

# Convolutional Neural Network-Based Deep Urban Signatures with Application to Drone Localization

Karim Amer Mohamed Samy Reda ElHakim Mahmoud Shaker Mohamed ElHelw Center for Informatics Science

Nile University

{k.amer, m.kasem, r.mostafa, m.serag, melhelw}@nu.edu.eg

# Abstract

Most commercial Small Unmanned Aerial Vehicles (SUAVs) rely solely on Global Navigation Satellite Systems (GNSSs) - such as GPS and GLONASS - to perform localization tasks during the execution of autonomous navigation activities. Despite being fast and accurate, satellite-based navigation systems have typical vulnerabilities and pitfalls in urban settings that may prevent successful drone localization. This paper presents the novel concept of "Deep Urban Signatures" where a deep convolutional neural network is used to compute a unique characterization for each urban area or district based on the visual appearance of its architecture and landscape style. Such information is used to identify the district and subsequently perform localization. The paper presents the methodology to compute the signatures and discusses the experiments carried out using Google maps and Bing maps, where the latter is used to simulate footage captured by SUAVs at different altitudes and/or using different camera zoom levels. The results obtained demonstrate that Deep Urban Signatures can be used to successfully accomplish district-level aerial drone localization with future work comprising accurate localization within each identified district.

# 1. Introduction

When viewed from above, different urban areas have dissimilar appearances due to varying building shapes and construction patterns, road network topology, vegetation and space allocation/utilization density. As humans, it is easy to identify the picture in the lower part of Fig. 1 as being taken for Heliopolis district in Cairo due to its starshaped mesh of straight roads, the proximity of buildings, and the high degree of development. On the top of Fig. 1, the picture is taken form the upscale Sheikh Zayed City on the outskirts of Cairo where in contrast to Heliopolis, Sheikh Zayed is characterized by gated residential compounds, curved roads, and still to be developed areas. Cairo also suffers from many slum districts that grew sumptuously without proper urban planning. When observing aerial imagery of different slum districts shown in Fig. 2, it is evident that each slum has its own distinctive visual appearance due to used building materials, economic factors and demand levels. The unique circumstances of each location results in a unique visual fingerprint that differentiates each district and which we term "Deep Urban Signatures". In case of drone losing GPS navigation signals, these signatures can be computed on-board to autonomously guide the drone to regions with better GPS coverage and/or away from dense residential areas.



Figure 1: Shows the difference between images from Zayed City (top) and Heliopolis (bottom).

## **1.1. GNSS-Denied Environments**

A Global Navigation Satellite System (GNSS) is a Medium Earth Orbit (MEO) satellite constellation that provides autonomous geo-spatial positioning. GNSS allows receiver devices to determine their location (longitude, latitude, and altitude) using timed radio signals transmitted along a line of sight from satellites. Common GNSS Systems include GPS, GLONASS, Galileo, Beidou and other regional systems. The term GPS, which is the United States' owned Global Positioning System, is commonly used interchangeably with GNSS in UAV navigation literature. The GPS satellites orbit at 20,180 km above sea level and transmits L1 signals for civilian usage at 1575.42 MHz [1]. By the time it reaches earth, and due to the large distance traveled, the L1 signal power is limited -160 dBW when measured by receiver units, which is below the ambient background noise of many places on earth especially in urban locations in and around certain areas of large cities. In addition to background noise, L1 signals are also vulnerable to radio interference and loss of Line of Sight (LOS).

# 2. Deep Urban Signatures

In this work, we use a satellite image as reference map for localization instead of GPS. The typical scenario involves a UAV capable of capturing top-down images to be compared to a pre-stored onboard reference map. This scenario has two challenges: 1) the search space in the reference map is massive even if we already know the city above which the drone is flying, and 2) there is significant difference in appearance between an image captured by a drone and the corresponding image patch in the satellitebased reference map.

This paper aims to reduce the search space in the reference map by identifying the district above which the drone is flying by using the concept of 'Deep Urban Signatures'. After district identification, a second phase is carried out, called Neighborhood Retrieval, to achieve drone localization within recognized district. Typically, the drone image covers a neighborhood or a subarea of the district and the Neighborhood Retrieval phase computes the closest neighborhood to the location of the drone. The paper explains how Deep Urban Signatures are computed using a Convolutional Neural Network (CNN) [2]. In fact CNNs have been extensively used in tasks such as object recognition [3], object localization [4], and semantic segmentation [5], and it will be shown how they extract a range of descriptive features from aerial images sufficient to discriminate different districts and achieve enhanced localization.

#### 3. Related Work

The idea of geo-localization in GPS denied environments using aerial images has been studied in [6]. The objective was to estimate the relative position for autonomous navigation and the absolute position for increasing the reliability of the system. A query image is captured from the on-board camera to be matched with reference images using Hausdorff distance [7]. Then, the absolute location is inferred to reduce the error accumulated during the calculation of the relative position. In [8], a query image is used for matching with a georeferenced image to calculate the absolute position based on the normalized cross-correlation of intensity values of both the reference and the query images. A more recent approach for geo-localization is developed in [9] where mutual information [10] is used for image registration.

The above research uses template based techniques for image registration; the query image is passed over every possible position in the geo-referenced images, where the latter are typically stitched to form a map. In this work, we use a CNN to extract unique identifiers for different districts that are used during testing to classify the district to which the query image belongs. This approach has the advantage of being more robust and scalable compared to template based systems. While the latter scales linearly during testing and suffers from a sharp decrease in reliability, in the CNN based approach the test time remains constant as the size of the database increases and adapts better to variable query image size/resolution.

Additional related work can be found in the literature that uses other input data than aerial images and its localization objective is within a specified region. In image geo-localization [11]-[13], a vision-based technique is used to estimate the geographic location given a street level image not an aerial image. It is a more challenging problem due to the difference in viewpoints between training and testing images. Similar to [13], our work takes a classification approach to geo-localization using CNNs. However, [13] uses street level images which, unlike aerial maps now available for most parts of the world and being updated regularly, are not accessible for many locations especially in developing countries. In Simultaneous Localization and Mapping (SLAM) [14] [15], UAV localization is achieved within the mapped environment and [16] [17] describe how SLAM is used for autonomous navigation in GPS-denied environments. However, SLAM algorithms typically use other sensors than the camera for 3D mapping and the objective of localization is restricted to the local environment.

Recently, CNNs gained considerable popularity in image classification by winning the ImageNet competition [18] by AlexNet [3] in 2012. AlexNet architecture consisted of multiple consecutive layers of convolutional

and max pooling layers followed by fully connected layers. Many CNNs have emerged since then with different architectures and deeper models such as VGGNet [19] and ResNet [20]. Feature extraction with CNNs is commonly used in visual analysis tasks where it has proved to outperform previously-used global descriptors [21]. Another important practice in CNNs is using models pretrained on different dataset for different tasks other than classification. Pretrained VGGNet has been used in image segmentation [4] [17] and multiple object detection [4] whereas [23] uses AlexNet for object localization. In [22], a pretrained VGGNet is applied for segmentation of satellite images. In our work, pretrained models are used for classification, and in conjunction with neighborhood retrieval, provide an integrated framework for aerial drone localization.

## 4. Dataset

We generated a dataset of satellite images from Google maps [24] and Bing maps [25]. Google maps are considered our reference maps, and due to lack of real drone imagery, Bing maps are used to generate a simulated drone footage. The dataset contains images representing seven different districts in Cairo, Egypt with resolution of 1.19 meter/pixel. The districts and their characteristics are as follows:

- 1. Sheikh Zayed City: a modern district located in the west of Cairo and distinguished with curved roads, gated compounds, under-devolvement, and green areas.
- 2. Nasr City: a developed district located east of Cairo and is famous for its Manhattan road planning.
- 3. Heliopolis: a developed district located east of Cairo and planned as a collection of roundabouts that encompass triangular blocks.
- 4. Fatimid Cairo: a historical old district in the middle of Cairo with the most distinctive features being large ancient mosques scattered all over the area.
- Manshit Naser: a slum area east of Cairo. It is a very crowded area with poor or no planned infrastructure and extremely narrow clearances between buildings.
- 6. Al-Khsus: another slum area east of Cairo. While being very crowded like Manshit Naser, this slum expansion formed a pattern of long narrow alleys.
- The 5<sup>th</sup> Settlement: a modern upscale underdeveloped district south of Cairo. In this district, buildings are organized in back to back rows with many small green areas around.

For each district, images from both Google and Bing maps are partitioned into fixed-size tiles. The tiles were downloaded and stitched back into larger image which we will call District Map. Table 1 shows different image dimensions of each District Map in pixels and in meters.

District Name	Dimensions in pixels	Dimensions in meters
Sheikh Zayed City	4413 x 7917	5251 x 9421
Nasr City	3328 x 4096	3960 x 4874
Heliopolis	3328 x 2304	3960 x 2741
Al Fatimid Cairo	1024 x 1280	1945 x 1523
Manshit Naser	1792 x 1792	2132 x 2132
Al-Khsus	2819 x 2048	3354 x 2437
The 5 <sup>th</sup> Settlement	4096 x 4352	4874 x 5178

Table 1: Dataset District Maps with dimensions in pixels (middle) and meters (right).

Fig. 2 provides sample images from the District Maps dataset. As seen in figure, there is a difference in appearance between Google and Bing maps. Also, both sources didn't record their images at the same time which resulted in different degree of development in corresponding areas manifested in streets and buildings. These differences will help test model generalization when using Deep Urban Signatures.

## 5. Experiments

Enhanced drone localization is achieved by firstly carrying out district-level localization, *i.e.* recognizing the district over which the drone is flying, and secondly neighborhood-level localization, *i.e.* finding drone location within the recognized district using Neighborhood Retrieval. This Section presents experiments related to these two levels of localization and discusses obtained results.

#### 5.1. District-Level Localization

District classification is the first level of localization and the output of this phase is the district above which the drone is flying. It thus narrows down the drone location to one of the District Maps in the dataset. An important factor in district classification is the area needed to be able to reliably capture Deep Urban Signatures. Therefore, the each of the District Maps was partitioned using different window sizes to investigate the effect of this factor on model performance. Table 2 shows the covered area in meters by each window.

Window Dimensions (pixels)	Window Dimensions (meter)
200 x 200	238 x 238
300 x 300	357 x 357
400 x 400	476 x 476
500 x 500	595 x 595
600 x 600	714 x 714

Table 2: Window dimensions used in district classification in pixels and meters.

Commercial drones such as Phantom DJI have cameras with 94-degree field of view [26]. The corresponding height of the drone to capture the biggest window size (714x714) is approximately 333 meters. This height should be easily reached with most drones but some countries set a limit for flying height to be less than 333 meters.

#### 5.1.1 Data Configuration

District maps are partitioned using different window sizes to test the area of the region to be covered by the drone image in order to capture a robust Deep Urban Signature. Since the districts have different sizes, applying non-overlapping windows will generate different number of patches for each district map. To balance the number of patches from different districts, overlapping windows were used where the degree of overlap is increased for smaller districts. The same partitioning is applied on district maps



Figure 2: Images of seven districts in the dataset. Districts are (from left to right): Sheikh Zayed City, Nasr City, Heliopolis, Fatimid Cairo, Manshiet Nasr, AlKhsos and the 5<sup>th</sup> Settlement.

from Google and Bing, and a separate model is trained for each window size. Google district maps are used for training and validation whereas Bing district maps are used for testing. To expand the dataset size, data augmentation is applied on training patches by rotating them with random angle in the range of [-10, 10] degrees and by horizontal and vertical flipping.

#### 5.1.2 District Localization Model

A deep convolutional neural network is used for district level localization that consists of pretrained convolutional layers of VGG16 [19] followed by two fully connected layers. VGG16 layers are frozen whereas the fully connected layers are trained from scratch using Adam optimizer [27]. The objective of our model is to classify the district of an image patch. For the different window sizes, the same model architecture was used. VGG16 achieved high accuracy on the ImageNet challenge. It is shallower than other deep models such as Inception [28] and ResNet [20] which achieved higher accuracies but have more time and memory requirements. Global average pooling is used after the last convolution layer as in [29] for enhanced time and memory efficiency. Global average pooling is subsequently followed by a fully connected

layer then a Softmax layer for classification.

#### 5.1.3 District Localization Results

Comparing different window sizes quantitatively isn't straightforward as each window size generates different image patches. But we can consider the overall accuracy on each window size as the localization success rate that will face a flying drone. So, the window size that has higher accuracy should be better to use. Fig. 3 shows that accuracy gets higher by increasing the window size from 100x100 to 500x500 as the Deep Urban Signature appears better. However, accuracy starts to decrease after 500x500. Our explanation for this behavior is that bigger window sizes generate smaller datasets, which in effect makes the training more difficult.



Figure 3: Plot shows the model accuracy percentage versus different window sizes in pixels.



Figure 4: Google maps versus Bing maps. Left column shows images from Bing maps and second column are the corresponding images on Google. There is a difference in appearance between the two sources. In addition, some buildings in Google maps do not exist in Bing maps.

The Confusion matrix of 500x500 window model is presented in Fig. 5 and provides the conflict between

classes. One conflict area is between Sheikh Zayed, the 5<sup>th</sup> Settlement and Nasr City which is sensible as these three districts are quite new and have green areas and swimming

pools between the building blocks. Another reasonable conflict is between the 5<sup>th</sup> Settlement and Heliopolis districts.

The results show that models pretrained on human-view images can be used successfully in satellite images which has been previously studied in [22]. However, we still have some conflict areas between different districts due to using a pretrained network on street view images from ImageNet [30]. Results will be improved when we have sufficient data to train a neural network (or tune some layers in a pretrained one). Another way to improve the results should be using a pretrained neural network on areal images.



Figure 5: Confusion matrix of district classification model on 500x500 Window.

#### 5.2. Neighborhood-Level Localization

This is the second level of localization after district recognition. The output is the closest neighborhood to the drone's location within the recognized district.

# 5.2.1 Data Configuration

Google map of each district is divided into patches of size 500x500 pixel with stride 50. For testing, the Bing map is also divided using the same approach but the stride is perturbed with a uniform random integer in the range of [-50, 50] in order to randomly simulate random location.

#### 5.2.2 Neighborhood Localization Model

Each patch from Google map is fed to the VGGNet to extract the feature map of the 14<sup>th</sup> and 18<sup>th</sup> convolutional layers to be used as a database for retrieval. Each patch from Bing map runs through a nearest neighbor test to get the closest patch in Google map. 14<sup>th</sup> and 18<sup>th</sup> convolutional layer outputs are tested separately as the model is pretrained on ImageNet data [30] that is different

from the satellite data. The early convolutional feature map corresponds to mid-level features and the deep convolutional feature map corresponds to high-level features as mentioned in [31]. The number of features in the  $14^{\text{th}}$  convolutional layer is 14\*14\*512 = 100,352 features which is four times the number of features in the  $18^{\text{th}}$  convolutional layer with 7\*7\*512 = 25,088 features.

Error is calculated as the Euclidean distance in meters between the coordinates of the top left pixel of the true patch and the predicted patch. The queried image captured by the drone is assumed to cover an area of 354,025 (595\*595) squared meters at least. Due to memory limitations, we were able to apply this experiment on all districts except for Sheikh Zayed district. The second largest district, The 5<sup>th</sup> Settlement, requires 4.2 GB of RAM for the nearest neighbor retrieval. The time per image test sample is around 1 second. Other districts are smaller than The 5<sup>th</sup> Settlement and have less time and memory requirements.

#### 5.2.3 Neighborhood Localization Results

Table 3 shows the error obtained using the 18<sup>th</sup> convolutional layer over 6 districts. Table 4 shows the accuracy obtained using the 14<sup>th</sup> convolutional layer over 6 districts. Sheikh Zayed City patches could not be generated because the patches' size was bigger than that of the available CPU memory.

District	Error in meters
Nasr City	533.41
Heliopolis	394.93
Fatimid Cairo	46.67
Manshiet Naser	628.98
Al Khsus	144.92
The 5 <sup>th</sup> Settlement	582.71

Table 3: Results of using the 18<sup>th</sup> convolutional layer in the VGGNet for neighborhood retrieval.

District	Error in meters
Nasr City	529.70
Heliopolis	111.50
Fatimid Cairo	18.87
Manshiet Naser	123.18
Al Khsus	34.06
The 5 <sup>th</sup> Settlement	387.19

Table 4: Results of using the 14<sup>th</sup> convolutional layer in the VGGNet for neighborhood retrieval.

Surprisingly, the mean error over the six districts is 200.75 meters for the 14<sup>th</sup> convolutional layer and 388.60 meters for the 18<sup>th</sup> convolutional layer, which could be

attributed to the difference in the dataset distribution between ImageNet and the satellite dataset as ImageNet contains different objects with different features than the aerial images. For example, ImageNet has a large number of dogs and cats with high-level features such as a circular face and a body, unlike aerial images that consists of roads and blocks of buildings. It can thus be concluded that midlevel features achieve higher accuracy than high-level features from pretrained models. However, they suffer from an increase in memory requirements, which is four fold in the current case.

### 6. Conclusions

In this paper, we presented the concept of Deep Urban Signatures and discussed how they can be used to locate the district above which the drone is flying. Knowing the district narrows the search space when using satellite maps as a reference for localization. A second phase of localization is applied to retrieve a neighborhood within recognized district and achieve enhanced localization accuracy. Our best model which makes use of pretrained VGG16-Net can discriminate between 7 different districts with average accuracy of 91.2%. For neighborhood retrieval an overall localization error of 200.75 meters on 6 districts has been achieved.

#### 7. Future Work

In this work, VGGNet was used to generate initial promising results. Fine tuning upper layers in VGGNet is expected to produce better results as this network is fully trained on traditional images not areal images. However, such tuning will require collecting larger dataset in order to enable the network to generalize well. Other future work includes testing different deep architectures such as Inception and ResNet. One concern to raise about our pipeline is how the model will classify the areas located on the boundaries between different districts. To address this issue, a voting mechanism should be devised to resolve ambiguity between subsequent batches captured by the flying drone.

#### References

- T. Layh and D. Gebre-egziabher, "GPS-Denied Navigator for Small UAVs Final Report Authors:," no. 2008, 2014.
- [2] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.
- [3] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," pp. 1–9, 2012.
- [4] R. Girshick, "Fast R-CNN," Proc. IEEE Int. Conf.

Comput. Vis., vol. 11-18-Dece, pp. 1440-1448, 2016.

- [5] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.
- [6] D. Sim et al., "Using Aerial Image Sequences," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 1, pp. 1– 18, 2002.
- [7] D. G. Sim and R. H. Park, "Two-dimensional object alignment based on the robust oriented Hausdorff similarity measure," *IEEE Transactions on Image Processing*. 2001.
- [8] G. Conte and P. Doherty, "Vision-based unmanned aerial vehicle navigation using geo-referenced information," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 1–19, 2009.
- [9] A. Yol, B. Delabarre, A. Dame, J. É. Dartois, and E. Marchand, "Vision-based absolute localization for unmanned aerial vehicles," *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 3429–3434, 2014.
- [10] A. Alahi, P. Vandergheynst, M. Bierlaire, and M. Kunt, "Cascade of descriptors to detect and track objects across any network of cameras," *Comput. Vis. Image Underst.*, 2010.
- [11] J. Choi and G. Friedland, "Multimodal location estimation of videos and images," *Multimodal Locat. Estim. Videos Images*, pp. 1–191, 2015.
- [12] C. Paper, "Learning Deep Representations for Groundto- Aerial Geolocalization," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, no. JUNE, p. 270, 2015.
- [13] T. Weyand, I. Kostrikov, and J. Philbin, "Planet photo geolocation with convolutional neural networks," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, vol. 9912 LNCS, pp. 37–55.
- [14] H. Durrant-Whyte and T. Bailey, "Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms," *IEEE Robot. Autom. Mag.*, vol. 2, p. 2006, 2006.
- [15] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-whyte, and M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem," *IEEE Trans. Robot. Autom.*, vol. 17, pp. 229–241, 2001.
- [16] J. Kim and S. Sukkarieh, "SLAM aided GPS/INS navigation in GPS denied and unknown environments," *Int. Symp. GNSS/GPS*, vol. 4, no. December, pp. 6–8, 2004.
- [17] J. Collier *et al.*, "Autonomous Navigation and Mapping in GPS-Denied Environments at Defence R & D Canada," *Proc. NATO Symp. SET*, pp. 1–12, 2012.
- [18] "J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and

L. Fei-Fei. ILSVRC-2012, 2012. URL http://www.image-net.org/challenges/LSVRC/2012/.".

- [19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv Prepr. arXiv1409.1556, pp. 1–14, 2014.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition."
- [21] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, "SUN Database: Exploring a Large Collection of Scene Categories," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 3–22, 2016.
- [22] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks," *arXiv cs.CV*, vol. 9, p. 6846, 2016.
- [23] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 fps with deep regression networks," *Eur. Conf. Comput. Vis.*, 2016.
- [24] maps.google.com. Google, Google Maps, [Online] Available: http://maps.google.com. [Accessed: 24-Aug-2017].
- [25] www.bing.com. Microsoft, Bing Maps, [Online] Available: http://www.bing.com/maps. [Accessed: 24-Aug-2017].
- [26] www.dji.com. DJI, Phantom 3 pro specs, [Online] Available https://www.dji.com/phantom-3-pro/camera. [Accessed: 24-Aug-2017].
- [27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv Prepr. arXiv1412.6980, pp. 1–15, 2014.
- [28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," 2016.
- [29] M. Lin, Q. Chen, and S. Yan, "Network In Network," arXiv Prepr. arXiv1312.4400, p. 10, 2013.
- [30] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conf. Comput. Vis. Pattern Recognit., pp. 248–255, 2009.
- [31] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2018–2025, 2011.