GAMa: Cross-view Video Geo-localization

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

Abstract. The existing work in cross-view geo-localization is based on *images* where a ground panorama is matched to an aerial image. In this work, we focus on **ground videos** instead of images which provides additional contextual cues which are important for this task. There are no existing datasets for this problem, therefore we propose **GAMa dataset**, a large-scale dataset with ground videos and corresponding aerial images. We also propose a novel approach to solve this problem. At **clip-level**, a short video clip is matched with corresponding aerial image and is later used to get **video-level** geo-localization of a long video. Moreover, we propose a hierarchical approach to further improve the clip-level geo-localization. On this challenging dataset, with unaligned images and limited field of view, our proposed method achieves a Top-1 recall rate of 19.4% and 45.1% @1.0mile. Code & dataset are available at this link.

1 Introduction

Video geo-localization is a challenging problem with many applications such as navigation, autonomous driving, and robotics [11,25,5]. The problem to estimate geolocation of the source of a ground video is also faced by first respondents now and then. There are two main formulations to address this problem; same-view geo-localization [20,18,3] and cross-view geo-localization [13,36,17]. In same-view geo-localization, the query ground image is matched with a street view image from the reference set or gallery. Research in **videos is limited to same-view geo-localization** where a frame-by-frame approach is followed. This approach relies on availability of ground view images for all the locations which may not be possible always.

In such scenarios, cross-view geo-localization is more useful where the query image is matched with the corresponding aerial or satellite image. However, cross-view geo-localization is a more difficult problem since there is a large domain shift between ground and aerial views. The limited field-of-view in the ground view makes the problem even harder and it is sometimes difficult even for humans to identify the correct location of an image.

The existing works in cross-view geo-localization follow an image based approach where a ground image is matched with an aerial image [17,30,27]. In such an approach, the contextual information available with the video is lost. We propose to focus on the geo-localization of ground videos instead of images to take advantage of the context, i.e. how the view in one frame is related/located w.r.t.

Fig. 1. A) Sample ground video frames and corresponding aerial image from the proposed GAMa dataset. The stars represent locations of video frames every second and mark the trajectory of the video; B) An example of larger aerial region for a video where stars represent the GPS points labeled every second and marked as X_1, .. X_n. The range μ is derived from the GPS points and corresponds to side of the bounding box/square. In histogram, we show frequency of number of sample videos at a given value of μ, in training set. The red dotted lines in histogram show the lower and and upper threshold (i.e. 0.001 and 0.004) for selecting the videos

another frame. To the best of our knowledge there are no existing datasets which are publicly available and can be used for this problem. Thus, we propose a new dataset, named as *GAMa (**G**round-video to **A**erial-image **Ma**tching)*, which contains ground videos with GPS labels and corresponding aerial images. It consists of $\sim$ 1.9M aerial images and 51K ground videos where each video is around 40 seconds long. An example video with representative frames and corresponding aerial image is shown in Figure 1.

We propose GAMa-Net as a benchmark method to solve this problem at *clip-level* where we match every 0.5 second (short clip), from a long video, with the corresponding aerial view. A frame-by-frame approach can be a straightforward method for video geo-localization, where a 2D convolutional network is used to get the spatial features frame-by-frame. However, as mentioned before a frame-by-frame approach ignores the rich contextual information available in a video. Thus, we propose a 3D-convolution based approach to learn the spatial-temporal features from a ground video. The proposed approach is trained using an *image-video contrastive loss* which aims to match the ground view with the aerial view. This provides a *clip-level* geo-localization which is unexplored.

Next, we propose a *hierarchical approach* which helps in improving the clip-level geo-localization performance while providing a *video-level* geo-localization with the help of clip-level predictions. Corresponding to each video, we get a set of small aerial images as matched with 0.5 second clips. We match this set of aerial images and against a larger geographical area. Therefore, we make use of the contextual information, i.e. location of smaller aerial images w.r.t. each other in a larger aerial region, to improve the geo-localization.

We evaluate the proposed approach on GAMa dataset and demonstrate its effectiveness for clip/video geo-localization. We provide an analysis and propose a set of baselines to benchmark the dataset. We make the following contributions:

- A novel problem formulation *i.e.* cross-view *video* geo-localization and a large-scale dataset, *GAMa*, with ground videos and corresponding aerial images. This is the *first* video dataset for this problem to the best of our knowledge.
- We propose *GAMa-Net*, which performs *cross-view video geo-localization* at *clip-level* by matching a ground video with aerial images using an *image-video contrastive* loss.
- We also propose a novel *hierarchical approach* which provides *video-level* geo-localization and utilizes aerial images at different *scales* to improve the clip-level geo-localization.

2 Related works

Traditional classical features like SIFT, were earlier used for image matching in geo-localization [21,34]. As deep learning has proven successful in feature learning most of the recent studies have adopted CNN based approach for learning discriminative features for image matching [26]. The problem of image or video geo-localization is solved using either the same view, which is mostly the ground view, or cross-view. **Same-view geo-localization** makes use of the large collections of geo-tagged images available online [16,35,20,18,3,2]. The problem is approached with the assumption that there is a reference dataset consisting of geo-tagged ground images and there is an image corresponding to each query image. There is some research in video geo-localization as well which is solved at frame level and is followed by trajectory smoothing [6,18,3]. However, a more complete coverage by overhead reference data such as satellite/aerial imagery has spurred a growing interest in cross-view geo-localization.

Cross-view Geo-localization. Most of the recent work adopt CNN based approaches. Several studies have explored CNN architectures for matching ground-level query images to overhead satellite images [29,27,37,12,8,30]. Triplet loss is mostly used optimization function [8,17]; certain studies however report better results with contrastive loss[16]. Field-of-view (FOV) also plays an important role in deciding the recall rate and ground panorama is highly accurate as compared to limited FOV [8,36]. Similarly, videos also contain more visual information as seen through the trajectory of the camera and can be expected to provide more accurate geo-localization as compared to images or frames with similar FOV.

Current works on cross-view geo-localization follow image based approach since the existing datasets only contain image pairs for ground and aerial view [26,17,9,32,19]. However, some papers do report testing their model on videos using a frame-by-frame approach [9]. Most popular datasets, Cross-View USA (CVUSA) dataset [31], CVACT [13], and Vo et al. [29] contain ground panorama aligned with corresponding aerial image. Recent publications have shown very high recall rate on these datasets while using panoramas however these values

Table 1. Statistics of the proposed GAMa dataset. Note: Large aerial images are at 1792×1792 resolution whereas small aerial images or tiles are at 256×256 resolution

Parameter	**Train**	**Test**	**Train(day)**	**Test(day)**
Videos	45029	6506	21144	3103
Large aerial images	45029	6506	21144	3103
Clips	1.68M	243k	790k	116k
Centered (CN) small aerial images	1.68M	243k	790k	116k
Uncentered (UCN) small aerial images	2.21M	319k	1.04M	152k

are quite low when using limited FOV and unaligned images, i.e. top@1 recall of upto 14% [23,32]. VIGOR [36] dataset also contains panorama however being unaligned it is more realistic. All these datasets use ground panorama which is not realistic from video geo-localization, as videos have limited FOV, neither do they have time series data required for such training. It is possible to get unaligned images and limited FOV from these datasets however, there is no existing dataset with ground videos and aerial image pairs to solve **cross-view** geo-localization in videos and the proposed dataset addresses this gap.

Cross-view is also used for fine geo-localization of UAVs or robots. Camera feed (also frame-by-frame) and a known small region/map, of about a mile is used to find a more exact location in the given map [7]. Sometimes a prior is given to estimate the vehicle pose [10] or multiple modalities are used [14]. However, our focus is on coarse geo-localization where the gallery spans over multiple cities.

A frame-by-frame geo-localization of videos is also possible with the proposed GAMa dataset where 2D convolutional networks are used to extract the spatial features from each frame. However, the contextual information as available from a video is ignored with frame-by-frame processing. Fusion of features obtained from 2D-CNN is also possible however it is more memory intensive as compared to a 3D-CNN network. To address these limitations and challenges, we propose a videos based cross-view geo-localization.

3 GAMa dataset

The proposed GAMa (**G**round-video to **A**erial-image **Ma**tching) dataset comprises of select videos from BDD100k [33] and aerial images from apple maps.

Ground video selection: Most of the videos in BDD100k dataset are 40 sec long and usually have GPS label every second. The selection of videos from dataset was based on the range of latitude and longitude for a given video where we use a range parameter μ. We label the GPS points at n^{th} second as X_n (lat_n, $long_n$), where the corresponding latitude is lat_n and longitude is $long_n$ (Figure 1B, Aerial image). Thus, for the whole video we have GPS points as $X_1(lat_1, long_1)$, $X_2(lat_2, long_2)$,.., $X_n(lat_n, long_n)$. If max latitude = lat_k and min latitude = lat_l, then Latitude range = $lat_k - lat_l$. Also, if max longitude =

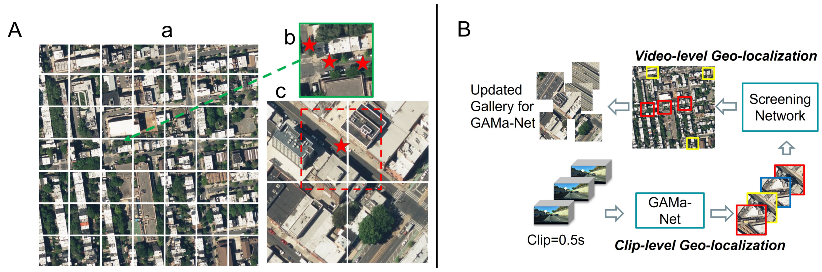

Fig. 2. **A**: Sample aerial images from GAMa dataset. a) A large aerial image corresponding to a video, b) Small UCN aerial image (Zoomed-in) and possible locations (stars) for a matching clip, and c) small centered (CN) aerial image as centered around the GPS label; **B**: An outline of the proposed approach. From a given ground video, clips of 0.5 sec are input to GAMa-Net, one clip at a time is matched to an aerial image. The sequence of aerial images thus obtained for a video is input to the Screening network, to retrieve the large aerial region for video-level geo-localization. Top predictions of larger aerial regions provide the updated gallery for GAMa-Net

$long_p$ and min longitude $= long_q$, then Longitude range $= long_p - long_q$. Range, $\mu = \max$(Latitude range, Longitude range) In order to eliminate stationary and very fast videos, we select videos with μ from 0.001 to 0.004. Figure 1B. shows the distribution of training videos with range, μ. The distribution was similar for training and test sets. This selection left us with 46596 training and 6728 testing videos which were further screened based on the availability of GPS labels.

Aerial images: For the selected videos, aerial images are downloaded as tiles from Apple maps at 19× zoom [1]. The dataset comprises of one large aerial image (1792×1792) corresponding to each video of around 40 sec. and 49 uncentered small aerial images (256×256) for these large aerial regions. Table 1 summarizes the dataset statistics. Since, most of the videos have a GPS label every second we divide the videos into smaller clips of 1 sec. each and for each clip we have an aerial tile. In Figure 1A, we see an example of a large aerial region, along with the video frames at each second.

Aerial image centering: We have a centered (CN) and an uncentered (UCN) set of small aerial images, as reported in Table 1. As shown in Figure 2Aa, UCN aerial images are obtained by dividing the large aerial image into smaller tiles. The GPS label in these UCN smaller aerial images can be anywhere besides the center, Figure 2Ab. There are three labels in the figure and for each of these GPS points the same tile will be considered as the ground truth. For making a CN set, as shown in Figure 2Ac, we take a crop centered around the corresponding GPS point however it leads to overlap among the aerial images since in some videos we have a distribution where GPS points are nearby. In the dataset, we still have a one-to-one correspondence among aerial images and short ground clips. The overlap among the aerial images increases the difficulty level for top-1 retrieval, thus we evaluate with distance thresholds.

Table 2. Comparison of GAMa dataset with existing cross-view geo-localization datasets. Note that the previous datasets do not contain ground videos

	Vo[29]	CVACT[13]	CVUSA[31]	GAMa (proposed)
Ground videos	no	no	no	51535
Panorama	450,000	128,334	44,416	no
Aerial images	450,000	128,334	44,416	1.92M
Aerial img resolution	-	1200×1200	750×750	256×256 & 1792×1792
Multiple cities	yes	no	yes	yes

The ground videos in the proposed GAMa dataset are selected from BDD100k dataset[33]. They are distributed all over the US and also from middle eastern region. However, most of the videos are from four US cities; New York, Berkeley, San Francsico, and Bay area. They show different weather conditions, as well as different times of the day including day and night.

There is occlusion in videos and shadows of the skyscrapers in aerial images. Limited FOV in videos and all stated characteristics bring it closer to a realistic scenario, making it a difficult dataset for geo-localization. In Table 2, we show a comparison with existing datasets for cross-view geo-localization.

4 Method

An overview of the proposed approach which works on clip and video level is show in Figure 2B. Ground-video to Aerial-image Matching Network (GAMa-Net) learns features for ground view clips and aerial images; and bring the matching pair closer in the feature space by applying a contrastive loss. This provides a clip-level geo-localization for a long video. In addition, we propose an hierarchical approach, where we perform video-level geo-localization and use it to update the gallery of aerial images by selecting top matched large aerial regions (Figure 2B). The reduced gallery helps improve the clip-level geo-localization by screening out some of the visually similar however incorrect aerial images.

4.1 GAMa-Net: *Clip-level Geo-localization*

The proposed network takes as input a short clip from a ground video and matches it with corresponding aerial image. An overview of the proposed network is shown in Figure 3.

Visual encoders In GAMa-Net, we have a video encoder i.e. Ground Video Encoder (GVE) to get features from ground video frames and an image encoder for aerial image features i.e. Aerial Image Encoder (AIE). GVE uses 3D-ResNet18 as backbone and a two layer transformer encoder. Given a ground video C, GVE provides features for a 8 frame clip C_i at the i^{th} second of the video. There is a skip-rate of one frame so we are covering around half a second

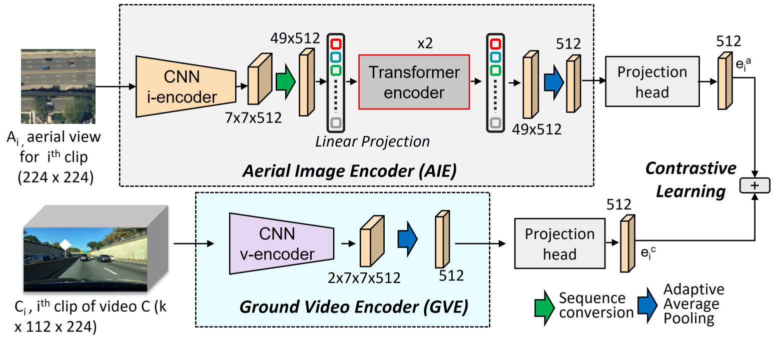

Fig. 3. Network diagram of **GAMa-Net** proposed for clip-level geo-localization. We use 3D-CNN as our base network for learning features from a ground clip (around 0.5sec). Similarly, for aerial image features, we use a 2D CNN backbone. Since only some parts of aerial images are covered by the video feed, using a transformer encoder improves the learning. Number of frames, k=8 with skip rate=1

in the clip, C_i. AIE on the other hand uses 2D-ResNet18 as the backbone. It takes the corresponding aerial image, A_i as the input and learns the features to match with that of the ground video.

Transformer encoding: In the ground video all the visual features are not of equal importance for matching with the aerial view. This is true for aerial images as well and few features are more important when matching with ground videos. For example, the top of a building as seen from aerial images is not visible from a ground video and hence cannot be used to match the pair. To address this, we experimented with multi-headed self attention. We input the features obtained from convolutional networks to a transformer encoder framework which comprises of 4 heads and 2 layers, and uses positional encoding. A small neural network i.e. projection head is used to map the representations to the space where contrastive loss is applied. We use a MLP with one hidden layer to obtain the ground video and aerial image feature vectors ${e^c}_i$ and ${e^a}_i$, respectively.

Image-video contrastive loss: We utilize contrastive loss formulation, base on NT-Xent [4,15,24], to train our network. For a given ground video the corresponding aerial image is considered a positive sample and rest of the samples are considered as negatives. This is a image-video contrastive loss applied on features from two different visual modalities i.e. ground videos and aerial images. In the loss formulation, the focus is on reducing the distance between the positive pair. We have a total of $2(N)$ data points in any mini-batch with N examples. The image-video contrastive loss for a pair of positive examples is defined as,

$$l_{c,a} = -\log \frac{\exp(\mathrm{sim}({e^c}_i, {e^a}_i)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}({e^c}_i, {e^c}_k)/\tau) + \sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}({e^c}_i, {e^a}_k)/\tau)}, \tag{1}$$

where ${e^c}_i$ and ${e^a}_i$ is a positive pair, ${e^c}_i$ and (${e^c}_k$ or ${e^a}_k$) are negative pairs, $\mathbb{1}_{[k\neq i]} \in \{0,1\}$ is an indicator function with value 1 if $k \neq i$, τ is a temperature parameter, sim is the cosine similarity between a pair of features. The final loss is computed for all the positive pairs, both (e^c, e^a) and (e^a, e^c).

4.2 Hierarchical approach

In this approach, we introduce video-level geo-localization in contrast to clip-level geo-localization which also helps in reducing the search space for GAMa-Net (Figure 2B). The clips of a given video are temporally related and provide contextual information to help improve the geo-localization. Similarly, while looking at this problem from aerial image view point, the sequence of aerial images corresponding to clips from a given video are also related geographically and some of the correct prediction at clip-level can be used to update the gallery. Using a smaller gallery also reduces the possibility of error in feature matching.

Approach: We have four steps in this approach. In Step-1, we use GAMa-Net which takes one clip (0.5 sec) at a time and matches with a small aerial image. Using multiple clips of a video, we get a sequence of aerial images for the whole video, i.e. around 40 images. In Step-2, we use predictions of aerial images and match them to the corresponding larger aerial region. We use a screening network to match the features in the same view i.e aerial view. In Step-3, we use screening network predictions to reduce the gallery size (i.e. search space) by selecting top ranked large aerial regions corresponding to a video. These large aerial regions define our new search area for a given video. In Step-4, we use GAMa-Net i.e. the same network as in Step-1, however localize using the updated gallery. More explanation in supplementary.

Visually correct predictions from Step-1 which may not be the ground truth are used to reduce the gallery using this approach. This approach helps improve the clip-level geo-localization since the reduction of the gallery is based on the fact that all the clips of a given video are geolocated nearby. Thus, the aerial images predicted by GAMa-Net can be used to find that large aerial region where all these clips have been captured. In this case, the probability of finding all those visually correct aerial images in the same region is higher when we are searching at the correct geolocation. Hence, it is likely to match with correct large aerial region provided that meaningful and enough information is available.

Screening Network: Video-level Geo-localization We use the screening network to match a sequence of smaller aerial images with the corresponding larger aerial region. The network is similar to GAMa-Net, however the sequence of aerial images is input to Small Aerial Encoder (SAE) with 2D-ResNet18 backbone and the sequence of features are later averaged to obtain a 512D feature vector (Figure 4A). We also experimented with a 3D-ResNet18 backbone which is discussed later. The feature vector thus obtained is matched with the feature vector of corresponding large aerial image from Large Aerial Encoder (LAE). We apply contrastive loss similar to Equation 1. The predictions from this network are used to update the gallery for GAMa-Net.

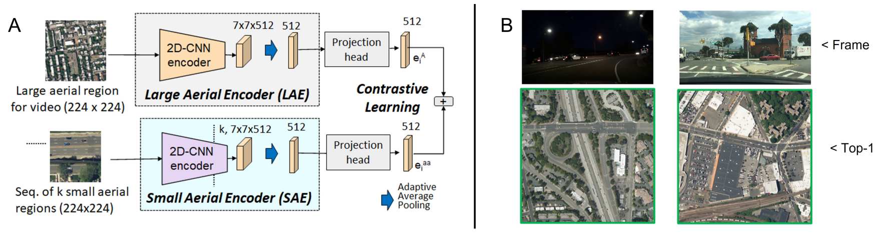

Fig. 4. A) Network diagram of the Screening network used for video-level geo-localization in our proposed Hierarchical approach. B) Predictions at video level (2 samples), where a representative frame of video (top-row) is shown along with correct matched large aerial region (bottom-row)

5 Experiments and Results

Implementation and training details: We implement the proposed GAMa-Net and screening network using PyTorch [15] and train the networks using Adam optimizer with a learning rate (lr) of 8e-5. We use a lr scheduler with lr decay rate of 0.1. The screening network is trained in two step; first with the actual ground truth sequences, and then finetuned with predictions from GAMa-Net. The finetuning step allows the network to adapt to noisy aerial sequences which will be used during inference. We use pre-trained weights from Kinetics-400 for 3D-ResNet18 and ImageNet weights for 2D-ResNet18. The ground videos in the proposed dataset are from different times of the day. For faster training, we used **only day** videos in our experiments unless stated otherwise.

Evaluation: We use top-k recall for clip-level and video-level matching at different values of k. Given a video query, its closest k reference neighbors in the feature space are retrieved as predictions. Similar to image based geo-localization methods we use recall rates at top-1, top-5, top-10, and top 1%. More details can be found in [29,8,22]. We have UCN and CN sets of aerial images corresponding to each clip. In UCN set there is one-to-one correspondence between the clip and aerial image. The GPS point can be anywhere within the aerial image however in the CN set there is an overlap among the aerial images. To keep the evaluation similar to UCN set it is considered a correct match if the predicted GPS is within the range of 0.05 mile of the correct location. We also report top-1@t rate where t is a distance threshold to be used for correct prediction.

Baselines: We propose several baselines for comparison. For ground image based baselines we use the center frame of the clip as an input. Image-CBn, our proposed baseline uses 2D-CNN ResNet18 model to encode the ground video frame with similar contrastive loss formulation as GAMa-Net. We use two different loss functions for video based baselines (Triplet-Bn and CBn); margin triplet loss [28] and contrastive loss [4]. In these baselines, we utilize 2D-CNN ResNet18 for aerial images and 3D-CNN ResNet18 for ground videos.

Table 3. Comparison with proposed baselines; CBn(Contrastive Baseline) & Bn(Baseline). Recall rates in parentheses() are for uncentered set. Top three rows show results with image based methods

Model	Video	R@1	R@5	R@10	R@1%
Image-CBn	x	9.5(1.5)	18.8(5.5)	24.6(9.0)	87.7(50.7)
Shi et al.[23]	x	9.6	18.1	26.6	71.9
L2LTR [32]	x	11.7	20.8	28.2	87.1
Triplet-Bn	✓	<1(<1)	<1(1.2)	1.1(2.4)	33.1(33.8)
CBn	✓	11.5(3.2)	21.7(10.9)	27.8(16.7)	89.2(65.0)
GAMa-Net	✓	15.2	27.2	**33.8**	**91.9**
GAMa-Net (Hierarchical)	✓	**18.3**	**27.6**	32.7	-

5.1 Results

The evaluations of GAMa-Net and a comparison with baselines is shown in Table 3. We observe a Top-1 recall rate of 18.3% and 15.2%, using GAMa-Net with and without the hierarchical approach, respectively. On UCN set, we observe poor performance as compared to CN set which was expected.

Comparison: In first row, we show results with our image-based baselines which uses a single frame from the ground video. We also compare with other image based methods using a single frame as input and observe that the proposed approach outperforms all these baselines. While training with a single image is faster we do not have the temporal information which can be perceived as relative positioning or contextual information. As the camera moves along a path or trajectory we can see the buildings or objects pass-by giving an idea of their respective location. The information of distance/relative-positioning as seen by a 3D-CNN is the additional information when we train with videos. As shown in the Table 3, using a video provides better results as compared to images.

Few studies have reported better performance with contrastive loss [16], however triplet loss is also frequently used for geo-localization [36,18]. We also observe better performance with contrastive loss as compared to triplet loss (Table 3). In GAMa-Net, we have features from two different visual modalities i.e. one from ground **videos** and other from aerial **images**, which is different from traditional contrastive loss. Similarly, training with triplet loss also use image-video features and this difference is likely responsible for the poor training with triplet loss.

Qualitative analysis: Figure 5 show sample Top-5 predictions with different models where the leftmost is Top-1 and the rightmost is 5th. The combined model, i.e. GAMa-Net without transformer encoder, makes visually meaningful predictions. In the leftmost example, camera is passing under a fly-over and the predictions show a similar locations. The middle example is of a road without any crossings or red lights in sight, and right-most example is of a city street with crossing markings on road. The predictions by combined model match these

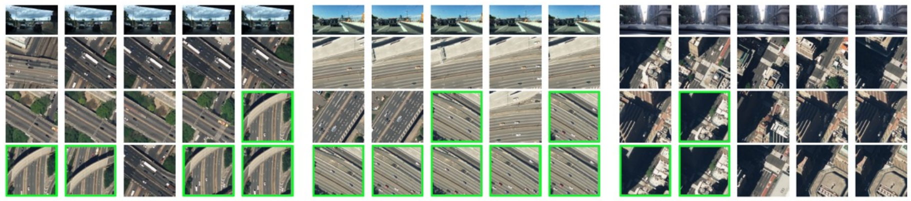

Fig. 5. Geo-localization results for three query clips, using different models. Top-row shows frames of the query clips. Second row is for combined model (Top-5 predictions), and third row is GAMa-Net. Bottom row shows predictions by GAMa-Net with Hierarchical approach (gallery reduced to 1% of larger aerial regions). Correct predictions have a green outline. Owing to close GPS labels there are multiple correct aerial images

Table 4. Screening network matches large aerial regions using a sequence of matched aerial images. In this Table, we compare 3D-CNN and 2D-CNN backbone for aerial image sequence of 8 or 32 images

Model	R@1	R@10	R@1%	R@10%
3D-CNN (seq of 8 aerial)	7.5	24.9	39.8	67.9
3D-CNN (seq of 32 aerial)	3.8	19.8	36.1	74.9
2D-CNN (seq of 8 aerial)	**12.2**	**35.3**	**49.3**	77.2
2D-CNN (seq of 32 aerial)	8.2	20.9	29.3	**83.8**

specifications. However, in these samples, the ground truth is in top-1% but not in top-5 images. The predictions by GAMa-Net, with multi-headed self-attention improves the network performance and correct prediction moves up in the top-5.

Hierarchical approach The predictions from the screening network are used for video-level geo-localization and to reduce the gallery for GAMa-Net. We experimented with both 2D-CNN and 3D-CNN backbone for aerial image sequence. As shown in Table 4, better results were achieved with 2D-CNN backbone likely due to better spatial features. A comparison of number of input aerial images show that 2D-CNN with 8 frames retains more ground truths at Top-1 (12.2%), Top-10 (35.3%) and Top-1% (49.3%). We do not have 32 samples in some of the samples which likely effects the performance. However, when we evaluated GAMa-Net after gallery reduction, Screening network i.e. **2D-CNN network with 32 aerial images** provided the best results. Figure 4B shows qualitative results for video-level geo-localization. It shows one frame of the video input to GAMa-Net, the predicted seq of aerial images thus obtained was used to identify the larger aerial regions. The bottom row shows the correct matched large aerial regions for video-level geo-localization. More results are in supplementary.

Gallery sizes: In Table 5, we discuss the results with GAMa-Net using various gallery sizes, i.e. Top-10, Top-1%, and Top-10% of large aerial regions,

Table 5. Results with **GAMa-Net**, without and with the hierarchical approach

Gallery size	R@1	R@5	R@10	R@1%
Full	15.2	27.2	33.8	**91.9**
	With Hierarchical Approach			
Top-10	16.2	22.8	25.5	-
Top-1%	**18.3**	27.6	32.7	-
Top-10%	16.6	**28.2**	**34.6**	76.6

Table 6. Results with our best model and ablations for comparison. Here we show Top-1@threshold for comparison, with threshold values of 0.1, 0.2, 0.5 and 1.0 mile. CBn-UCN and CBn-CN, represent the contrastive baselines with uncentered and centered aerial images, respectively. Hierarchical approach uses one percent of the gallery

Model	Recall @							
	Top-1	Top-5	Top-10	Top-1%	Top-1@0.1	Top-1@0.2	Top-1@0.5	Top-1@1.0
CBn-UCN	3.2	10.9	16.7	65.0	3.6	5.9	11.3	18.6
CBn-CN	11.5	21.7	27.8	89.2	15.7	19.0	25.1	32.1
Combined	11.6	23.4	30.4	**92.4**	15.6	19.0	25.0	32.7
Video-Tx	14.1	25.4	31.8	90.8	18.7	22.1	28.1	35.7
GAMa-Net	15.2	27.2	**33.8**	91.9	19.6	23.0	28.7	36.1
Dual-Tx	14.6	26.1	32.7	91.8	19.0	22.2	28.2	35.5
Hierarchical	**18.3**	**27.6**	32.7	-	**23.5**	**27.8**	**34.9**	**43.6**

identified using screening network. We see better Top-1 results with Top-1% gallery as compared to Top-10 and Top-10%. There is a trade-off between reduced search space which improves matching and retaining the ground truth. As observed from Table 4, in Top-10 we have ground truth for only 20.9% of videos which increases to 83.8% at 10% gallery. However, this percentage will be different if we consider clips since we get different number of clips from each video, as per GPS labels. We evaluated GAMa-Net, which is clip level, using a reduced gallery. Even when ground truth was not available for many samples we see a Top-10 recall rate of 25.5%. When gallery is reduced to Top-10 and Top-1% of larger aerial regions we do not have enough clips to make upto 1% of the total gallery. In Figure 5, we show three example predictions and it is evident that results by GAMa-Net are improved with hierarchical approach.

5.2 Ablations

Combined model: In Table 6, we observe better performance with combined model as compared to baselines, and a Top-1% recall of 92.4%. As expected, the performance improves as we increase the distance threshold. In the combined model, we have various augmentations which includes spatial and temporal centering, and random crop. Since images are not aligned in GAMa we stochastically rotate the aerial view (0, 90, 180, 270 deg) for view-point invariance. All these augmentations have been reported to help improve geo-localization. We also include hard negatives for better training however transformer encoder is not a part of combined model.

Transformer encoder: Similar to an aerial image, not all visual features in the ground video have the same importance for cross-view geo-localization. Thus, we implemented transformer(Tx) encoder on ground video(Video-Tx) and aerial images(Aerial-Tx), individually. In both cases, we observe an increase in recall @Top-1. The performance is better with GAMa-Net(i.e. Aerial-Tx) and we observe 15.2% Top-1 recall which is higher than 14.1% with Video-Tx (Table 6). Observing an improvement in both cases we used transformer encoder on both sides(Dual-Tx) however it did not perform better than the GAMa-Net.

Hierarchical approach: We use hierarchical approach to improve the performance of GAMa-Net by reducing the gallery. Top-1@threshold (Table 6) shows that using the hierarchical approach makes the matching more effective by predicting the aerial images closer to the correct GPS location or ground truth. The **difference** in Top-1 recall, with and without hierarchical approach, is even higher at higher thresholds i.e 7.5% at 1.0 mile vs 3.9% at 0.1 mile. An increase in recall to 43.6% at 1.0 mile threshold shows that the ground truth is not very far from the Top-1.

6 Discussion

Videos and contextual information: In videos, we have more information which can be considered as contextual information from geo-localization point of view. It enables the network to locate a given frame or view with respect to the other frames in the video. One frame may contain features to complement another and together they are likely to provide more or complete information required for geolocating. In Table 3, we have compared video based method with frame based baselines. Better recall rate with videos show that the network is able to utilize the additional information available with videos.

Centered vs Uncentered: In an UNC aerial image the corresponding GPS point can be anywhere in the tile. In cases where the GPS point is near the boundary, the visual information from the video is less likely to match the corresponding tile. As expected, after centering (CBn-CN) we observed an improvement in the recall rate as compared with the uncentered set (i.e. CBn-UCN).

Full Dataset: GAMa dataset is a large dataset which has its pros and cons. With large amount of data networks are better trained however this also increases the training time and memory requirement. Here we discuss results with models trained on full dataset (Table 7) i.e. both day and night videos. There is an improvement of 1-4% in recall at all k. Thus, using the hierarchical approach our best R@Top-1 and R@Top-1@1.0 is 19.4% and 45.1%, respectively.

Table 7. Training on full dataset and evaluation on day videos

Model	Recall @			
	Top-1	Top-1%	Top-1@0.1	Top-1@1.0
Combined	14.7	**94.8**	19.2	36.7
GAMa-Net	17.5	94.7	22.3	39.3
Hierarchical	**19.4**	-	**25.0**	**45.1**

Comparison: Our results are comparable to existing image based methods. One recent study reports **R@1=13.95% on CVUSA** when images are unaligned and have around 70% field of view (FOV) [32]. Most of the cross-view geo-localization datasets such as CVUSA report high R@k while using ground panorama and aligned aerial images; which is unrealistic since a normal camera lens has a FOV of around 72% and alignment is also not possible always. In GAMa dataset, the aerial images and ground videos are **unaligned**, and ground videos have a **limited FOV** which makes it a more realistic and difficult dataset.

Challenges: In GAMa dataset among the two sets of small aerial images, i.e. CN and UCN, the UCN set is more realistic however difficult for geo-localization. Lack of alignment increase this difficulty however, the orientation information can be extracted from the GPS information and is likely to improve the performance [13]. Also, the **ground videos** have varying lengths and in some videos GPS label is not available every second. Thus, the video length available for geo-localization is less than 40 sec. Additionally, there is occlusion because of the car hood and other objects. Such cases are more likely to appear in fail cases.

Limitations: The proposed hierarchical approach performs better with longer videos (8 sec. or more), however it can be used with shorter clips as well. The aim behind using a hierarchical approach is to filter out confusing samples to improve the retrieval rate. However, this sometimes leads to filtering of the ground truth from the gallery. The model is also likely to fail with indoor videos since it will not be possible to match the features with an aerial image. However, this limitation is common to all cross-view geo-localization methods.

7 Conclusions

In this work, we focus on the problem of video geo-localization via cross-view matching. We propose a new dataset for this problem which has more than 51K ground videos and 1.9 million satellite images. The dataset spans multiple cities and is a more realistic dataset, unaligned and limited FOV. We believe that this dataset will be useful for future research in cross-view video geo-localization. Our proposed GAMa-Net effectively makes use of the rich contextual information available with video. In addition, we propose a hierarchical approach which also utilize the contextual information to further improve the geo-localization.

References

1. Satellite images. https://www.apple.com/maps/, accessed: 2021-01
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
3. Chaabane, M., Gueguen, L., Trabelsi, A., Beveridge, R., O'Hara, S.: End-to-end learning improves static object geo-localization from video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2063–2072 (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. Journal of Field Robotics **37**(3), 362–386 (2020)
6. Hakeem, A., Vezzani, R., Shah, M., Cucchiara, R.: Estimating geospatial trajectory of a moving camera. In: 18th International Conference on Pattern Recognition (ICPR'06). vol. 2, pp. 82–87. IEEE (2006)
7. Hosseinpoor, H., Samadzadegan, F., Dadras Javan, F.: Pricise target geolocation and tracking based on uav video imagery. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences **41** (2016)
8. Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7258–7267 (2018)
9. Hu, S., Lee, G.H.: Image-based geo-localization using satellite imagery. International Journal of Computer Vision **128**(5), 1205–1219 (2020)
10. Kim, D.K., Walter, M.R.: Satellite image-based localization via learned embeddings. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 2073–2080. IEEE (2017)
11. Li, A., Hu, H., Mirowski, P., Farajtabar, M.: Cross-view policy learning for street navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8100–8109 (2019)
12. Lin, T.Y., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5007–5015 (2015)
13. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5624–5633 (2019)
14. Miller, I.D., Cowley, A., Konkimalla, R., Shivakumar, S.S., Nguyen, T., Smith, T., Taylor, C.J., Kumar, V.: Any way you look at it: Semantic crossview localization and mapping with lidar. IEEE Robotics and Automation Letters **6**(2), 2397–2404 (2021)
15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)

16. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)
17. Regmi, K., Borji, A.: Cross-view image synthesis using geometry-guided conditional gans. Computer Vision and Image Understanding **187**, 102788 (2019)
18. Regmi, K., Shah, M.: Video geo-localization employing geo-temporal feature learning and gps trajectory smoothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12126–12135 (2021)
19. Rodrigues, R., Tani, M.: Are these from the same place? seeing the unseen in cross-view image geo-localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3753–3761 (2021)
20. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
21. Senlet, T., Elgammal, A.: Satellite image based precise robot localization on sidewalks. In: 2012 IEEE International Conference on Robotics and Automation. pp. 2647–2653. IEEE (2012)
22. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. Advances in Neural Information Processing Systems **32**, 10090–10100 (2019)
23. Shi, Y., Yu, X., Campbell, D., Li, H.: Where am i looking at? joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4064–4072 (2020)
24. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)
25. Tian, X., Shao, J., Ouyang, D., Shen, H.T.: Uav-satellite view synthesis for cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology (2021)
26. Tian, Y., Chen, C., Shah, M.: Cross-view image matching for geo-localization in urban environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3616 (2017)
27. Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6488–6497 (2021)
28. Vassileios Balntas, Edgar Riba, D.P., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: Richard C. Wilson, E.R.H., Smith, W.A.P. (eds.) Proceedings of the British Machine Vision Conference (BMVC). pp. 119.1–119.11. BMVA Press (September 2016). https://doi.org/10.5244/C.30.119, `https://dx.doi.org/10.5244/C.30.119`
29. Vo, N.N., Hays, J.: Localizing and orienting street views using overhead imagery. In: European conference on computer vision. pp. 494–509. Springer (2016)
30. Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zhenga, B., Yang, Y.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology (2021)
31. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
32. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. Advances in Neural Information Processing Systems **34** (2021)

33. 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)
34. Zamir, A.R., Shah, M.: Accurate image localization based on google maps street view. In: European Conference on Computer Vision. pp. 255–268. Springer (2010)
35. Zemene, E., Tesfaye, Y.T., Idrees, H., Prati, A., Pelillo, M., Shah, M.: Large-scale image geo-localization using dominant sets. IEEE transactions on pattern analysis and machine intelligence **41**(1), 148–161 (2018)
36. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)
37. Zhu, Y., Sun, B., Lu, X., Jia, S.: Geographic semantic network for cross-view image geo-localization. IEEE Transactions on Geoscience and Remote Sensing (2021)