GAMa: Cross-view Video Geo-localization *(Supplementary Material)*

Shruti Vyas, Chen Chen, and Mubarak Shah

Center for Research in Computer Vision, University of Central Florida, USA
{shruti,chen.chen,shah}@crcv.ucf.edu

1 Overview

In this supplementary, we have included additional qualitative results. In Section 2, we show and discuss some qualitative results for video-level geo-localization. In Section 3, we have included additional qualitative results for clip-level geo-localization and ablations.

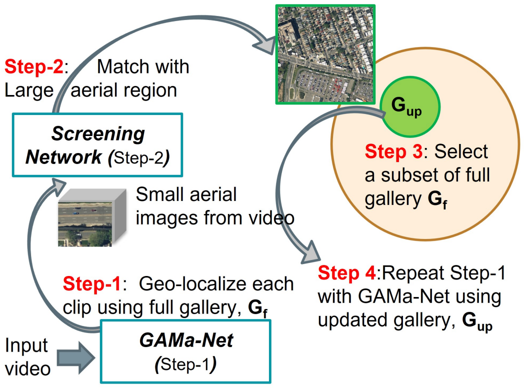
Step-2: Match with Large aerial region
Screening Network (Step-2)
Small aerial images from video
Step-1: Geo-localize each clip using full gallery, G_f
Input video
GAMa-Net (Step-1)
G_{up}
Step 3: Select a subset of full gallery G_f
Step 4:Repeat Step-1 with GAMa-Net using updated gallery, G_{up}

Fig. 1. An outline of the proposed approach. Given a ground video, in Step-1 0.5 sec clips from the video are input to GAMa-Net. It takes one clip at a time and matches it with an aerial image. In Step-2, the sequence of aerial images obtained from GAMa-Net is input to the Screening network to retrieve the corresponding larger aerial region. This is our video-level geo-localization. In Step-3, top predictions of these larger aerial regions provides the updated gallery for a video. Step-4, is prediction by GAMa-Net while using the updated gallery

An **outline** of the proposed approach in Figure 1 shows how video level geo-localization is used to improve the clip-level results. The GAMa-Net outputs aerial image predictions at clip-level using the full gallery i.e. G_f. A video comprises of a number of clips (upto 40 clips per video), thus a sequence of aerial images is obtained from each query video, this sequence is then input to the screening network. The screening network uses this sequence of small aerial images to predict a large aerial region corresponding to the query video. Then we select Top-1% large aerial images to update the gallery (the updated gallery G_{up}) for GAMa-Net and reduce the search space. *We are hopeful that with further research this approach has the potential to be generalized to even larger scale.*

Our current evaluation/dataset gallery spans over multiple cities since we selected the videos from BDD100k[1]. In Figure 2, we show the trajectory as predicted by GAMa-Net following an hierarchical approach where the search space was reduced using the predictions from screening network. It is not usual to show trajectory in cross-view geo-localization as it is coarse geo-localization. In Figure 2, top-1 predictions are marked as boxes with numbers representing the time of the respective clip. The numbers show an oscillating trajectory however we are on the correct path which is the same track as the ground truth. There are some outliers (e.g. 5, 12 and 19) which were geolocalized in the nearby region and are not shown in the figure.

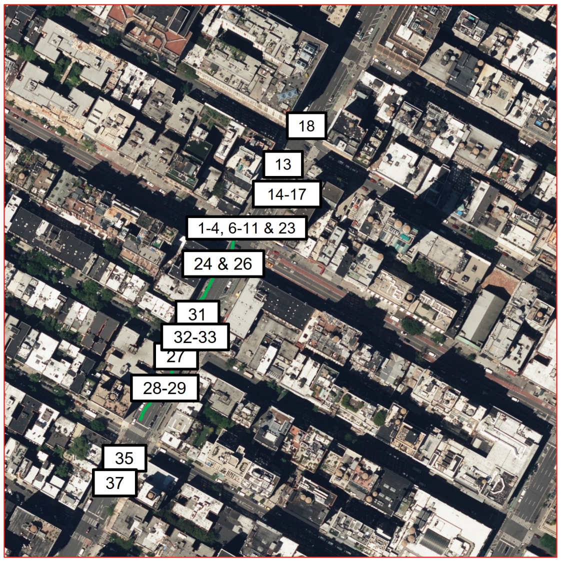

Fig. 2. Ground truth trajectory of a video sample is marked as a green line and trajectory as predicted by clip-level geo-localization (top-1) is marked by time(second). Please note that in our approach a clip is matched with corresponding small aerial image.

As shown in main paper we observe an improvement in Top-1 recall using this approach. The screening of the locations at large region level removed the confusing samples.

2 Qualitative Results: *Video-level Geo-localization*

In Figure 3, we show some of the predictions of larger aerial regions as ranked by the screening network. For each sample, the first row shows frames of the video spanning over 35 sec. of time duration. In the subsequent row, we have the ground truth larger aerial region which is followed by Top-5 predictions by the screening network. The examples shown here are mostly correct predictions where Top-1 is same as the ground truth.

In the first sample (Figure 3), we see that the video frames are **partially outdoors and partially indoors**. Indoor setting appears to be a parking lot or an underpass. The screening network is able to correctly localize the larger aerial region while using the information available from all the clips. The prediction when used to reduce the gallery of GAMa-Net is likely to enable a better prediction by screening-out the far away regions. Frame-by-frame it would have been difficult to localize the indoor frames. However, using the hierarchical approach the network is able to use the context from outdoor frames. We also observed this from the predictions, with GAMa-Net only 12 clips out of the 38 clips had ground truth prediction in top-10 and after reducing the gallery using the prediction from screening network this number increased to 28 clips. In the second sample (Figure 3), we can see that all Top-5 predictions are visually similar. These predictions appears to be from the same region and most are from around a mile radius of the ground truth larger aerial region.

In the third sample, because of a car, there is **occlusion** in part of the video. GAMa-Net correctly localizes the initial clips however fails in the clips with occlusion. After screening the gallery using the correct larger aerial prediction, most of these clips were correctly geolocalized by GAMa-Net in Top-1, Top-5 or Top-10. Similarly, in fourth and fifth sample the occlusion is in all the frames either because of the car hood or rear view mirror. In the fourth sample, similar improvement in clip-level geo-localization was observed with GAMa-Net because of correct screening of the larger aerial region at the video-level. We see correct video-level Top-1 with fifth sample, however the improvement in the clip-level predictions by GAMa-Net was less and only 3-4 more clips had predictions added to top-10 as compared to retrieval from the full gallery. The second top prediction of this sample is visually similar to the ground truth however a closer look shows that it is a different image. In the last or sixth sample we can see that multiple correct predictions because of the visual similarity are in Top-5.

With all these video samples, because of some of the correct clip-level predictions by GAMa-Net, the screening network is able to localize at video-level and identify the correct larger aerial region. However, the last sample had correct clip-level prediction for a single clip out of 31. It is likely that the **visual similarity** of the incorrect aerial image predictions helped with correct video-level geo-localization at Top-1 using screening network.

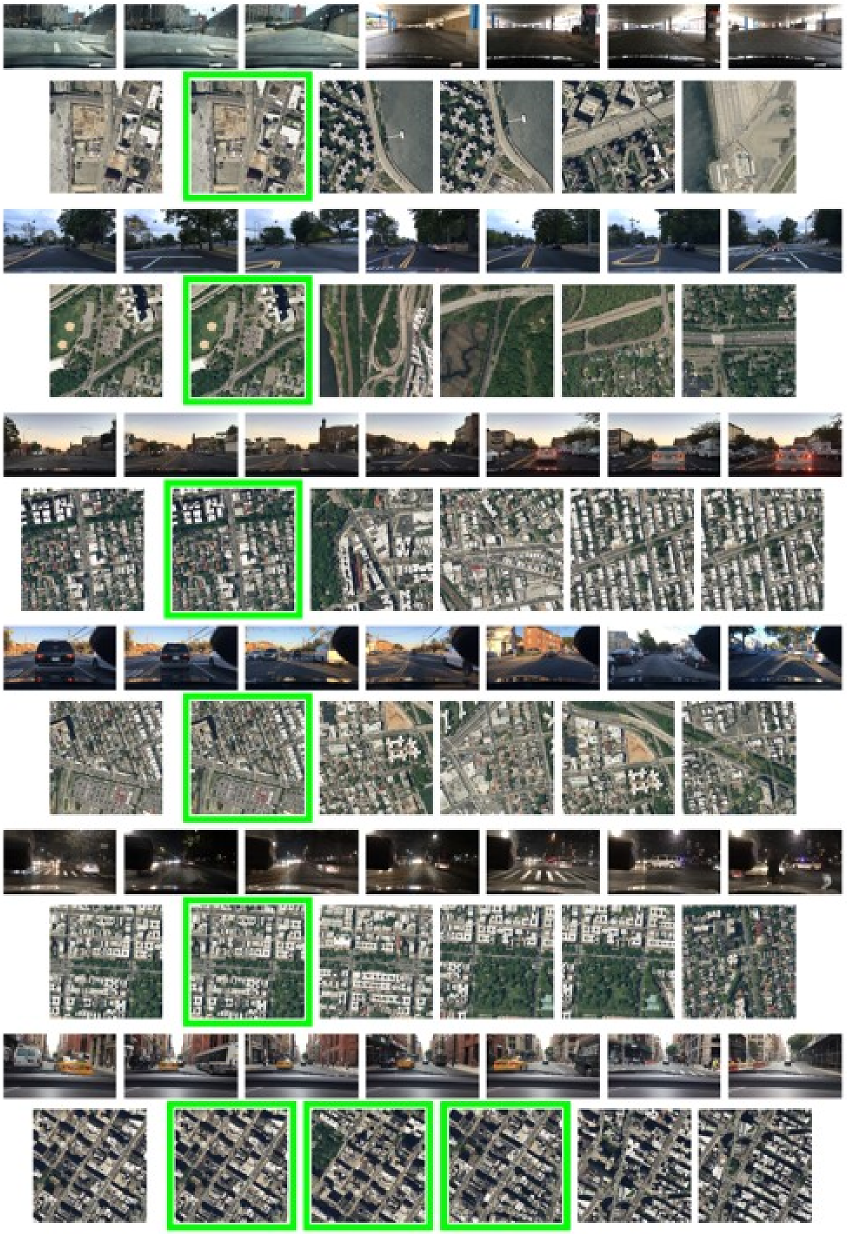

Fig. 3. Here, we show results for video-level geo-localization using the Screening network. In each sample, we show seven frames of the query video, followed by the ground truth aerial image and top-5 predictions of larger aerial regions.

3 Additional Qualitative Results: *Clip-level Geo-localization*

In Figure 4 and Figure 5, we show additional results using the proposed GAMa-Net with Hierarchical approach, where we use the video-level predictions to improve clip-level geo-localization. Figure 4 shows examples of correct Top-1 predictions. We can see that most of the Top-5 predictions are nearby the ground truth. Because of the nearby GPS labels in video clips we have overlapping in aerial images and all these images appear in the Top-5 predictions along with the ground truth. Figure 5 shows examples of fail cases, we can see that most of the fails are due to shadows or occlusion or poor quality of aerial images. The model makes meaningful predictions however fails in difficult or confusing samples e.g. in the last sample traffic lights are visible in the video and predictions are with zebra crossings however does not retrieve the correct aerial image in Top-5.

In Figure 8 and Figure 9, we show some additional results with the combined model which does clip-level geo-localization by retrieving a matching aerial image. The network however is an ablation of the proposed GAMa-Net and does not have a transformer encoder.

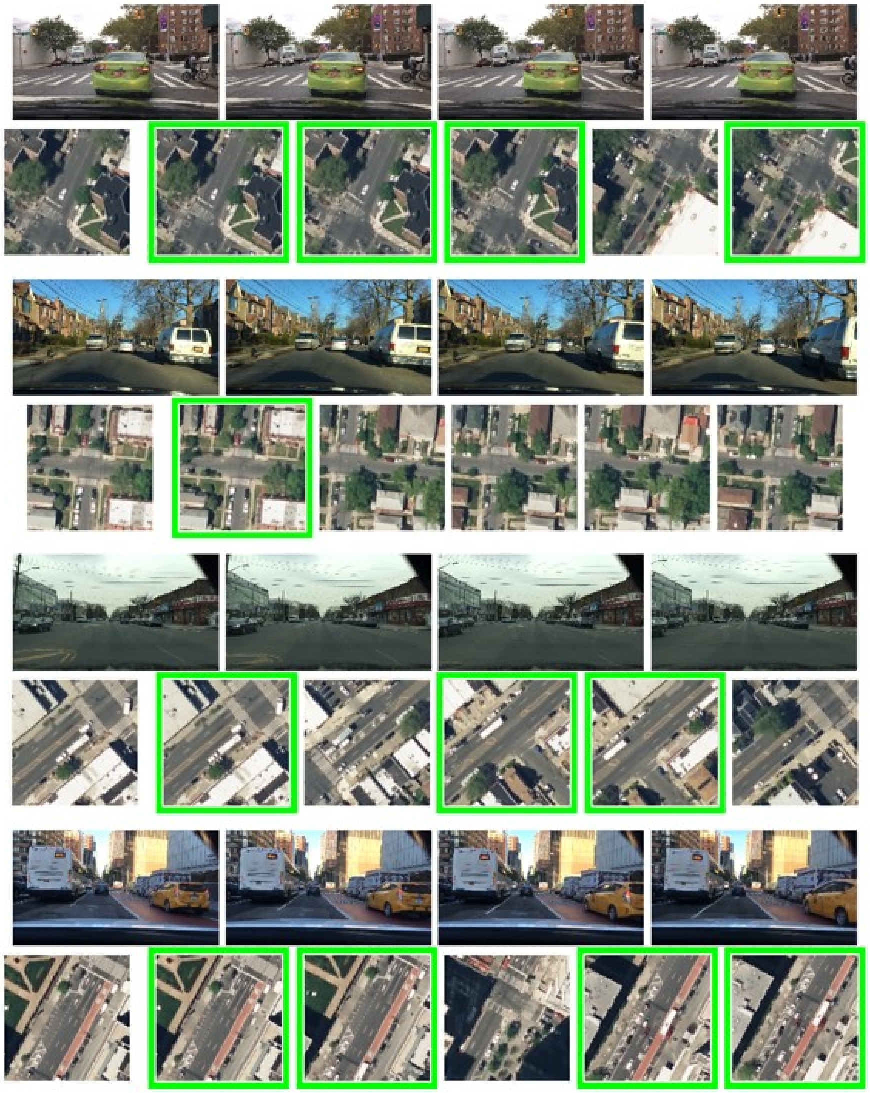

Fig. 4. Here we show sample *success cases* with GAMa-Net using Hierarchical approach where Top-1 prediction is correct. In each sample, four frames of the query clips are followed by ground truth aerial image and Top-5 predictions

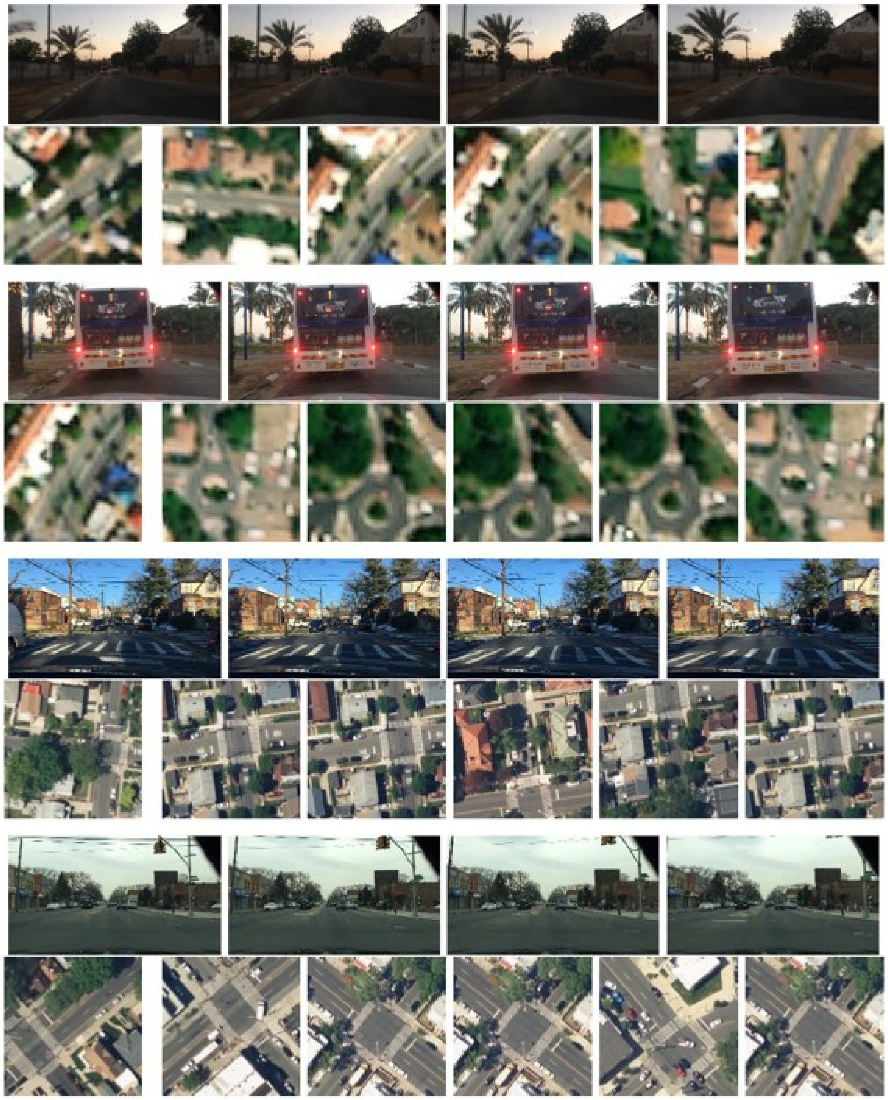

Fig. 5. Here we show sample *fail cases* with GAMa-Net using Hierarchical approach where Top-1 prediction is correct. In each sample, four frames of the query clips are followed by ground truth aerial image and Top-5 predictions

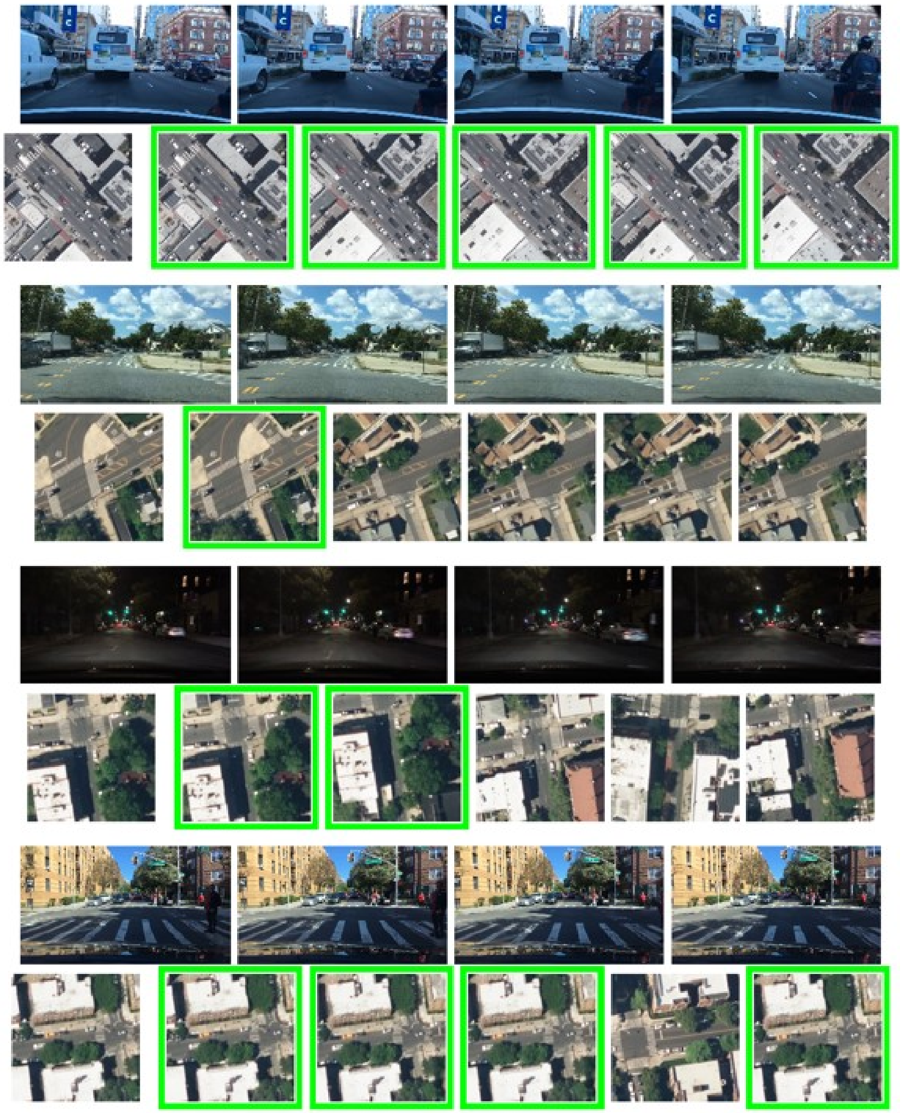

Fig. 6. Here we show *success cases* using GAMa-Net without Hierarchical approach where top-1 prediction is correct. In each sample, four frames of the query clips are followed by the ground truth aerial image and Top-5 predictions

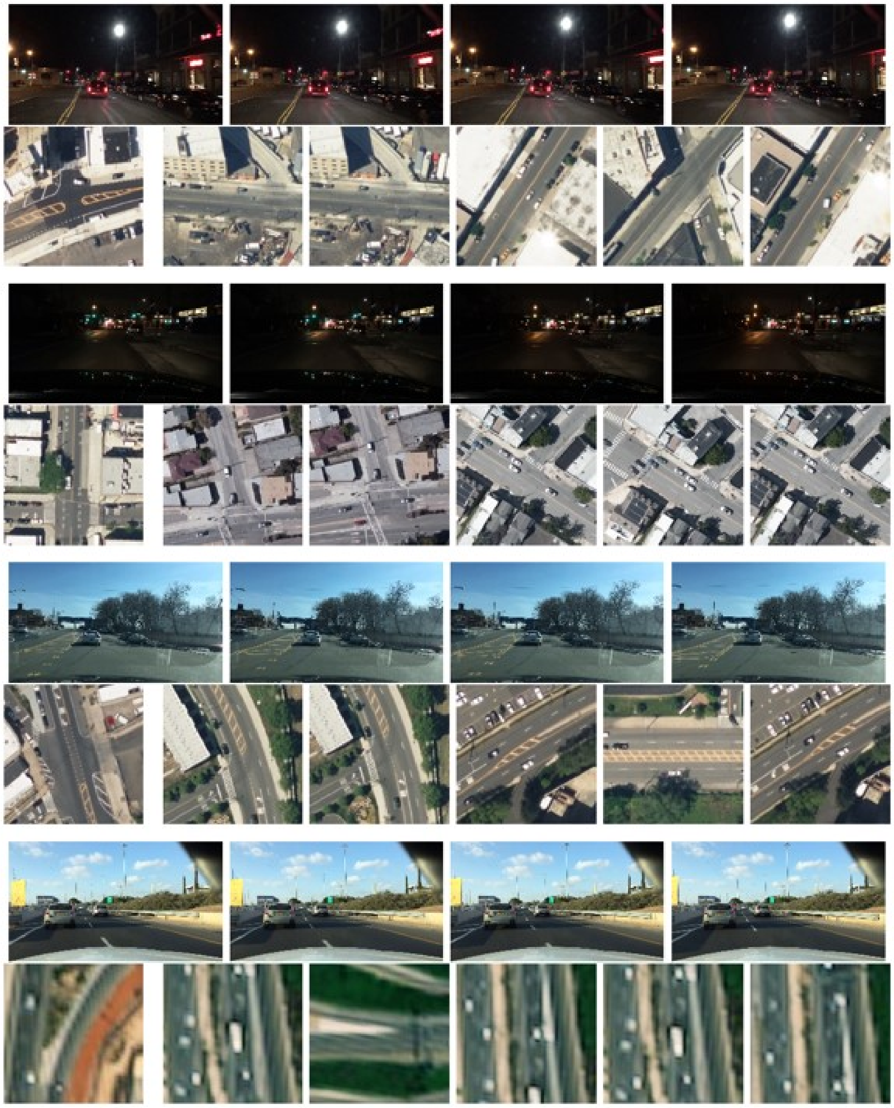

Fig. 7. Here we show *fail cases* using GAMa-Net without Hierarchical approach. In each sample, four frames of the query clips are followed by the ground truth aerial image and top-5 predictions

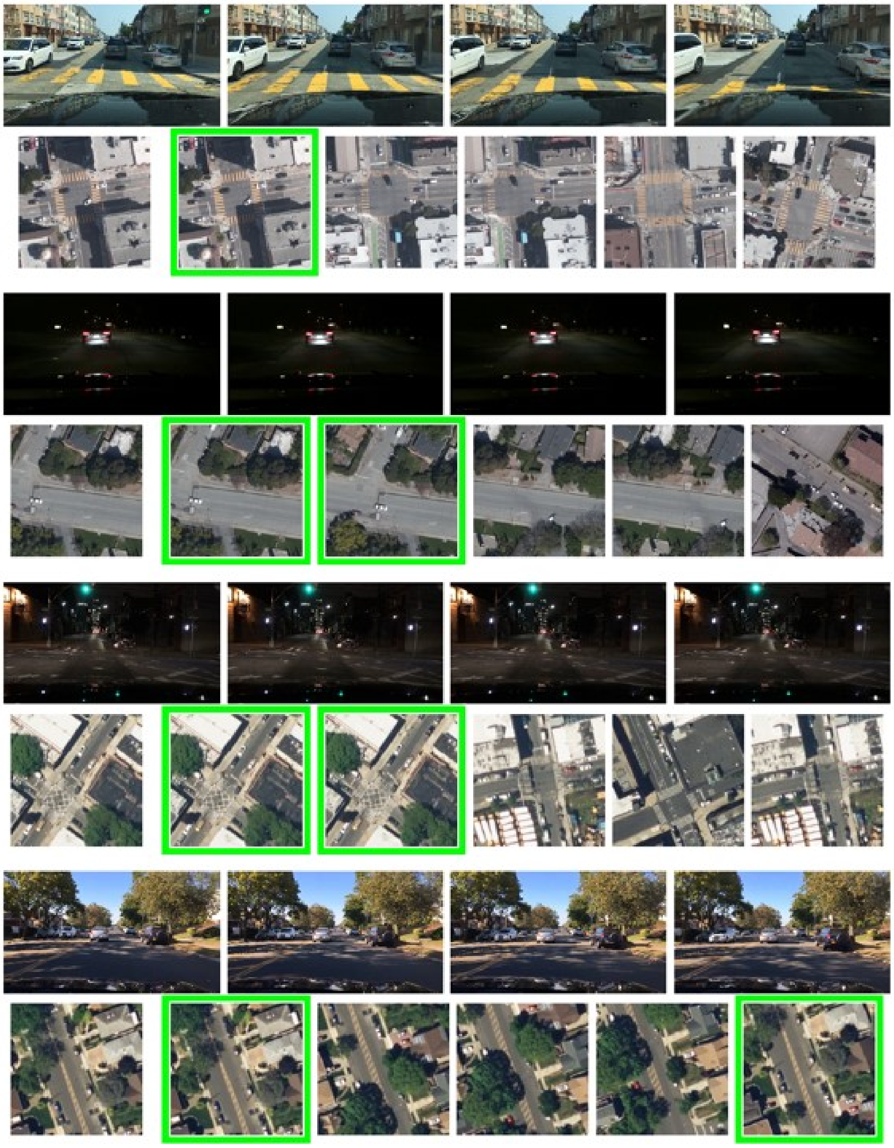

Fig. 8. Here we show *success cases* for combined model where Top-1 prediction is correct. In each sample, four frames of the query clips are followed by ground truth aerial image and top-5 predictions

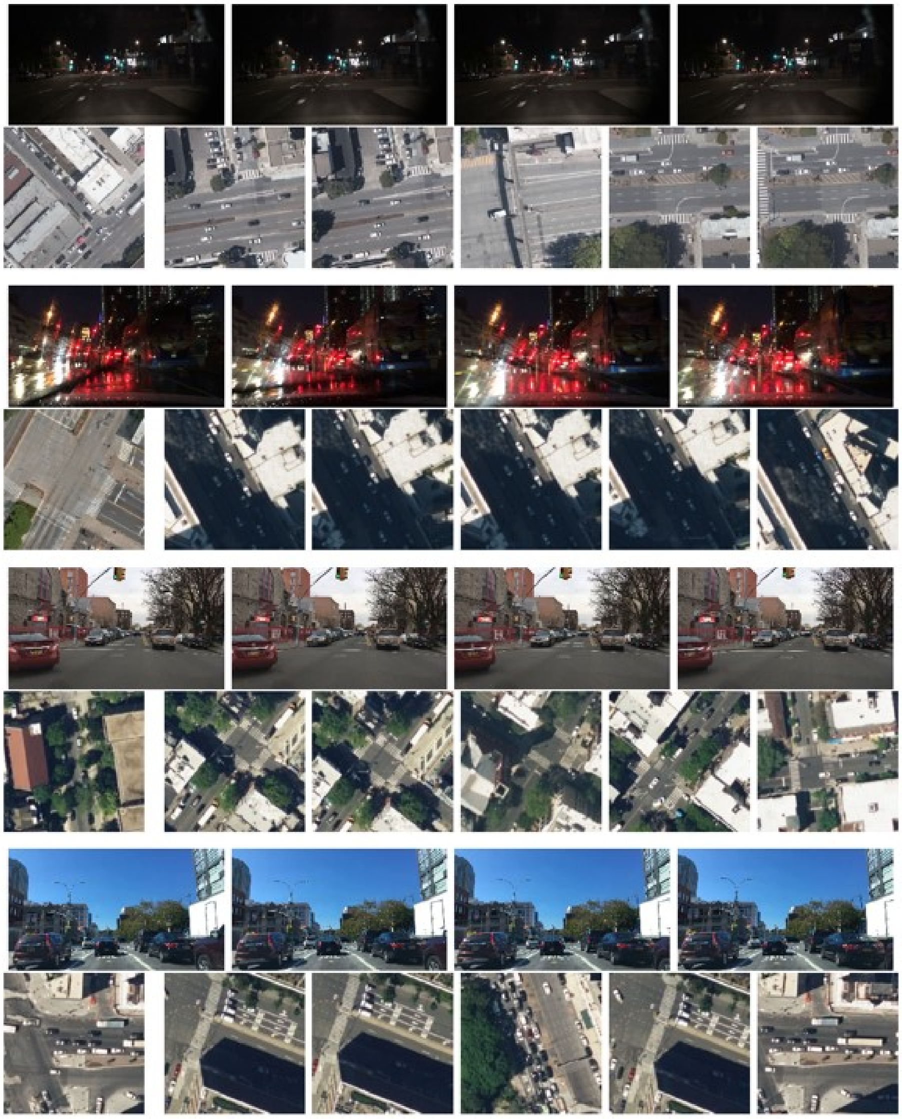

Fig. 9. Here we show *fail cases* for combined model. In each sample, four frames of the query clips are followed by ground truth aerial image and Top-5 predictions

References

1. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)