Cross-year multi-modal image retrieval using siamese networks

Margarita Khokhlova, Valérie Gouet-Brunet, Nathalie Abadie, Liming Chen

CROSS-YEAR MULTI-MODAL IMAGE RETRIEVAL USING SIAMESE NETWORKS

Margarita Khokhlova[1,2], *Valérie Gouet-Brunet*[1], *Nathalie Abadie*[1], *Liming Chen*[2]

[1]LASTIG, Univ Gustave Eiffel, ENSG, IGN, F-94160 Saint-Mande, France,
[2]LIRIS/Lyon Centrale, 36 av Guy de Collongue, 69134 Écully, France

ABSTRACT

This paper introduces a multi-modal network that learns to retrieve by content vertical aerial images of French urban and rural territories taken about 15 years apart. This means it should be invariant against a big range of changes as the (natural) landscape evolves over time. It leverages the original images and semantically segmented and labeled regions. The core of the method is a Siamese network that learns to extract features from corresponding image pairs across time. These descriptors are discriminative enough, such that a simple kNN classifier on top, suffices as final geo-matching criteria. The method outperformed SOTA "off-the-shelf" image descriptors GEM and ResNet50 on the new aerial images dataset.

Index Terms— Siamese networks, multi-modal CBIR.

1. INTRODUCTION

Aerial images, such as images from satellites or other aerial imaging devices, are distinctly different from image datasets such as CIFAR [1], Imagenet [2], etc. These images are much more semantically similar in composition as they capture natural and urban landscapes, which are all made up of visually near-identical elements such as vegetation and man-made structures. Lately, a great volume of historical images was digitized, among them many aerial images through national surveys mainly from mapping agencies [3]. They are a unique resource to study landscape evaluation, urbanization, land usage, historical events, and others. Alegoria project [4] aims to create content-based image retrieval (CBIR) tools to help end users accessing such volumes of images. The difficulty is that many photographic materials are scarcely, or not at all annotated, which makes it hard to link them to modern photographic images of the same territory.

State-of-the-art (SOTA) approaches for image retrieval [5, 6] are designed to deal with cross-view and multi-modality challenges but are trained and tested on benchmarks composed of important and distinct man-made architectures [7, 8], which are very different from aerial images, if only by the resolution and composition. It is, therefore, an open research question whether these methods can be used for cross-time aerial image retrieval. An additional question is whether the semantic information from cartographic maps can be beneficial for such a cross-time image retrieval task and if so, how it can be exploited and encoded by a descriptor? This paper aims to match urban and rural vertical aerial images, taken 15 years apart. Obviously, these images differ as the landscape evolves over time, which makes matching non-trivial. The core idea is to learn an embedding that retains all information required to recover the scene appearance through time and under varying acquisition conditions. The dataset contains images in pairs, one from each decade, next to manually labeled segmented semantic masks. An example of the data can be seen in Figure 1. The sought-after method must be able to distinguish between images that are semantically close, as all contain similar elements, yet also must be robust against change, the appearance and disappearance of objects over time and even seasonal effects. Lastly, as we are learning to match image pairs, each image pair is a class by itself. Unlike Imagenet object recognition tasks containing at least 500 images per class, we deal with only 2 training images per correspondence.

Our contributions are the following. Firstly, we evaluate the performance of existing methods for a new cross-time retrieval task using a dataset created for this purpose. We determine the most important data modality and evaluate several scenarios for multi-modal fusion. Secondly, we propose a novel descriptor for multi-modal data and fine-tune it on our dataset. The core of our method is formed by a Siamese network that takes image pairs as an input. The image pairs contain not only natural images but also the semantic labels corresponding to each image, which makes our approach multi-modal. The net outputs a single descriptor per image pair that captures the similarity whilst being robust against all changes occurred over time. This descriptor is low-dimensional but powerful enough such that the final classification whether an image pair is a temporal match may be done by a simple unsupervised k-nearest neighbors (kNN) method.

2. PROBLEM STATEMENT AND BACKGROUND

The dataset originates from French Mapping Agency (IGN) [9] and contains vertical aerial images taken from three

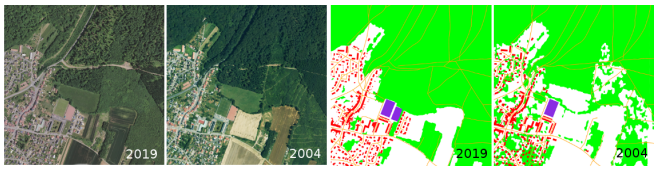

Fig. 1. Image and semantic data evolution 2004-2019. Note also the seasonal changes and lighting condition differences.

data type	color RGB	# 2004	# 2019
road	(255,165,0)	380731	326882
church & chapel	(255,255,0)	1292	2195
fort & blockhaus	(128,128,128)	481	734
other building	(255,0,0)	251294	3475104
water resource	(0,0,255)	28043	12040
sport ground	(138,43,226)	1409	2859
cemetery	(75,0,130)	928	1299
vegetation zone	(0,255,0)	224101	164435
railroad	(255,0,255)	3308	3972

Table 1. Main semantic categories stats. Moselle 2004-2019.

French regions (Moselle, Bas-Rhin, and Meurthe-and-Moselle) in 2004 and 2019. Consequently, we call it FR-0419. The extra modality is formed by per pixel semantic annotations, similar to those found in traditional cartographic maps. The number of image pairs for the regions is 6000, 4430 and 5855.

We selected several categories of semantic objects (See Table 2). Note that the number of annotated objects may differ significantly, partly due to the different annotation strategies and partly due to landscape evolution. The aerial images are 50cm/pixel and used in patches of a square kilometer. We use perfectly matching image coordinates between the years - *i.e.* the image regions are aligned. This scenario is not realistic but allows us to test and demonstrate the robustness of descriptors against changes and evolutions in the landscape.

Problem statement. This paper ascertains to what extent existing descriptors may be used to match aerial images through time, and which information (visual, semantic, or both) is more relevant for this task. We test different fusion strategies to encapsulate all modalities into a single multi-dimensional descriptor. Images from 2019 are used queries against their 2004 counterparts, and different geographical regions are selected to form a training, validation, and testing sets. Lastly, we introduce our new Siamese net-based descriptor.

Background. Recently with the progress of segmentation Convolutional Neural Network (CNN) architectures the task of fully automated scene segmentation became possible [10]. Semantic maps are an incredible additional source of information that can potentially improve geolocalisation, cross-view and cross-time retrieval. Combining different information sources and modalities to improve (CNN) models was explored in [11, 12, 13, 14]. However, to the best of our knowledge, using multi-modal data to retrieve aerial images taken in different years, is novel. A related research problem is visual localization, where acquisitions differ in viewing conditions and suffer from a wide range of distortions [5, 6, 15] or extreme view-point changes [16, 17]. Traditionally the goal is to represent image features as robust feature vectors, contemporary work focuses on learning based methods [14, 5, 6]. However, current methods are designed to handle object-specific features and are not tailored to retrieve images over time, where the scenes might not contain a single outstanding key object being composed of repetitive man-made structures instead. Hence these methods cannot be applied straightforwardly in the case of large landscape changes over time along with non-characteristic (distinct) image features.

Aerial image descriptors can greatly benefit from the data of other modalities. This was successfully demonstrated by Audebert et al. [13], where better segmentation maps were obtained using an encoder-decoder architecture and images along with semantic data originating from OpenStreetMap [18]. Up to our knowledge, this is the only work that directly uses semantic labels along with images in an aerial image context. Li et al. [19] proposed a multi-modal late feature fusion-based framework to improve the geographic image annotation. However, they source from a single image, where the different modalities are simply features extracted by different algorithms. Chen et al. [20] propose a CBIR method benefiting from multi-modal (spatial and spectral) information content of Remote Sensing images, however, they use hand-crafted descriptors and focus on the retrieval task.

Siamese network architectures aim to construct an embedding, where two extracted features corresponding to the same identity are likely to be closer, than features from different identities [21]. They are a popular choice for problems dealing with so-called one-shot learning problems, when a single training sample is available for each class. The efficiency of Siamese networks was previously demonstrated for visual object tracking [22], person reidentification [23], cross-view image matching [24] and other tasks.

We propose to use a custom Siamese architecture to obtain an embedding, which encodes the visual features and learn to ignore any temporal induced changes. It allows us to train descriptors on our single-pair correspondence dataset, whilst the backbone architecture is simultaneously designed to benefit from multi-modal information.

3. PERFORMANCE BASELINE

Up to our knowledge, there are no publicly available pre-trained image descriptors, that are fine-tuned on aerial images, nor are there any quantitative studies comparing image presentations specifically for multi-modal aerial images. We, therefore, establish our own baseline performance benchmark using existing image descriptors.

The baseline is formed by Resnet [25] and GEM [5], both global in their nature. They are pre-trained on Imagenet [26] and Oxford5k [26] respectively. We use the output of the last convolutional layer followed by max-pooling to obtain a descriptor for Resnet and the pre-trained framework as provided by the authors for GEM. The first step is to comparatively evaluate methods, their combinations, and parameters for accurate cross-year image matching. The baseline study using the algorithm(s) is schematically depicted in Figure 2.

Cross-time image retrieval setup. As a benchmark, we evaluate the retrieval accuracy of the "off-the-shelf" descriptors for different data modalities. We resize the input images to the size of 512x512x3 for Resnet50 and 1024x1024x3 for GEM [5]. We evaluate three scenarios:

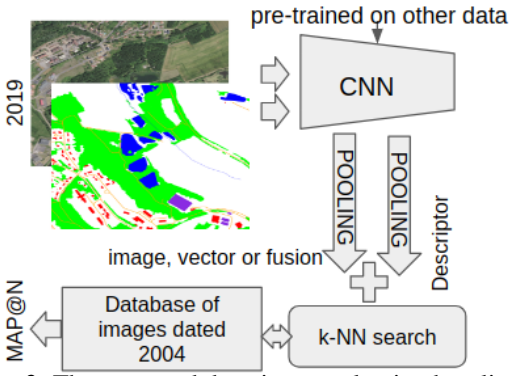

Fig. 2. The proposed descriptor evaluation baseline.

- concatenation of multi-modal descriptors;
- prior to CNN image fusion via convolution;
- late fusion.

The weights of the convolutional fusion layer are pre-defined.

We use the Mean Average Precision (map@N) to evaluate the results: $map@N = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{r}{n}}{M}$, where N is 5, r is equal to 1 if the retrieved image is correct and 0 otherwise and n is the order of retrieved images, M the total number of images. Tables 2 and 3 summarise the averaged map@N results. We experimented the retrieval with different distances (Euclidean, Cosine), RGB and grayscale semantic masks, different types of re-scaling techniques (L_1, L_2, standardization), and concatenation or sum of the obtained descriptors. We observed that different parameter combinations affect the final map@5 accuracy significantly yielding up to 5% variations. However, we did not establish a single common trend for the three regions tested apart of cosine distance constantly outperforming the euclidean in the kNN matching stage. We, therefore, report the best map@5 score and a parameter set.

Baseline results. Firstly, descriptors based on semantic information give better results in terms of map@5 value than descriptors based on natural images. The best map@5 scores were obtained using both modalities which confirms that the additional information is beneficial for the cross-time image retrieval task. Moreover, the combination of visual and semantic data at the late stage allows improving the results even further for the descriptors tested. Overall, the obtained results are not good, showing the limitations of existing CNN-based approaches when dedicated training datasets are not available.

4. SIAMESE ARCHITECTURE

Our multi-modal Siamese network architecture is schematically illustrated in Figure 3. The architecture has two copies of the function G_W, which share the same set of parameters W, consists of two branches and and a distance module. A loss module is placed on top of the architecture.

The network architecture is designed to handle the multi-modal input and fine-tune the descriptors for the image retrieval task. The input to the CNN is a pair of multi-modal images $(X_1, L_1; X_2, L_2)$ and a label Y. One branch processes corresponding image pairs originating from the same geographical zone through time (2019-2004), the other processes a non-corresponding pair from the same year (2019). The images are passed through, yielding two outputs $G(X_1, L_1)$ and $G(X_2, L_2)$. The cost module then generates the distance $D_W(G_W(X_1, L_1), G_W(X_2, L_2))$. The loss function combines the probability p predicted by the classification layer with the sigmoid activation based on D_W with label Y to produce the scalar loss value:

$$\mathcal{L}(W) = -\frac{1}{M}\sum_{m=1}^{M} L(W, (Y, X_1, X_2, S_1, S_2)^m) \quad (1)$$

$$D_W(X_1, S_1, X_2, S_2) = |G_W(X_1, S_1) - G_W(X_2, S_2)| \quad (2)$$

$$\mathcal{L}(W, (Y, X_1, X_2, S_1, S_2)^m) = Y log(p) + (1-Y) log(1-p) \quad (3)$$

where m is the number of pairs. The first layer of the network is a convolutional layer with pre-defined trainable weights which serves to warm-up the training. The backbone of the network is Resnet50 pre-trained on ImageNet. The output of the last convolutional layer of Resnet is passed through 3 convolutional layers C_1-C_3 followed by a fully connected layer. We found that using the $tanh$ activation and batch normalization in all the added convolutional layers gives the best result.

Mining so-called hard image pairs is essential to make proper training of Siamese nets possible [27]. We adopted the following learning strategy: every 5 epochs, the map@5 score is calculated for the training dataset. Hard image pairs are the ones that have wrong retrieved images (*i.e* retrieved images do not correspond to the same geographical zone).
The code and weights of the trained model are available[1]

5. EXPERIMENTS

In this section, we describe the experimental setup used to compare our method to the baseline results obtained with "off-the-shelf" image descriptors on FR-0419 benchmark dataset. We fine-tune the proposed architecture in an end-to-end fashion using the images from the Moselle region for training, whereas Bas-Rhin is a validating set and Meurthe and Moselle form the testing set. Throughout all experiments, the architecture from Figure 3 is used and the input size of both aerial and semantic images is 256x256. The fusion by convolution is deployed to combine the multi-modal data and allow end-to-end training. The first convolutional layer is pre-initialized. The 3 convolutional layers C_1-C_3 on top of the Resnet have 3×3 kernels and the number of filters equals 1024, 512, and 256. The final descriptor dimension is D^{128}, next to testing the values of 256, and 512. See Table 5, R=128 dimensions generalizes the best.

The network learns to predict whether two multi-modal descriptors correspond to the same geographical zone through time, based on the L_1 distance between them. The idea is similar to contrastive loss [28] commonly employed in Siamese networks, but we found this approach to work better.

We use BCE loss, next to Adam optimizer with a fixed lr 8e-4 and decay 8e-7. We re-determine hard samples after every 5 epochs based on the training set map@5 scores. The kNN algorithm with cosine distance is used to retrieve

[1] https://github.com/margokhokhlova/siamese_net

data	norm	fusion	distance	R	mean average precision @5 Moselle	Bas-Rhin	Meurthe & Moselle	average
visual image	n/a	none	cosine	2048	0.48	0.70	0.65	0.60
semantic image	n/a	none	cosine	2048	0.67	0.57	0.72	0.65
vis + semantic	mean std	D concatenation	cosine	4096	0.66	0.69	0.71	0.69
vis + RGB sem	none	image conv	cosine	2048	0.62	0.64	0.64	0.63
vis + RGB sem	none	late fusion	cosine	2048	**0.76**	**0.75**	**0.84**	0.79

Table 2. Off-the-shelf descriptor map@5 for pre-trained Resnet50 [25]. Best scores per per department are shown in bold.

data	norm	fusion	distance	R	mean average precision @5 Moselle	Bas-Rhin	Meurthe & Moselle	average
visual image	mean std	none	cosine	2048	0.54	0.63	0.59	0.60
semantic image	mean std	none	euclidean	2048	0.63	0.64	0.61	0.63
vis + semantic	none	D concatenation	cosine	4096	0.66	0.69	0.71	0.68
vis + RGB sem	none	image fusion	cosine	2048	0.51	0.52	0.56	0.53
vis + RGB sem	none	late fusion	cosine	2048	0.75	0.73	0.84	0.77

Table 3. Off-the-shelf descriptor map@5 for pre-trained Gem [5], architecture resnet101-gem-reg-whiten.

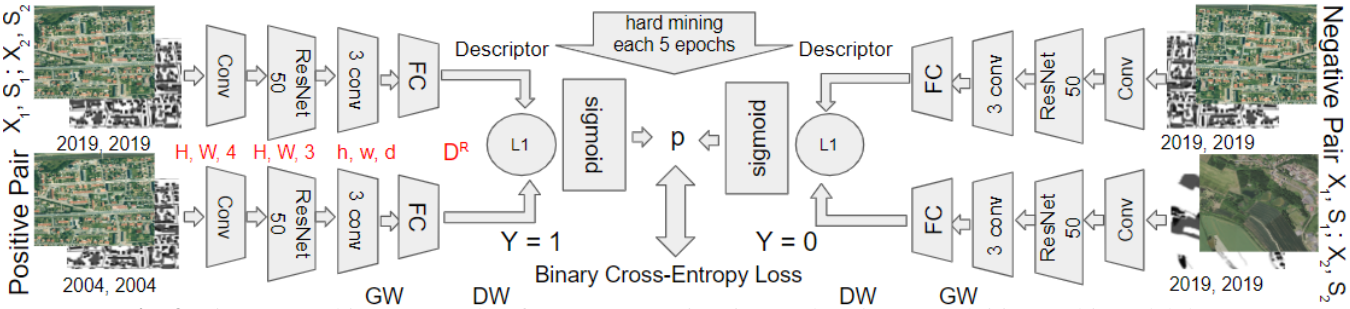

Fig. 3. Siamese Architecture used to fine tune cross-time image descriptors exploiting multi-modal data.

the most similar images given a query. During training, each batch is composed of 12 pairs of positive and negative images, half of which are randomly selected and half are the hard images. In each epoch we go through all training set images, each time selecting random negatives and adding hard-mining samples into a batch. Data augmentation consisted solely of vertical and horizontal image flips. The final model was trained 120 epochs. The map@5 score on the validation set determines the best descriptor parameters. We also cross-validated by swapping the regions for training/validation and test and re-training the net from zero. Once the net is tuned, we use one branch to calculate a multi-modal input descriptor.

Table 5 summarises the final map@5 scores obtained. They demonstrate that the proposed Siamese architecture successfully improves the baseline results and is capable to deal with temporally misaligned images. We attain the map of 0,90 for our validation and training datasets which is a 10% improvement over the best baseline results. Moreover, the resulting descriptor is >10 times more compact than 'off-the-shelf' counterparts having just 128 dimensions instead on 2048 (or even 4096 if concatenated), which allows it to better scale for large databases and reduce the retrieval time.

6. CONCLUSION

In this study we tackle cross-time aerial image retrieval. We introduced a novel approach for learning from multi-modal data to fine-tune any CNN-based image descriptor. We performed a comprehensive comparison of different strategies to use multi-modal information and proposed a custom Siamese network architecture. The resulting descriptor is powerful enough to distinguish between images that are semantically close and is robust against evolutionary landscape changes through time. Experiments show that our method improves the baseline and outperforms SOTA image descriptors. We demonstrated how to use both image and semantic modalities in a single descriptor. In addition, the method is generalizable to any CNN-based feature extractor.

R	map@5 training Moselle	map@5 validation Bas-Rhin	testing M Moselle
128	0.92	0.88	0.94
256	0.93	0.91	0.90
512	0.86	0.70	0.62

Table 4. Map@5 precision obtained with different R in D^R.

data	baseline map@5	set	tuned map@5
Moselle	0.76	training	**0.87**
Bas-Rhin	0.75	validation	**0.88**
Meurthe-and-Moselle	0.84	testing	**0.94**

Table 5. Baseline best vs fine-tuned model performance.

7. ACKNOWLEDGEMENTS

This work is supported by ANR, the French National Research Agency, within the ALEGORIA project, under Grant ANR-17-CE38-0014-01.

8. REFERENCES

[1] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," Tech. Rep., Citeseer, 2009.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[3] "Catalogue collectif de france. fonds lapie de photographies aériennes," `https://ccfr.bnf.fr/portailccfr/ark:/06871/0033535`.

[4] "Alegoria: Advanced linking and exploitation of digitized ge0graphic iconographic heritage," `http://www.alegoria-project.fr`.

[5] Filip Radenović, Giorgos Tolias, and Ondřej Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[6] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3456–3465.

[7] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5706–5715.

[8] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *IEEE conference on computer vision and pattern recognition*, 2008, pp. 1–8.

[9] "Le portail IGN," `https://geoservices.ign.fr/`.

[10] Pascal Kaiser, Jan Dirk Wegner, Aurélien Lucchi, Martin Jaggi, Thomas Hofmann, and Konrad Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6054–6068, 2017.

[11] Krishna Regmi and Ali Borji, "Cross-view image synthesis using conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3501–10.

[12] Kaiqiang Chen, Kun Fu, Xin Gao, Menglong Yan, Wenkai Zhang, Yue Zhang, and Xian Sun, "Effective fusion of multi-modal data with group convolutions for semantic segmentation of aerial imagery," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019, pp. 3911–14.

[13] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre, "Joint learning from earth observation and openstreetmap data to get faster better semantic maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 67–75.

[14] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler, "Semantic visual localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6896–6906.

[15] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li, "Optimal feature transport for cross-view image geo-localization," *arXiv preprint arXiv:1907.05021*, 2019.

[16] Liu Liu and Hongdong Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5624–5633.

[17] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers, "Image-based localization using lstms for structured feature correlation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 627–637.

[18] "OpenStreetMap," `https://www.openstreetmap.org`, 2019.

[19] Ke Li, Changqing Zou, Shuhui Bu, Yun Liang, Jian Zhang, and Minglun Gong, "Multi-modal feature fusion for geographic image annotation," *Pattern Recognition*, pp. 1–14, 2018.

[20] Osman Emre Dai, Begüm Demir, Bülent Sankur, and Lorenzo Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 7, pp. 2473–2490, 2018.

[21] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, 1994, pp. 737–744.

[22] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8971–8980.

[23] Dahjung Chung, Khalid Tahboub, and Edward J Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1983–1991.

[24] 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*, 2018, pp. 7258–7267.

[25] 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*, 2016, pp. 770–778.

[26] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[27] Ben Harwood, BG Kumar, Gustavo Carneiro, Ian Reid, Tom Drummond, et al., "Smart mining for deep metric learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2821–2829.

[28] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006, vol. 2, pp. 1735–1742.