Are These from the Same Place? Seeing the Unseen in Cross-View Image Geo-Localization

Royston Rodrigues
Biometrics Research Laboratories
NEC Corporation
Kawasaki, Kanagawa, Japan
r-rodrigues@nec.com

Masahiro Tani
Biometrics Research Laboratories
NEC Corporation
Kawasaki, Kanagawa, Japan
masahirotani@nec.com

Abstract

In an era where digital maps act as gateways to exploring the world, the availability of large scale geo-tagged imagery has inspired a number of visual navigation techniques. One promising approach to visual navigation is cross-view image geo-localization. Here, the images whose location needs to be determined are matched against a database of geo-tagged aerial imagery. The methods based on this approach sought to resolve view point changes. But scenes also vary temporally, during which new landmarks might appear or existing ones might disappear. One cannot guarantee storage of aerial imagery across all time instants and hence a technique robust to temporal variation in scenes becomes of paramount importance. In this paper, we address the temporal gap between scenes by proposing a two step approach. First, we propose a semantically driven data augmentation technique that gives Siamese networks the ability to hallucinate unseen objects. Then we present the augmented samples to a multi-scale attentive embedding network to perform matching tasks. Experiments on standard benchmarks demonstrate the integration of the proposed approach with existing frameworks improves top-1 image recall rate on the CVUSA data-set from 89.84 % to 93.09 %, and from 81.03 % to 87.21 % on the CVACT data-set.

1. Introduction

Consider the panoramic photographs in Figure 1. Is it possible to discover their location? A classical way to determine their location is to match them against a database of geo-tagged aerial images covering a broad geographic region. An accurate match would successfully localize these photographs. On the face of it, matching images across drastically different viewpoints seems intimidating, since visual appearances and spatial correspondence of the two views lie far apart. Recent progress in deep learning [12, 20, 18] shows that it is possible to match images across different views and localize panoramic images satisfactorily. More often, the cross-view image geo-localization problem is posed as an image retrieval task [12, 20, 18]. A common setup involves a training objective to generate feature embeddings of image pairs from the same place closer while those from different places far apart, solely based on visible image content.

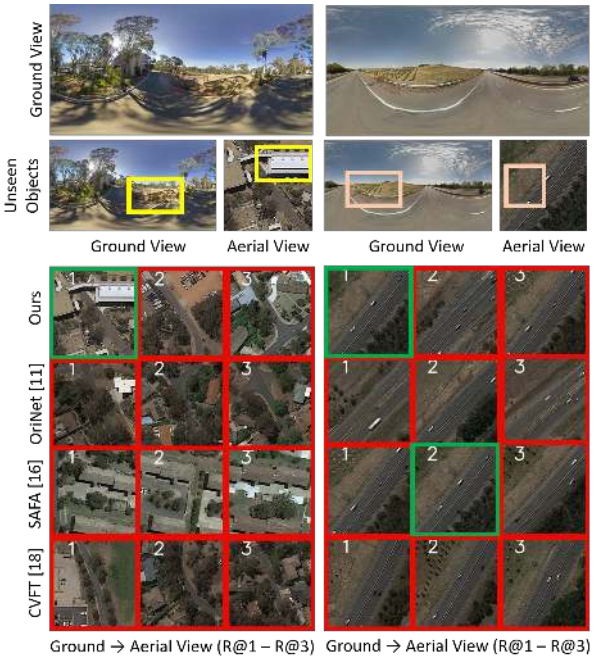

Figure 1: Given a ground view panorama, we retrieve corresponding aerial views from a large database of geo-tagged images (*Top*). Note that there exists time differences between the two views, during which new buildings appear and vegetation patterns change (*Middle*). Our method retrieves the correct aerial view and is robust to these unseen changes (*Bottom*).

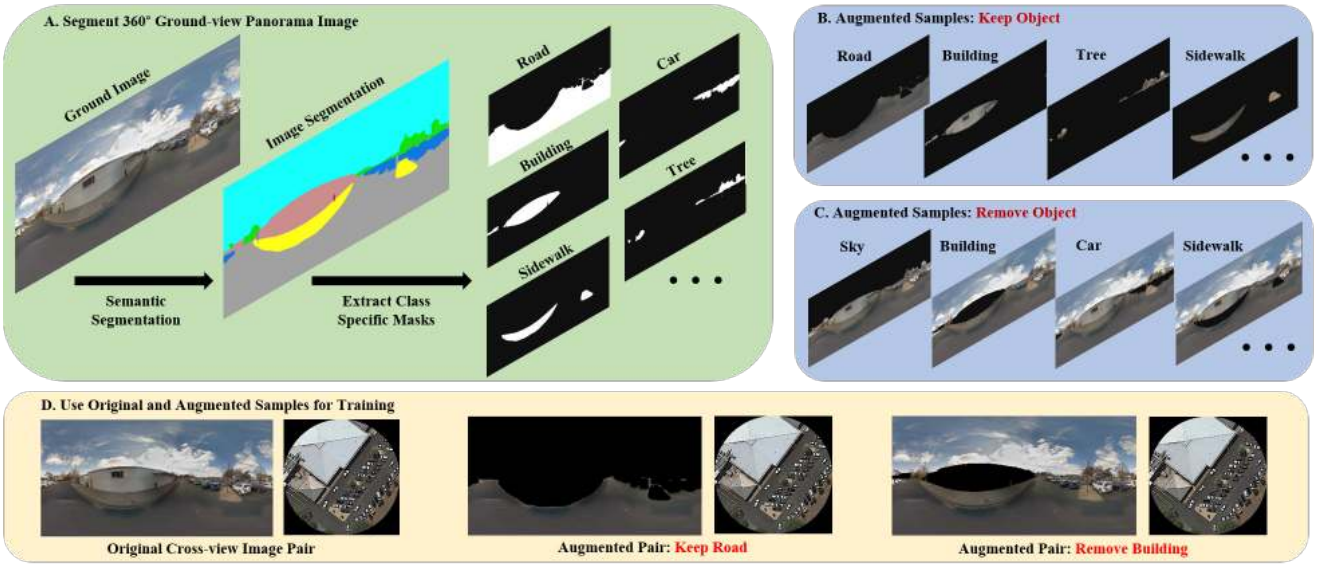

Figure 2: We propose a semantically driven data augmentation technique to facilitate unseen object matching. An off-the-shelf image segmentation module [28] is used to obtain class specific masks corresponding to the ground view image (*Green*). These masks are used to create augmented samples in *keep* and *remove* mode respectively (*Blue*). The augmented samples along with their original versions are used for training into a Siamese pipeline (*Yellow*). Note that the corresponding satellite image is kept unchanged. This is done to enable the matching network in figure 3 to match available regions and hallucinate unavailable ones to appropriate locations in the aerial view.

One major limitation of such an approach is that it does not take into account the temporal variation of places. For instance in figure 1 after the photographs have been captured new buildings might appear or vegetation patterns might change. This would mean capturing aerial images from every possible location and at every possible time. Practically its impossible to store such an enormous data-set, limiting the applicability of existing cross-view image geo-localization methods.

In this paper we extend the capabilities of standard cross-view image geo-localization systems by developing an image matching module capable of matching unseen objects. Such a module is important to match images with significant time differences and could save us from the herculean task of having aerial images from every possible location and at every possible time.

Might there be a way to simulate the disappearing and newly appearing characteristics of objects in temporally varying scenes, while being able to preform successful matching? We propose a data augmentation approach to the cross-view image matching paradigm: employing semantic segmentation to mask out image regions to *keep* and *remove* objects as indicated in figure 2. Since the objects being masked out can vary at different spatial scales (*i.e.*, large spatial scale for objects near the camera and small spatial scale for objects away from the camera), the samples generated are presented to a multi-scaled attention module. This extracts feature embeddings to perform cross-view image matching across different image scales. The proposed approach induces the abstraction of disappearing and newly appearing objects in temporally varying scenes, while being able to match images across the two views successfully.

Apart from solving the problem of matching across temporal variation in scenes, equipping cross-view image matching networks with hallucinating unavailable regions also benefits matching areas that are occluded or covered due to shadows (See Figure 8).

The contribution of this work are as follows:

- We propose a new data augmentation pipeline to address the challenge of temporal variation in scenes for cross-view image geo-localization. This gives matching networks the ability to hallucinate unseen objects and perform unseen object matching.

- We present a multi-scale attention module for matching task, our ablation studies demonstrate the effectiveness of using multiple scales compared to operating at a single scale.

- We confirm the efficacy of our approach by conducting extensive experiments on two established benchmarks. Integrating our approach with existing frameworks increases top-1 recall rates on the CVUSA [26] data-set from 89.84 % to 93.09 %, and from 81.03 % to 87.21 % on the CVACT [12] data-set.

2. Related work

In this section we briefly review existing works on cross-view image geo-localization.

Cross-view image geo-localization has been explored before prevalence of the deep learning age. Initial works by [4, 9, 17, 2] extracted meaningful handcrafted features for this task. Recently, the effectiveness of deep learning has motivated deep learning based approaches for this research. [24] provided a framework to make use of a pre-trained CNN by discovering location specific semantic information in deeper layers of a CNN. [25] studied the effect of fine-tuning pre-trained CNNs by imposing an objective function to place feature vectors of images from the same location together. [10] used a Siamese style architecture with a contranstive loss function [6]. [8] demonstrated the effectiveness of incorporating NetVLAD [1] for cross-view image matching tasks. [12] incorporated orientation based priors into the image matching process. This was accomplished by generating color-coded angle maps, encoding azimuthal and altitude angles. The color coded maps along with the original RGB image was fed in as a six channel signal for matching . [20] made use of an optimal feature transport strategy to align ground and aerial views in feature space. This lead to a more meaningful feature similarity and improved retrieval rates by a significant margin. [18] reduced the geometric gap between ground and aerial views by introducing an effective polar transform. The transform was developed by simple pixel re-ordering, it considered pixels lying on the same azimuth direction in an aerial view corresponded to pixels from the same vertical image column in the ground view. [18] also proposed a spatial aware embedding module to determine a spatially dependent embedding map. This lead to significant improvements in the recall rates for cross-view image geo-localization. [15] proposed an image generation approach to reduce the visual differences between the two views, this was done by using conditional GANs [14]. Their approach could generate plausible aerial view images from corresponding ground view images and consider it for matching. [19] considered challenges involved in cross-view image matching when moving from a full field of view configuration to a limited field of view setup. [27] used a classification based approach. [21] considered predicting orientation by incorporating regression. [22] used a square ring partition of the signal and learnt part-wise representation to encode context information. Our approach is different from the above approaches as we consider the problem of temporal variation in scenes. To the best of our knowledge this is the first work that considers scene changes with time and equips cross-view image matching networks with the capability to match unseen objects.

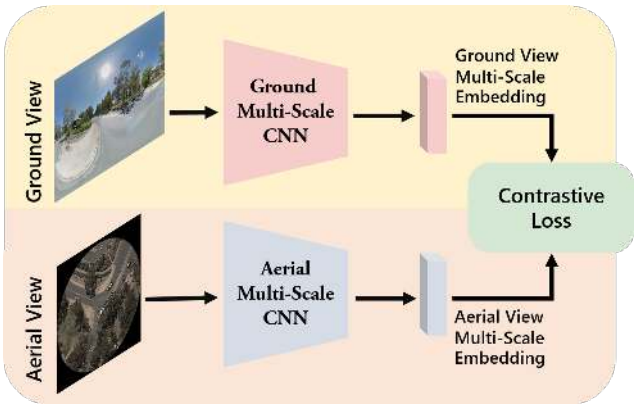

Figure 3: Our Siamese framework: The inputs to the two multi-scale CNN (Figure 4) branches are ground view and aerial view images, respectively. The Multi-Scale embedding vectors corresponding to ground and aerial views are learnt by a contrastive loss function.

3. Approach

In this section, we introduce our proposed object based data augmentation method (Figure 2), we then present our multi-scale attentive backbone network that is used within a Siamese framework (Figure 3).

3.1. Object based data augmentation

In order to account for the time differences in scenes, we propose to equip Siamese networks with the ability to hallucinate unseen objects while being able to perform the matching task successfully. We enforce such capabilities into our Siamese pipeline by proposing an object based data augmentation approach. We make use of a semantically driven data augmentation technique to simulate the disappearing and newly appearing object peculiarities in temporally varying scenes (Figure 2).

We make use of an off-the-shelf semantic segmentation module [28] for obtaining segmentation maps corresponding to ground view. The classes that we considered for this task were building, car, grass, ground, road, sidewalk, tree and sky. These masks are used to create augmented samples in *keep* and *remove* mode respectively. Original pixel values for a particular class are retained in keep mode whereas they are masked by black pixels when operated in *remove* mode. All object classes were considered for the remove mode, but the classes sky and car were not considered in keep mode. This is because regions from these classes are not available in the aerial view for matching. When the ground view is modified, we do not modify the aerial view. This is done to enforce matching networks with the ability to identify what parts are visible and what parts need to be hallucinated to carry out cross-view image matching successfully.

A similar object driven data augmentation method can be performed in the aerial view, however we decided not carry it out since no off-the-shelf semantic segmentation module transferred well on the aerial images of CVACT [26] and CVUSA [12] datasets.

3.2. Proposed model

As illustrated in figures 3 and 4 we make use of a Siamese network pipeline trained with a contrastive loss function (Eq.1), to solve for our image retrieval task. The two CNN branches corresponding to ground and aerial views have the same architecture (Figure 4) but are distinct and do not share weights.

Input representation: We use a spatial resolution of 416 x 208 pixels for the ground view image and 300 x 300 pixels for the aerial view image. The aerial view image is circular cropped with a radius of 150 pixels (Figure 3). This done to exploit the 360° nature of the panoramic ground view. As rotation around the center in the aerial view corresponds to a circular shift along the horizontal axis in the ground view. Pre-processing the inputs in such a manner helps create more samples for training by merely rotating the aerial view along its center and circular shifting the ground view by a corresponding amount. This was done to facilitate extra sample creation by incorporating such a style of data representation.

Multi-scale CNN backbone: The backbone network used to extract feature representations is a modified version of the ResNet-18 [7] network pre-trained on Imagenet [5]. In-order to capture features from different spatial scales, we consider intermediate outputs of the ResNet-18 network representative of features from different scales. Early layers of the network are considered as low-scale feature extractors and later layers of the network are regarded as high-scale feature extractors. Our approach is motivated by [16] here we adapt it for the two dimensional case. This is different from common multi-scale approaches such as feature pyramid [11] used for object detection. Considering intermediate outputs of deep neural networks leads to capturing features at different spatial scales. These intermediate values are passed through independent convolutional block attention modules [23], in-order to apply attention along the spatial and channel dimensions. This is done so that spatial and channel correspondences at a particular scale can be maintained across the two views by selecting salient features while suppressing the features that are not relevant at that scale.

Feature aggregation: We also consider aggregating multiple attentive features by using multiple convolutional block attention modules [23] and extracting multiple features at each scale. Our best performing system makes use of three such modules at every scale. Our aggregation method is different from [13] as we use attention.

Parameter initialization: The learnable parameters of ResNet-18 are initialised from Imagenet pre-training, while that of the convolutional block attention module are randomly initialised. Our network is trained end to end, all the parameters of the model are updated during backward pass.

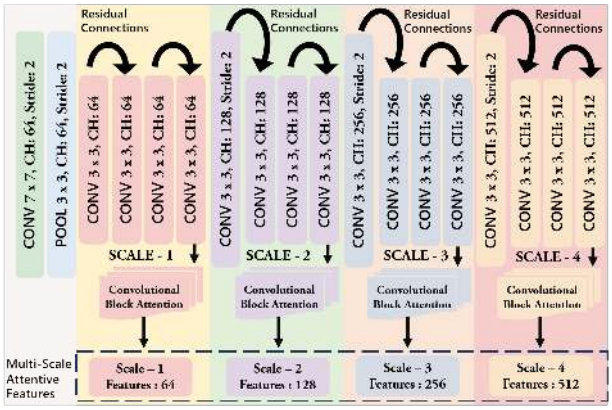

Figure 4: Architecture details of the proposed multi-scale CNN model. Intermediate results of a residual neural network are consider as multi-scaled features. Outputs of the neural network before stride 2 operations are tapped for representations at individual scale. The curved lines with arrows represent residual connections. Convolutional block attention module [23] is used to en-corporate channel and spatial attention. Multiple attentive features are captured by using multiple such modules at each scale.

3.3. Objective function

We employ a metric learning objective to learn multi scale feature representations for both ground and aerial views. One common objective is the contranstive loss (Eq. 1 and 2). The aim of contranstive loss is to learn representations that bring matching pairs closer and drive non-matching pairs away. We use the below objective function:

$$L = (1 - Y)\frac{1}{2}(D_w)^2 + (Y)\frac{1}{2}\{\max(0, m - D_w)\}^2 \quad (1)$$

$$D_w = ||F_g(I_g) - F_a(I_a)||^2 \quad (2)$$

where, m is a margin parameter. F_g is a multi-scale ground CNN responsible to extract image features from ground view image I_g. F_a is a multi-scale aerial CNN responsible to extract image features from aerial view image I_a. D_w represents euclidean distance and Y is a binary value indicating weather the ground view image I_g and aerial view image I_a belong to the same pair or not.

4. Experiments

We demonstrate the effectiveness of our proposed approach with qualitative and quantitative experiments. Qualitatively, we show some success and failure cases of our method in comparison to SAFA [18] (See Figure 8). Our method can match unseen objects that might newly appear due to the presence of time differences between the two views. Our method also performs well in the case of occluded landscape patterns and road structures. SAFA [18] does well when all objects are present in both the views.

Data-sets: Our method is evaluated using two standard benchmark data-sets *i.e*. CVUSA [26] and CVACT [12]. Here ground views consist of 360 degree panorama images which are matched against a large data-set of geo-tagged aerial views. The split for each of the two data- sets consists of 35,532 training image pairs and 8,884 validation image pairs, respectively. We also report the performance of image geo-localization at scale by evaluating on an additional large scale test split of CVACT [12] consisting of 92,802 image pairs. The ground and aerial view images from the above data-sets are captured at different time.

Figure 5: Ground view (*left*) and corresponding aerial view (*right*) from CVACT (*top*) and CVUSA (*bottom*).

Implementation Details: We make use of ResNet-18 model initialised with pre-trained Imagenet weights. Outputs from intermediate layers of this model are extracted to capture features at multiple scales. Features at each scale are passed through a convolutional block attention module which applies attention across both channel and spatial dimensions. The networks are trained end-to-end. We make use of adam as our optimiser and use 10^{-4} as our learning rate. The margin m (Eq. 1) in our contrastive loss framework is set to 1. We make use of B positive image pairs and $B(B-1)$ negative image pairs in a single batch during training. B is set to 128 for our experiments. To accommodate a large batch size we make use of gradient accumulation. This involves making optimizer updates after several forward and backward passes and accumulating the gradients for each pass, until the effective batch size is attained.

Table 1: Comparison with existing works on CVACT-val [12]. R@K indicates top-K retrieval rate.

Methods	CVACT-val [12]			
	R@1	R@5	R@10	R@1%
CVM-Net [8]	20.15	45.00	56.87	87.57
OriNet [12]	46.96	68.28	75.48	92.01
CVFT [20]	61.05	81.33	86.52	95.93
SAFA [18]	81.03	92.80	94.84	98.17
Ours	73.19	90.39	93.38	97.45
Ours + SAFA [18]	**87.21**	**95.01**	**96.39**	**98.69**

Our method was implemented on a single NVIDIA Quadro RTX 8000 GPU and it took a week of training time.

Table 2: Comparison with existing works on CVUSA [26]. We use '-' when the metric for comparison is unavailable.

Methods	CVUSA [26]			
	R@1	R@5	R@10	R@1%
MCVPlaces [25]	-	-	-	34.30
Zhai [26]	-	-	-	43.20
Vo [21]	-	-	-	63.70
CVM-Net [8]	22.47	49.98	63.18	93.62
OriNet [12]	40.79	66.82	76.36	96.12
Zheng [27]	43.91	66.38	74.58	91.78
Regmi [15]	48.75	-	81.27	95.98
Siam-FCANet [3]	-	-	-	98.30
CVFT [20]	61.43	84.69	90.49	99.02
SAFA [18]	89.84	96.93	98.14	99.64
Ours	75.95	91.90	95.00	99.42
Ours + SAFA[18]	**93.09**	**98.14**	**98.94**	**99.80**

Evaluation Metric: We follow top-K as our evaluation metric similar to [8, 12, 20, 18] and compare it with the methods [25, 26, 21, 8, 12, 27, 15, 3, 20, 18]. We consider a ground view image as correctly localized if its corresponding aerial view is within the nearest top-K retrieved images.

4.1. Comparison with state-of-the-art methods

We compare our method with [25, 26, 21, 8, 12, 27, 15, 3, 20, 18] on CVUSA [26] and with [12, 20, 18, 8] on CVACT [12]. We make use of codes open sourced by corresponding authors for the case of fusion with SAFA [18]. We report recall rates at top-1, top-5, top-10 until top-1% and enumerate it in tables 1 and 2. It can be seen in tables 1 and 2 that integrating our approach with SAFA [18] provides a large performance gain. This can also be observed by the recall rates in figures 6 and 7.

Fusion with SAFA [18]: Our goal (*i.e*. matching unseen objects) being complementary to the objective of SAFA (*i.e*. matching visible objects), we consider fusion of the retrieval scores from the two systems. We use a late fu-

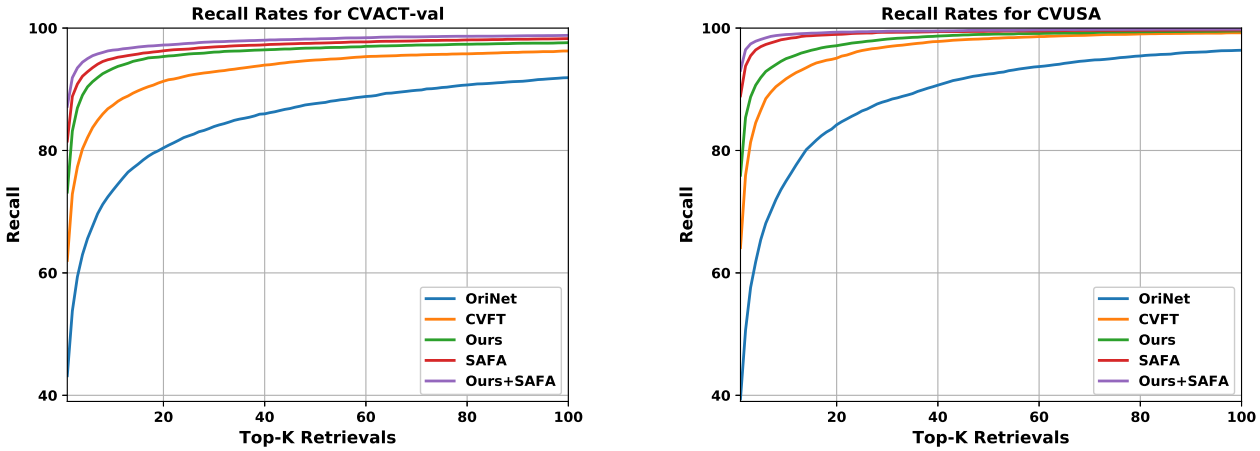

Figure 6: Recall rates on CVACT-val [12] and CVUSA [26]. Integrating our approach with SAFA [18] offers better recall.

sion approach which is performed when the two systems are operated in test mode. The final retrieval score is a weighted combination of retrieval scores obtained from individual systems. We allow equal contribution from both the systems and set the weight as 0.5 for our experiments.

Results on CVACT-val: As indicated in table 1, we compare the proposed method with other existing approaches on CVACT-val. The proposed method achieves 73.19% top-1, 90.39% top-5, 93.38% top-10 and 97.45% top-1% retrieval rates. Our method surpasses most existing methods [12, 20, 8]. SAFA [18] when operated independently performs best compared to all methods. We report an increase in the top-1 retrieval rate from 81.03% to 87.21% when combined with SAFA [18]. Recall rates in figure 6 demonstrates the performance of our system compared to existing approaches.

Results on CVACT-test: CVACT-test [12] is a densely sampled cross-view image geo-localization data-set and includes 92,802 image pairs from Canberra, Australia. This benchmark is representative of cross-view image geo-localization at scale. In figure 7 we show recall performance for top-K retrievals. We compare our method with [18, 12, 20]. Integrating our approach with SAFA [18] leads to a significant improvement in performance.

Results on CVUSA: We compare the performance of our method with other competent methods on the CVUSA [26] data-set in table 2. The proposed method achieves 75.95% top-1, 91.90% top-5, 95.00% top-10 and 99.42% top-1% retrieval rates. We perform well compared to [25, 26, 21, 8, 12, 27, 15, 3, 20]. SAFA [18] when operated independently performs best compared to all methods. Fusion of the retrieval scores from the two systems increases the top-1 retrieval rate from 89.84% to 93.09%. Recall rates on CVUSA [26] are indicated in figure 6.

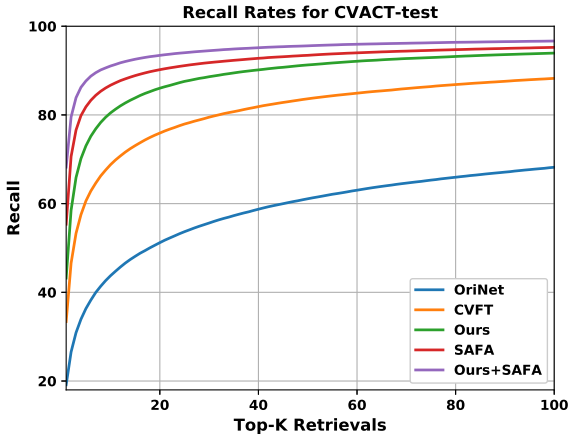

Figure 7: Recall rates on the large-scale CVACT-test [12] split. Combining our approach with SAFA [18] provides large performance gains.

4.2. Ablation Study

In this section, we provide ablation studies performed on CVUSA [26] and CVACT [12] data-sets. In particular, we compare our proposed object based data augmentation method with standard data augmentation techniques. We also highlight the importance of using multi-scaled attention modules for cross-view matching tasks.

Effects of Data augmentation: For illustrating the effects of data augmentation methods on cross-view image geo-localization performance, we consider random box based and blur based data augmentation along with the proposed object based data augmentation technique. Blur based data augmentation refers to convolution of the origi-

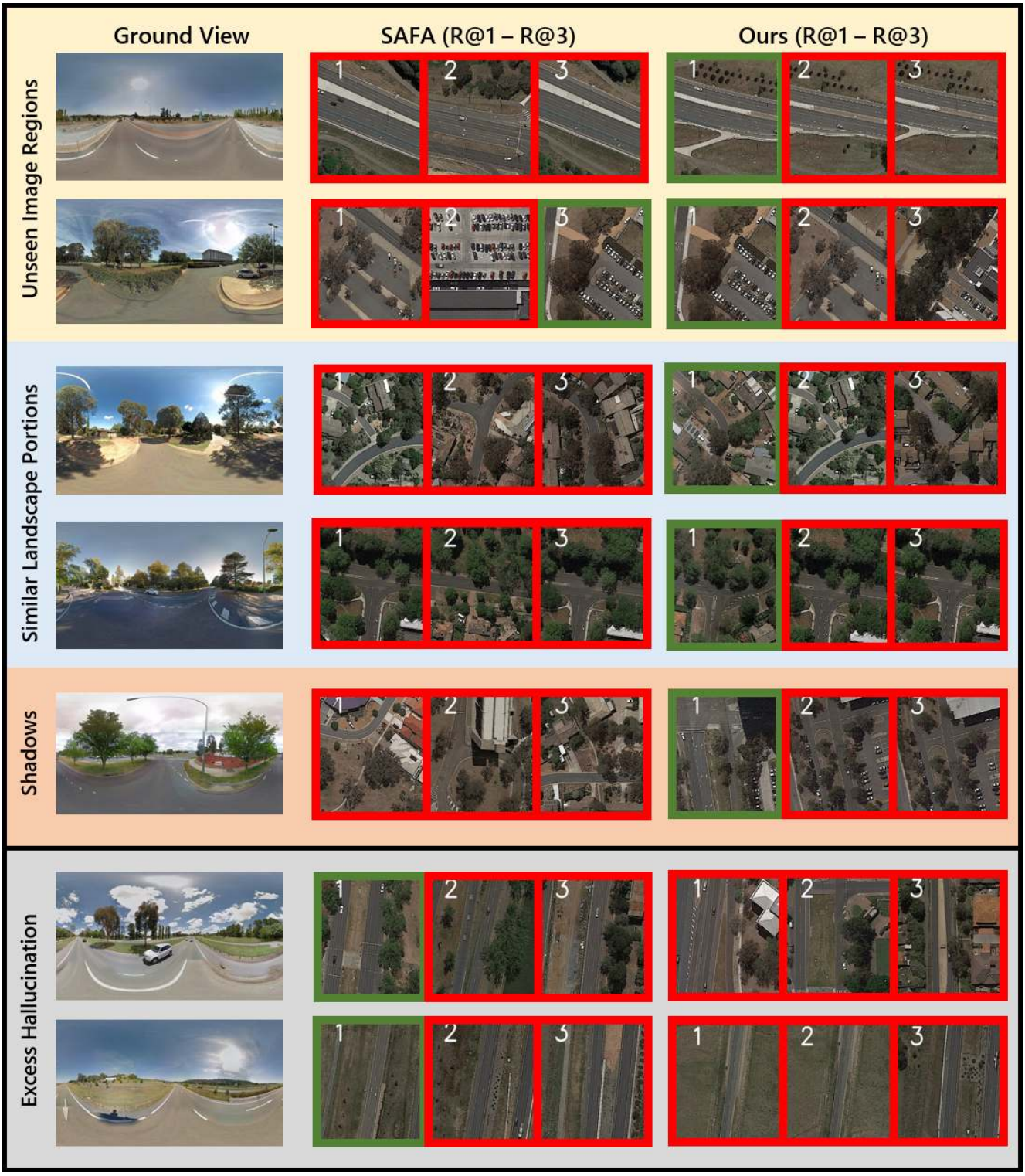

Figure 8: Examples of our results compared to SAFA [18] on the CVACT [12] data-set. Successful matching results are above the center black line in *Yellow, Blue and Red*. Common failure cases are shown below in *Grey*. Our method hallucinates unseen image regions such as occluded landscape patterns (*Yellow* 1^{st} *Row*) and road structures (*Yellow* 2^{nd} *Row*). We can successfully match among portions of similar landscape regions in the aerial image database (*Blue*). Hallucinating unseen objects also helps in matching aerial images with shadows (*Red*). Failure cases include over hallucination. Often all objects are present in both the views and there is no need for hallucination, SAFA [18] performs well on such cases (*Grey*). These examples indicate the complementary nature of the proposed approach compared to SAFA [18].

Table 3: Ablation study on the effect of combining common data augmentation techniques with the proposed approach.

Blur based for both views	Random box for both views	Object based for Ground view	Top-1 Rate CVUSA [26]	Top-1 Rate CVACT_val [12]
-	-	-	67.92	65.89
✓	-	-	69.75	66.03
-	✓	-	70.67	67.22
-	-	✓	74.80	70.94
-	✓	✓	75.51	71.34
✓	✓	✓	**75.95**	**73.19**

Table 4: Ablation study on the effect of using multi-scale attention modules. M denotes the use of multiple such modules.

Approach	CVUSA [26]				CVACT-val [12]			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
No attention	61.24	84.98	90.70	98.85	60.38	84.33	89.46	96.54
Single Scale attention	61.26	85.22	90.84	98.91	61.77	85.37	90.60	96.92
Multi-Scale attention (M = 1)	63.16	86.01	91.18	98.96	63.54	86.32	90.93	96.90
Multi-Scale attention (M = 2)	66.47	87.48	92.45	99.22	64.95	86.83	90.94	97.12
Multi-Scale attention (M = 3)	67.92	88.34	92.79	99.26	65.89	87.59	91.41	96.64
M = 3 with Data Augmentation	**75.95**	**91.90**	**95.00**	**99.42**	**73.19**	**90.39**	**93.38**	**97.45**

nal image with a kernel of fixed average blur. Random box based data augmentation closely resembles to our keep and remove mode of image regions, just that they are boxes of random size and do not contain any semantic information. Note that blur and random box based data augmentation are applied on both views, where as our object based data-augmentation is applied only on ground view. We decided not to segment aerial images as we could not find a suitable off-the-shelf image segmentation module that transferred well on aerial images of CVACT and CVUSA. The images used for training the baseline system (*i.e*. row 1 of table 3) contains no data augmentation. All systems indicated in table 3 are Multi-scaled attention systems (M = 3).

As shown in table 3 incorporating common data augmentation techniques leads to improvement in top-1 recall rates on CVUSA and CVACT benchmarks. Individual and joint contribution of data augmentation techniques are also illustrated in table 3.

Effects of multi-scaled attention modules: We show the effectiveness of our proposed multi-scale attention module in table 4 and compare it with no attention and single-scale attention equivalent models. We also show the advantage of aggregating features from multiple such attention modules. We study the behaviour of top-1, top-5, top-10 and top-1% recall rates on CVUSA and CVACT-val benchmarks. It can be observed from table 4 that incorporating attention at different scales and aggregating features from multiple such modules improves top-1 retrieval rate from 61.24% to 67.92% for CVUSA and 60.38% to 65.89% for CVACT-val data-set. Here the images used for training are not augmented. The last row in table 4 indicates the effectiveness of combining multi-scale attention modules with the proposed data augmentation. Combing the two increases top-1 retrieval rate from 67.92% to 75.95% for CVUSA and 65.89% to 73.19% for CVACT-val data-sets.

5. Conclusion

In this work we presented a method designed to address the challenges in cross-view image geo-localization that arise due to temporal variation in scenes. Successful localization of images captured at different time instants is advantageous as it reduces the burden of storing aerial images across time. Our proposed solution involves a two step approach. The first step uses a semantically driven data augmentation technique to simulate the disappearing and newly appearing object peculiarities in temporally varying scenes. We then present the augmented images to train a Siamese network with a multi-scaled attention backbone to hallucinate image regions that are not available for matching. Integrating our proposed model into existing frameworks boosts recall rates on standard benchmarks demonstrating the effectiveness of our proposed solution.

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] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. Geo-localization of street views with

aerial image databases. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1125–1128, 2011.

[3] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8391–8400, 2019.

[4] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 9–17, 2015.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. 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*, pages 7258–7267, 2018.

[9] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013.

[10] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5007–5015, 2015.

[11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[12] L. Liu and H. Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5617–5626, 2019.

[13] Xiaoqiang Lu, Hao Sun, and Xiangtao Zheng. A feature aggregation convolutional neural network for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):7894–7906, 2019.

[14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[15] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 470–479, 2019.

[16] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 2626–2634, 2020.

[17] Turgay Senlet and Ahmed Elgammal. A framework for global vehicle localization using stereo images and satellite and road maps. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2034–2041. IEEE, 2011.

[18] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems 32*, pages 10090–10100. Curran Associates, Inc., 2019.

[19] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020.

[20] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 11990–11997. AAAI Press, 2020.

[21] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *Proceedings of the European Conference on Computer Vision*, pages 494–509. Springer, 2016.

[22] Tingyu Wang, Zhedong Zheng, Chenggang Yan, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *arXiv preprint arXiv:2008.11646*, 2020.

[23] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[24] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 70–78, 2015.

[25] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[26] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017.

[27] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. *arXiv preprint arXiv:2002.12186*, 2020.

[28] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.