Video Geo-Localization Employing Geo-Temporal Feature Learning and GPS Trajectory Smoothing - Supplementary Materials

Krishna Regmi and Mubarak Shah
Center for Research in Computer Vision, University of Central Florida
krishna.regmi7@gmail.com, shah@crcv.ucf.edu

The supplementary material provides additional details on GPS loss, additional qualitative results for all four regions, ablations and feature visualizations. We also provide evaluations on the test dataset by Vaca-Castano *et al*. [2].

1. GPS Loss

In this work, we propose a novel GPS loss to train our Geo-Temporal Feature Learning (GTFL) network. The GPS loss is used together with Frame Triplet loss and Clip Triplet loss, as explained in the main paper. We believe that geographical regions have landmarks, landscapes and vegetation unique to the regions and holds true over a small geographical region. So, we utilized the GPS loss as an additional loss function to enforce geographical consistency on the learnt features. The appearance features between the query (Berkeley Driving Dataset [3]) videos and gallery (Google StreetView) images are vital during the retrieval and thus the GPS loss provides additional weak supervision while learning the features.

In Figure 1, we illustrate the relationship between the feature distances among the images and physical distances between their GPS locations for a subset of the training dataset. Each point in the scatter-plot represents the feature distance between a pair of images along x-axis and their physical distances along y-axis. We observe that a linear relationship can be established between the feature distances and the geographical distances and thus, the linearity can be imposed during the training of our network.

We report the impact of the GPS loss in Table 3 in the main paper. We observed that, the inclusion of the GPS loss during the training helps learn discriminative features and contributes to minimizing the localization error.

2. Additional Qualitative Results

In this section, we provide additional qualitative results for all four regions of evaluations, San Francisco, Berkeley, Bay Area and New York. Figure 2 visualizes three sample images from Bay Area (first row) and San Francisco (second row). Similarly, Figure 3 shows sample images from New York (first row) and Berkeley (second row). The green curves represent the ground truth trajectories and the red curves show the corresponding predicted trajectories in each image. We can observe that the predicted trajectories have a very high overlap with the ground truth trajectories, justifying that the network is able to localize the trajectories successfully.

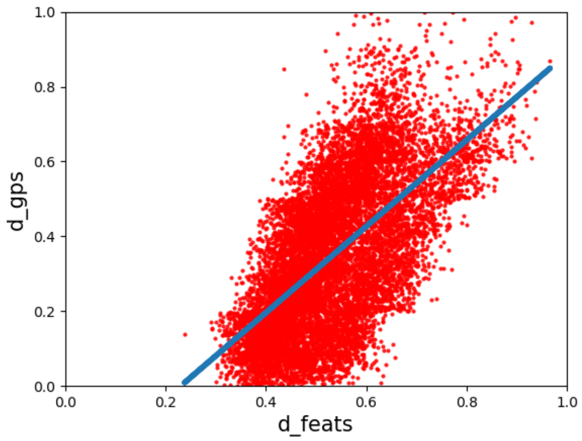
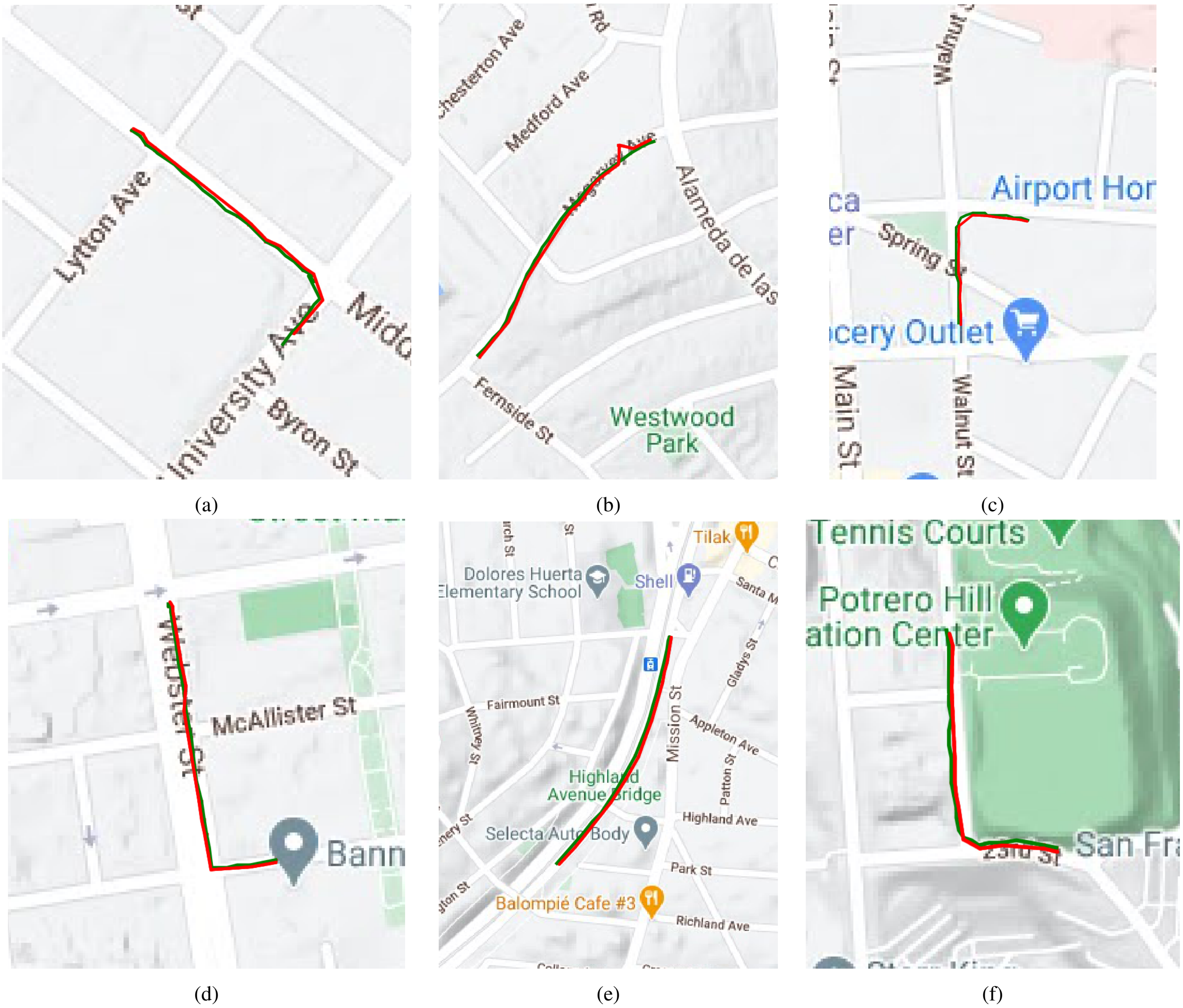


Figure 1: Scatterplot showing the relationship between feature distances (d_{feats}) and geographical distances (d_{gps}) for a subset of the dataset. Each point in the plot represents the feature distance between two images along x-axis and their geographical distance along y-axis. The blue line is the line of best fit through the scatterplot and shows a linear relationship can be established between the points; with a slope of 1.077 and intercept of -0.2313. We model the GPS loss to preserve the linear relationship between the feature distances and gps distances.

3. Additional Ablation Study

Here, we provide additional ablations conducted for our experiments.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2: Qualitative Results. Example Images showing the ground truth (green curves) and predicted (red curves) trajectories for Bay Area (first row) and San Francisco (second row).

Table 1: Ablation study on varying feature dimension size. We report the evaluations in terms of localization error in meters.

Feature dim.	**SF**	**Bay Area**	**Berkeley**	**NY**
d=1024	472.46	642.17	507.03	603.72
d=512	**300.47**	**524.28**	**424.79**	**493.43**
d=256	363.24	888.99	579.49	518.47

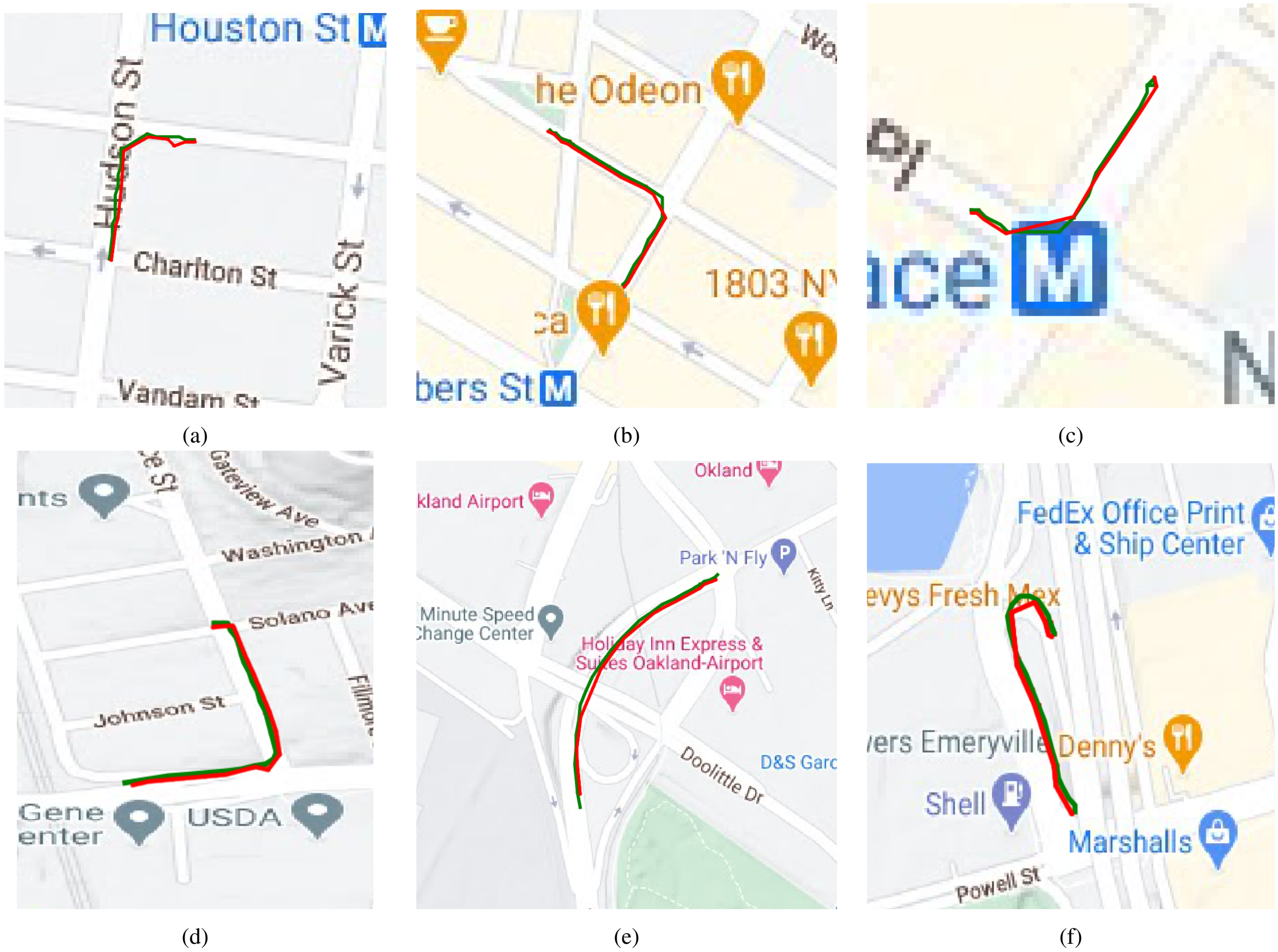
3.1. Ablation on Feature Dimensions

Feature dimensions are critical components in deep learning networks. The feature dimension represents the size of a feature vector representing each input image. Larger feature dimension means larger memory requirements to store them as well as more computations. Smaller dimension provides more compact representations but they may be insufficient for mapping the images to feature space.

We conducted experiments by varying the feature dimensions as 256, 512 and 1024 and report the results in Table 1. As observed, 512 dimensional feature representation works the best for our experiments.

3.2. Ablation with and without NetVLAD layer

NetVLAD [1] is a popular trainable pooling layer used to capture the information about the statistics of local de-



(a) (b) (c)

(d) (e) (f)

Figure 3: Qualitative Results. Example Images showing the ground truth (green curves) and predicted (red curves) trajectories for New York (first row) and Berkeley (second row).

scriptors aggregated over the image. NetVLAD learns the cluster centers and residuals. NetVLAD has be widely used in image retrieval problems and thus we use it in our framework as well. Here, for this ablation, we conduct experiments with and without NetVLAD layer and present the results in Table 2 . We can observe that the network with NetVLAD layer performs slightly better than the network without NetVLAD layer. This affirms that NetVLAD helps in retrieval problems; but the large improvement in results for our proposed method over 2D CNN (as reported in Table 2 in the main paper) is contributed by temporally learning of features and not necessarily due to the use of NetVLAD in our network.

Table 2: Ablation study on experiments with and without NetVLAD layer in the proposed network. We report the evaluations in terms of geo-localization error in meters.

Methods	SF	Bay Area	Berkeley	NY
without NetVLAD	531.53	655.49	736.41	715.98
with NetVLAD	300.47	524.28	424.79	493.43

4. Evaluation on Additional Dataset

We have conducted the evaluation on test sequences from the prior work by Vaca-Castano *et al*. [2]. Due to mismatch in the number of frames and the GPS annotations for the clips, we interpolated the sparse GPS annotations to obtain one-to-one correspondences with the test frames.

The average localization error for our proposed approach is 131.54 meters. After smoothing, the localization error reduces to 54.14 meters. The average localization error reported in the original paper for frame by frame evaluation is 268.6 meters. This justifies the contribution of our proposed method in learning coherent features for the video frames compared to frame based evaluation method. Their error after trajectory reconstruction is 9.94 meters. The reason for such low error is that they discard the outliers and only retain the correct predictions and compute the errors on the retained GPS locations. If we also discard the outliers as predicted by our prediction head and only keep the inliers and evaluate on them, the error is 11.79 meters. We believe these numbers are comparable.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 2

[2] Gonzalo Vaca-Castano, Amir Roshan Zamir, and Mubarak Shah. City scale geo-spatial trajectory estimation of a moving camera. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1186–1193. IEEE, 2012. 1, 3

[3] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1