CSC696H Project Proposal
Satellite View Generation from Ground View Image Sequence

Natnael T. Daba*
Department of Computer Science
University of Arizona
Tucson, AZ 85721
ndaba@arizona.edu

1 Project Summary

The goal of this project to explore generative modelling approach to generate satellite- or overhead-view image conditioned on a sequence of ground view images extracted from a video captured by camera mounted on a vehicle. Specifically, the cross-view video-based localization benchmark dataset, KITTI-CVL [5]. The task of estimating the overhead-view from a sequence of ground-view images is important because it can be used for downstream tasks such as cross-view video-based geo-localization where we want estimate the position (i.e. latitude and longitude) and orientation (i.e. heading angle) of vehicle. In this case, once we obtain a sequence of ground-view images, we then estimate the overhead-view image. Next, after some potential prepossessing step, we can compare the generated overhead-view image with a database of geo-tagged satellite images to estimate the position and orientation of where the ground-view image sequence was captured from.

2 Prior Work

Cross-view video geo-localization is a fairly new research area compared to related topics such as image-based geo-localization where the tasks is to estimate the location and orientation of a single ground view image as opposed to a sequence of images.

However, some closely related prior works include [1] where the problem tackled is the accurate pose estimation of ground-level images relative to a 3D model from satellite images, aiming to enhance 3D reconstructions. The method utilizes a cross-view SLAM solver that dynamically incorporates satellite model references during pose estimation, correcting non-rigid distortions and drifts common in monocular SLAM systems.

On the other hand, [4] tries to solve geo-localizing videos by learning geographical and temporal features to estimate GPS trajectories. It introduces a methodology based on a Geo-Temporal Feature Learning Network and employs GPS trajectory smoothing with a transformer encoder architecture to enhance localization accuracy. However, the approach doesn't estimate orientation.

Another work, published around the same time as [5] is [6] which introduces the GAMa dataset, a large-scale dataset with ground videos and corresponding aerial images. This work also introduces a baseline method that uses a hierarchical approach where a given video is localized first a clip-level (clip is a sequence of consecutive frames with a duration of 0.5 seconds) by extracting and matching features of the clip with the corresponding aerial image features that covers locations of those image sequences. Next, the results of the clip-level localization are then used to localize the entire video.

*

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3 Approach

The proposed method involves using a conditional denoising diffusion model [2, 3] to estimate the overhead-view image conditioned on a sequence of ground-view images. Figure 1 below shows a block diagram of the proposed method. A conditional denoising diffusion probabilistic model (DDPM) will be used to estimate overhead-view given a sequence of ground-view images and is trained using a simple reconstruction loss.

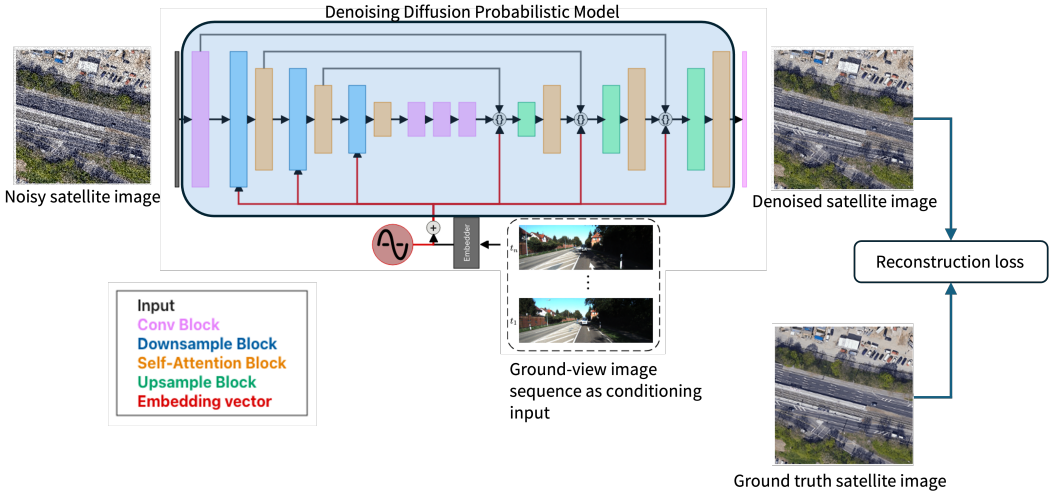

Figure 1: Proposed method

4 Evaluation Methodology

During training, the mean and standard deviation of the L2 reconstruction loss per minibatch will be used to evaluate the performance of the generative model both on the training set and a held-out validation set. Once training is done, The Inception Score (IS) and Fréchet Inception Distance (FID) will be used to evaluate how well the generative captured the data distribution of satellite images.

References

[1] Mostafa Elhashash and Rongjun 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 Journal of Photogrammetry and Remote Sensing*, 188:62–74, June 2022.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[3] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[4] Krishna Regmi and Mubarak Shah. Video geo-localization employing geo-temporal feature learning and gps trajectory smoothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12126–12135, 2021.

[5] Yujiao Shi, Xin Yu, Shan Wang, and Hongdong Li. Cvlnet: Cross-view semantic correspondence learning for video-based camera localization. In *Asian Conference on Computer Vision*, pages 123–141. Springer, 2022.

[6] Shruti Vyas, Chen Chen, and Mubarak Shah. Gama: Cross-view video geo-localization. In *European Conference on Computer Vision*, pages 440–456. Springer, 2022.