, C. Sun, Transformer-guided convolutional neural network for cross-view geolocalization, 2022, <http://dx.doi.org/10.48550/ARXIV.2204.09967>, arXiv.

- [38] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, P.M. Atkinson, UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, *ISPRS J. Photogramm. Remote Sens.* 190 (2022) 196–214, <http://dx.doi.org/10.1016/j.isprsjprs.2022.06.008>.
- [39] J. Zhao, Q. Zhai, P. Zhao, R. Huang, H. Cheng, Co-visual pattern-augmented generative transformer learning for automobile geo-localization, *Remote Sens.* 15 (9) (2023) 2221, <http://dx.doi.org/10.3390/RS15092221>.
- [40] Y. Shi, F. Wu, A. Perincherri, A. Vora, H. Li, Boosting 3-DoF ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, 2023, pp. 21459–21469, <http://dx.doi.org/10.1109/ICCV51070.2023.01967>.
- [41] K. Regmi, A. Borji, Cross-view image synthesis using geometry-guided conditional GANs, *Comput. Vis. Image Underst.* 187 (2019) 102788, <http://dx.doi.org/10.1016/j.cviu.2019.07.008>.
- [42] K. Regmi, M. Shah, Bridging the domain gap for ground-to-aerial image matching, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, 2019, pp. 470–479, <http://dx.doi.org/10.1109/ICCV.2019.00056>.
- [43] A. Toker, Q. Zhou, M. Maximov, L. Leal-Taixe, Coming down to earth: Satellite-to-street view synthesis for geo-localization, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2021, <http://dx.doi.org/10.1109/CVPR46437.2021.00642>.
- [44] S. Wu, H. Tang, X.-Y. Jing, J. Qian, N. Sebe, Y. Yan, Q. Zhang, Cross-view panorama image synthesis with progressive attention GANs, *Pattern Recognit.* 131 (2022) 108884, <http://dx.doi.org/10.1016/j.patcog.2022.108884>.
- [45] A. Durgam, S. Paheding, V. Dhiman, V. Devabhaktuni, Cross-view geo-localization: a survey, 2024, <http://dx.doi.org/10.48550/ARXIV.2406.09722>, arXiv.
- [46] Y. Zhuang, X. Sun, Y. Li, J. Huai, L. Hua, X. Yang, X. Cao, P. Zhang, Y. Cao, L. Qi, J. Yang, N. El-Bendary, N. El-Sheimy, J. Thompson, R. Chen, Multi-sensor integrated navigation/positioning systems using data fusion: From analytics-based to learning-based approaches, *Inf. Fusion* 95 (2023) 62–90, <http://dx.doi.org/10.1016/j.inffus.2023.01.025>.
- [47] M. Gadd, P. Newman, Checkout my map: Version control for fleetwide visual localisation, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2016, pp. 5729–5736, <http://dx.doi.org/10.1109/IROS.2016.7759843>.
- [48] R. Rodrigues, M. Tani, SemGeo: Semantic keywords for cross-view image geo-localization, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5, <http://dx.doi.org/10.1109/ICASSP49357.2023.10094763>.
- [49] V. Balaska, L. Bampis, A. Gasteratos, Self-localization based on terrestrial and satellite semantics, *Eng. Appl. Artif. Intell.* 111 (2022) 104824, <http://dx.doi.org/10.1016/j.engappai.2022.104824>.
- [50] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, Understanding convolution for semantic segmentation, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2018, <http://dx.doi.org/10.1109/WACV.2018.00163>.
- [51] S. Kluckner, T. Mauthner, P.M. Roth, H. Bischof, Semantic classification in aerial imagery by integrating appearance and height information, in: *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2010, pp. 477–488, [http://dx.doi.org/10.1007/978-3-642-12304-7\\_45](http://dx.doi.org/10.1007/978-3-642-12304-7_45).
- [52] S. Hu, G.H. Lee, Image-based geo-localization using satellite imagery, *Int. J. Comput. Vis.* 128 (5) (2019) 1205–1219, <http://dx.doi.org/10.1007/S11263-019-01186-0>.
- [53] M. Elhashash, R. Qin, Cross-view SLAM solver: Global pose estimation of monocular ground-level video frames for 3D reconstruction using a reference 3D model from satellite images, *ISPRS J. Photogramm. Remote Sens.* 188 (2022) 62–74, <http://dx.doi.org/10.1016/j.isprsjprs.2022.03.018>.
- [54] Y. Zhang, Y. Shi, S. Wang, A. Vora, A. Perincherri, Y. Chen, H. Li, Increasing SLAM pose accuracy by ground-to-satellite image registration, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 8522–8528, <http://dx.doi.org/10.1109/ICRA57147.2024.10611079>.
- [55] S. Workman, M. Zhai, D.J. Crandall, N. Jacobs, A unified model for near and remote sensing, in: 2017 IEEE International Conference on Computer Vision, ICCV, IEEE, 2017, <http://dx.doi.org/10.1109/ICCV.2017.293>.
- [56] M. Zhai, Z. Bessinger, S. Workman, N. Jacobs, Predicting ground-level scene layout from aerial imagery, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2017, pp. 4132–4140, <http://dx.doi.org/10.1109/CVPR.2017.440>.
- [57] J. Dai, X. Hao, S. Liu, Z. Ren, Research on UAV robust adaptive positioning algorithm based on IMU/GNSS/VO in complex scenes, *Sensors* 22 (8) (2022) 2832, <http://dx.doi.org/10.3390/S22082832>.
- [58] O.L.F. de Carvalho, O.A. de Carvalho Júnior, C.R.e. Silva, A.O. de Albuquerque, N.C. Santana, D.L. Borges, R.A.T. Gomes, R.F. Guimarães, Panoptic segmentation meets remote sensing, *Remote Sens.* 14 (4) (2022) 965, <http://dx.doi.org/10.3390/RS14040965>.
- [59] B. Pan, J. Sun, H.Y.T. Leung, A. Andonian, B. Zhou, Cross-view semantic segmentation for sensing surroundings, *IEEE Robot. Autom. Lett.* 5 (3) (2020) 4867–4873, <http://dx.doi.org/10.1109/LRA.2020.3004325>.
- [60] B. Zhou, P. Krahenbuhl, Cross-view transformers for real-time map-view semantic segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2022, <http://dx.doi.org/10.1109/CVPR52688.2022.01339>.
- [61] G. Zhou, A. Liu, K. Yang, T. Wang, Z. Li, An embedded solution to visual mapping for consumer drones, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, Columbus, USA, 2014, pp. 670–675, <http://dx.doi.org/10.1109/CVPRW.2014.102>.
- [62] S. Workman, M.U. Rafique, H. Blanton, N. Jacobs, Revisiting near/remote sensing with geospatial attention, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2022, pp. 1768–1777, <http://dx.doi.org/10.1109/CVPR52688.2022.00182>.
- [63] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2022, <http://dx.doi.org/10.1109/CVPR52688.2022.00135>.
- [64] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, 2023, <http://dx.doi.org/10.48550/ARXIV.2304.02643>, arXiv.
- [65] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryal, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K.V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, C. Feichtenhofer, SAM 2: Segment anything in images and videos, 2024, <http://dx.doi.org/10.48550/ARXIV.2408.00714>, arXiv.
- [66] S. Fu, M. Hamilton, L. Brandt, A. Feldman, Z. Zhang, W.T. Freeman, FeatUp: A model-agnostic framework for features at any resolution, 2024, <http://dx.doi.org/10.48550/ARXIV.2403.10516>, arXiv.
- [67] J. Kopf, M.F. Cohen, D. Lischinski, M. Uyttendaele, Joint bilateral upsampling, *ACM Trans. Graph.* 26 (3) (2007) 96, <http://dx.doi.org/10.1145/1276377.1276497>.
- [68] Z. Cui, P. Zhou, X. Wang, Z. Zhang, Y. Li, H. Li, Y. Zhang, A novel geo-localization method for UAV and satellite images using cross-view consistent attention, *Remote Sens.* 15 (19) (2023) 4667, <http://dx.doi.org/10.3390/RS15194667>.
- [69] A. Shetty, G.X. Gao, UAV pose estimation using cross-view geolocation with satellite imagery, in: 2019 International Conference on Robotics and Automation, ICRA, IEEE, 2019, pp. 1827–1833, <http://dx.doi.org/10.1109/ICRA.2019.8794228>.
- [70] Z. Ye, C. Bao, X. Liu, H. Bao, Z. Cui, G. Zhang, Crossview mapping with graph-based geolocation on city-scale street maps, in: 2022 International Conference on Robotics and Automation, ICRA, IEEE, 2022, pp. 7980–7987, <http://dx.doi.org/10.1109/ICRA46639.2022.9811743>.
- [71] D. Wilson, X. Zhang, W. Sultani, S. Wshah, Image and object geo-localization, *Int. J. Comput. Vis.* 132 (4) (2023) 1350–1392, <http://dx.doi.org/10.1007/S11263-023-01942-3>.
- [72] X. Zhang, X. Li, W. Sultani, Y. Zhou, S. Wshah, Cross-view geo-localization via learning disentangled geometric layout correspondence, *Proc. AAAI Conf. Artif. Intell.* 37 (3) (2023) 3480–3488, <http://dx.doi.org/10.1609/AAAI.V37I3.25457>.
- [73] Q. Zhang, Y. Zhu, Aligning geometric spatial layout in cross-view geo-localization via feature recombination, *Proc. AAAI Conf. Artif. Intell.* 38 (7) (2024) 7251–7259, <http://dx.doi.org/10.1609/AAAI.V38I7.28554>.
- [74] F. Deuser, K. Habel, N. Oswald, Sample4Geo: Hard negative sampling for cross-view geo-localisation, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, 2023, pp. 16801–16810, <http://dx.doi.org/10.1109/ICCV51070.2023.01545>.
- [75] P. Wang, Z. Yang, X. Chen, H. Xu, A transformer-based method for UAV-view geo-localization, in: *Lecture Notes in Computer Science*, Springer Nature Switzerland, 2023, pp. 332–344, [http://dx.doi.org/10.1007/978-3-031-44223-0\\_27](http://dx.doi.org/10.1007/978-3-031-44223-0_27).
- [76] Z. Xia, O. Booi, M. Manfredi, J.F.P. Kooij, Visual cross-view metric localization with dense uncertainty estimates, in: *European Conference on Computer Vision*, Springer, 2022, pp. 90–106, <http://dx.doi.org/10.48550/ARXIV.2208.08519>.
- [77] D. Yuan, F. Maire, F. Dayoub, Cross-attention between satellite and ground views for enhanced fine-grained robot geo-localization, in: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, IEEE, 2024, pp. 1238–1245, <http://dx.doi.org/10.1109/WACV57701.2024.00128>.
- [78] R. Cao, G. Qiu, Urban land use classification based on aerial and ground images, in: 2018 International Conference on Content-Based Multimedia Indexing, CBMI, IEEE, 2018, pp. 1–6, <http://dx.doi.org/10.1109/CBMI.2018.8516552>.
- [79] F. Fang, Y. Yu, S. Li, Z. Zuo, Y. Liu, B. Wan, Z. Luo, Synthesizing location semantics from street view images to improve urban land-use classification, *Int. J. Geogr. Inf. Sci.* 35 (9) (2020) 1802–1825, <http://dx.doi.org/10.1080/13658816.2020.1831515>.
- [80] I. Goodfellow, A. Courville, Y. Bengio, *Deep learning*, in: *Adaptive Computation and Machine Learning*, The MIT Press, Cambridge, Massachusetts, 2016, Includes bibliographical references and index.

- [81] N. Gong, L. Li, J. Sha, X. Sun, Q. Huang, A satellite-drone image cross-view geolocalization method based on multi-scale information and dual-channel attention mechanism, *Remote Sens.* 16 (6) (2024) 941, <http://dx.doi.org/10.3390/RS16060941>.
- [82] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, <http://dx.doi.org/10.48550/ARXIV.2010.11929>, arXiv.
- [83] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, <http://dx.doi.org/10.48550/ARXIV.1810.04805>, arXiv.
- [84] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, W. Liu, CrossFormer++: A versatile vision transformer hinging on cross-scale attention, *IEEE Trans. Pattern Anal. Mach. Intell.* 46 (5) (2024) 3123–3136, <http://dx.doi.org/10.1109/TPAMI.2023.3341806>.
- [85] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, 2021, <http://dx.doi.org/10.48550/ARXIV.2105.15203>, arXiv.
- [86] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: Transformers make strong encoders for medical image segmentation, 2021, <http://dx.doi.org/10.48550/ARXIV.2102.04306>, arXiv.
- [87] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, NeRF: representing scenes as neural radiance fields for view synthesis, *Commun. ACM* 65 (1) (2021) 99–106, <http://dx.doi.org/10.1145/3503250>.
- [88] X. Li, J. Tupayachi, A. Sharmin, M. Martinez Ferguson, Drone-aided delivery methods, challenge, and the future: A methodological review, *Drones* 7 (3) (2023) 191, <http://dx.doi.org/10.3390/DRONES7030191>.
- [89] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, 2021, <http://dx.doi.org/10.1109/ICCV48922.2021.00717>.
- [90] T. Lentsch, Z. Xia, H. Caesar, J.F.P. Kooij, SliceMatch: Geometry-guided aggregation for cross-view pose estimation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2023, pp. 17225–17234, <http://dx.doi.org/10.1109/CVPR52729.2023.01652>.
- [91] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision, 3DV, IEEE, 2016, pp. 565–571, <http://dx.doi.org/10.1109/3DV.2016.79>.
- [92] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, 2021, <http://dx.doi.org/10.1109/ICCV48922.2021.00951>.
- [93] S. Wazir, M.M. Fraz, HistoSeg: Quick attention with multi-loss function for multi-structure segmentation in digital histology images, in: 2022 12th International Conference on Pattern Recognition Systems, ICPRS, IEEE, 2022, pp. 1–7, <http://dx.doi.org/10.1109/ICPRS54038.2022.9854067>.
- [94] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, 2019, <http://dx.doi.org/10.48550/ARXIV.1912.01703>, arXiv.
- [95] Y. Peng, X. Lin, N. Ma, J. Du, C. Liu, C. Liu, Q. Chen, SAM-LAD: Segment anything model meets zero-shot logic anomaly detection, 2024, <http://dx.doi.org/10.48550/ARXIV.2406.00625>, arXiv.
- [96] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, <http://dx.doi.org/10.48550/ARXIV.1412.6980>, arXiv.
- [97] K. Berntorp, T. Hoang, S. Di Cairano, Motion planning of autonomous road vehicles by particle filtering, *IEEE Trans. Intell. Veh.* 4 (2) (2019) 197–210, <http://dx.doi.org/10.1109/ITV.2019.2904394>.