Aerial Cross-platform Path Planning Dataset

Md. Shahid
Indian Institute of Technology Hyderabad
ee15resch02005@iith.ac.in

Sumohana S. Channappayya
Indian Institute of Technology Hyderabad
sumohana@ee.iith.ac.in

Abstract

Self-localisation mechanism in an unknown territory has been an interest area for humans since ages. Image matching is an obvious contender due to advancements in imaging devices and compute technologies. Deep learning methods have proven to be state-of-art in recent times but require large volumes of relevant data. Aerial image matching remains a challenging task due to the quality of images (e.g. platform disturbances, atmospheric effects), multiple types of on-board sensors (e.g. visual, thermal), variations in scales and look angles etc. To address these challenges, we present a cross-platform path planning dataset composed of images acquired from an aircraft and the Google Earth Engine (GEE). The proposed dataset contains manually aligned frames, corresponding match region, and semantic labeling of the images. Multiple galleries representing historical and instantaneous paths are generated. Our dataset envisages several realistic scenarios in cross-platform matching and semantic segmentation. We evaluate the performance of state-of-the-art image matching and segmentation algorithms on the proposed dataset. We will make our dataset freely available at `https://www.iith.ac.in/~lfovia/downloads.html`. *Further, a case study on utilizing an existing open-source dataset for cross-platform path planning is also presented.*

1. Introduction

Sensors used in aircraft navigation systems are typical of two categories, namely self-contained sensors and external support sensors. Conventional self-contained sensors include Inertial Navigation System (INS), Altitude and Heading Reference System (AHRS), attitude sensors and so on. These systems compute aircraft position, velocity, and altitude by a Dead Reckoning (DR) mechanism concerning the initial reference but suffers with drift issue. On the other hand, external support sensors use the Global Navigation Satellite System (GNSS), Global Positioning System (GPS), etc. in the feedback/correction mechanism.

These sensor inaccuracies (angular drift or linear motion), frequent GPS unlock, electromagnetic interference, remote locations or other unforeseen factors are a matter of concern while traveling for a long duration or at high speeds, or both. This demands for alternate navigation systems which are self contained and passive in nature. With the recent advancements in imaging devices (in size, resolution, multi-spectrum, etc.) and computing resources, image-guided NAV systems are ideal candidates.Such image-based NAV systems typically work in a human-in-the-loop model where critical inputs and decisions are provided by an operator at control station. However, a difficulty with the human-in-the-loop approach is the potential unavailability of reliable communication channels due to the same factors that can lead to GPS outages as mentioned in addition to transmission delays. Therefore, automatic image/vision-based aerial NAV systems, built over robust and real-time matching algorithms become essential.

As a first step to address this problem, we construct a cross-platform dataset composed of images (video frames) taken from an aerial vehicle as well as a satellite. Our dataset envisages various realistic matching scenarios and is annotated with appropriate semantic labels to aid the matching task. While there is extensive and excellent literature on cross-domain image matching and segmentation, neither methods offer a solution that can be directly applied to the problem at hand (discussed in section 4). Therefore, we propose a cross-platform dataset to bridge the gap of deep networks for cross-domain application. To the best of our knowledge, such a cross-platform path-planning dataset is not available in the literature. The rest of the paper is organized as follows: related work is presented in section 2, and the proposed dataset is described in section 3. The proposed dataset is analyzed with recent state-of-arts in section 4 with discussions, followed by concluding remarks in section 5.

2. Related Work

We briefly review aerial image datasets followed by a review of traditional and deep learning-based image matching/segmentation approaches. Aerial image datasets dedicated to scene detection and recognition are far fewer in number compared to standard image datasets (e.g. Ima-

geNet [1]). Some of the publicly available aerial datasets include DOTA [2], HRSC2016 [3], VHR-10 [4], SSDD [5] etc. These datasets are mainly geared for object detection and are built using the Google Earth Engine (GEE) [6]. The DOTA [2] dataset has objects for detection in aerial images with oriented bounding box annotations. It contains 2806 large-size images. There are fifteen categories including *baseball, basketball, tennis court, helicopter* etc. The instances vary greatly in scale, orientation, and aspect ratio. The HRSC2016 [3] dataset is a benchmark for ship detection in GEE images. It contains 1061 images and has more than twenty categories of ships in various appearances. Images are resized to 512×800. The VHR-10 dataset [4] has two spectral bands with varying ground resolutions. It has images with visual resolution ranging from 0.5 m to 2 m, and color infrared spatial resolution of 0.08 m.

As mentioned earlier, these datasets are derived using GEE images and are annotated to solve the object detection problem in aerial imagery. However, these dataset contain images from another aerial vehicle such as an aircraft or images taken from other sensor types that are important for solving the cross-platform aerial image matching problem. Other than object detection datasets, two popular publicly available satellite image datasets for segmentation are the INRIA [7] and EuroSAT [8]. The INRIA dataset [7] has satellite images of a few North American and European cities with *building* and *non-building* labels. The EuroSAT [8] is a publicly available dataset that is composed of 27,000 patches spread over 13 spectral satellite bands. The size of each patch is 64×64, and the patches are divided into ten classes some of which are *annual crop, river, highway, residential/industrial building, etc.*. Apart from these satellite datasets, Muller et al. [9] proposed an UAV dataset where images are taken from a low altitude Micro-Aerial Vehicle (MAC) platform. It has 123 video sequences for target tracking performance evaluation with over 110,000 frames along with bounding box annotation. Aerial drone dataset [10] has 20 semantic class labels with high resolution imagery acquired from an altitude 5-30 meters.

In recent times, there has been an active interest in finding geo-location for street view images. It involves warping satellite images (orthographic view) and aerial images (oblique bird eye view) over streets. To automate this process and facilitate deep learning algorithms, several datasets have been proposed in the literature [11, 12, 13, 14, 15, 16]. The Zurich city [11] and Toronto city [12] datasets are two recent datasets covering the urban environment. The Zurich city [11] dataset has acquired high resolution aerial images from low altitude MAV (5-15m) for 2 sq. km region with time-synchronized aerial ground-level street view images. The Toronto city [12] dataset acquired images over a large area from airplanes, drones, and cars driving around the city. This dataset consists of various overhead perspectives images captured for four different years. Instead of manually aligning aerial images, the authors have used digital elevation maps captured by airborne LIDAR and an on-board high-precision navigation system. Tian et al. [13] present a dataset that has four street view images and four bird's eye view images taken at each GPS location. The street view images (at 0, 90, 180, and 270 degrees with respect to true north) are from Google and the overhead 45 degree bird's eye view images are from Bing. Image matching geo-localization is carried out by multiple nearest neighbors matching. Buildings in the query and reference (search) are detected using Faster RCNN [17] region proposals with known geo-locations. A Siamese network is employed for paired and unpaired buildings based on the contrastive loss function. A graph is built using global and local matches and the final output is the mean of matched buildings.

Kyrkou et al. [18] proposed the aerial image database for emergency response (AIDER). Four disaster events including flood, fire/smoke, collapsed buildings/rubble, and traffic accident are annotated and studied in this work. Khurshid et al. [14] proposed the CrossviewRet dataset which consists of six distinct classes (namely *freeway, mountain, palace, river, ship and stadium*) with 700 images per class. Deepsat [15] proposed patches of size 28×28 pixels, covering six land-cover-classes - *barren land, trees, grassland, roads, buildings and water bodies*. Chiu et al. [16] proposed an agriculture-vision dataset for pattern analysis covering nine relevant classes. We would like to note that these datasets are either composed of satellite or aerial imagery with the limited cross-band or cross-domain or cross-platform association. Further, they do not take into account variability due to environmental/developmental changes, spatial/temporal mismatches, etc. We address this lacuna in our work by constructing the proposed cross-platform aerial image dataset.

We now briefly review traditional and modern approaches to image matching and semantic segmentation. In traditional approaches, sparse key-points and dense flow techniques are popular for image matching. Deep learning (DL) approaches on the other hand are relatively recent and learn task-specific features automatically but require data apriori. DL methods have shown significant performance improvement not just in image matching but in myriad computer vision tasks.

Sarlin et al. [19] proposed feature matching called SuperGlue. SuperGlue [19] has two major components - an attentional graph neural network and an optimal matching layer. Key-point descriptor along with the position is encoded into a single vector, and then uses alternating self and cross attention layers to create a more powerful representation. The optimal matching layer creates a scoring matrix, which finds the optimal partial assignment using

the Sinkhorn algorithm. Radiation-invariant feature transform (RIFT) [20] uses phase congruency (PC) instead of image intensity for feature point detection. RIFT prepares a maximum index map (MIM) for feature description from the log-Gabor convolution sequence. Experimental results show that RIFT is much more superior to SIFT and SAR-SIFT. DeepMatching (DM) [21] is inspired by deep CNN architectures and computes dense correspondences between images. It relies on a hierarchical, multi-layer correlation architecture.

SimNet [22] is a neural network-based approach that exploits the learning of non-metric similarity functions for instance search. The authors proposed an end-to-end trainable approach for image retrieval. Feature extraction is done by a pre-trained network in a feed-forward manner followed by a visual similarity network for content-based image retrieval. The output of the max-pooling layer is L2 normalized and flattened to formulate the input for the visual similarity network. Yang et al. [23] proposed matching of aerial images by extracting robust features using a CNN. Authors [23] proposed a modification of the VGG16 network (e.g., the grid structure of size 8×8), and features are pooled from the second, third, and fourth layers. A Gaussian mixture model (GMM) is used for dynamic inlier selection. However, the experimented database is not available publicly. The authors have compared with four variants of the SIFT methodology and reported results in terms of precision.

A fully convolutional network (FCN [24]) based image classifier has been trained on the INRIA dataset. The generalization capability of the FCN classifier was demonstrated by testing over images from North American and European cities which were not a part of their train dataset. However, generalization performance over entirely different urban regions say in Asia or Africa not presented. SegNet [25] is a deep encoder-decoder architecture for multi-class pixel-wise segmentation with VGG16 model. It has been designed and trained for urban road scene segmentation. It consists of 12 classes including roads, trees etc. DeepLab [26] is a state-of-art deep learning model for semantic image segmentation, where the goal is to assign semantic labels to every pixel in the input image. DeepLabv3+ [27] employs the encoder-decoder structure where DeepLabv3 is used to encode the rich contextual information. It uses ResNet18 model trained over the Camvid dataset [28]. We describe next the proposed dataset that addresses the lacunae in existing datasets, and evaluate the same with state-of-the-art image matching algorithms.

3. Proposed cross-platform Dataset

We now describe the proposed cross-platform aerial image dataset in detail. We elaborate the data collection experiment, the procedure to generate cross-platform data, followed by an analysis of the dataset. We then present a summary of the proposed dataset.

3.1. Data Collection

We have collected airborne data (video and flight parameters) using a manned aircraft in a designated urban area from an height of around 5000' with speed of 50-60 meters per second. An HD camera with a spatial resolution of 1920×1080 pixels and temporal resolution of 60 frames per second (fps) is mounted on the aircraft's belly in a forward-looking direction with tilt (60° down from horizon) due to mounting constraints. Intrinsic parameters (field of view and look angle of the camera) and extrinsic parameters (GPS and navigation sensors) are jointly termed as metadata and saved during the flight. Instantaneous trajectory information is transmitted via a radio frequency (RF) link and stored at the base station(e.g. ground control station). Trajectory parameters include location, speed, look angle, altitude and attitude of the aircraft. Video data is stored on-board (the aircraft) and trajectory parameters are stored in the ground station. We denote this airborne video data as *DTV*. Due to logistical/resource constraints, there was no synchronization between the *DTV* video and telemetry data. Telemetry parameters are noisy due to sensor inaccuracies and RF link loss (e.g. line of sight/platform/environmental disturbances). These parameters are filtered and passed through appropriate aircraft profile constraints concerning each parameter. Given the computational load of processing full HD images (1080p), all DTV frames (of UAV) are re-sized to a resolution of 640×480 pixels before further processing, while retaining the temporal resolution of 60 fps.

3.2. Generation of Cross-platform Data

To envisage a realistic scenario where the target image is from a different sensor or platform, we rely on the Google Earth Engine (GEE) [6]. GEE provides us the flexibility to fetch an image from a given latitude/longitude, look angle, the field of the video, etc. We simulate the aircraft flight trajectory in GEE and record a video for this flight trajectory using GEE. We denote this video as *SAT*. To further associate navigational sensors issues like translation or rotational drift, altitude/speed mismatch, etc., we modify the intrinsic and extrinsic parameters of trajectory data and generate further test data. This data simulates possible trajectory deviations in terms of aircraft altitude/heading, camera zoom/look angle, time of the day, date of image acquisition, and so on. We generate these videos at a spatial resolution of 640×480 pixels at 60 fps for 12 years. Sample frames from the proposed dataset are shown in Fig. 1. We have named a few DTV frames that contain clearly identifiable regions, e.g., a frame containing a clearly visible blue roof as *DTVblue* and similarly, a clearly visible red roof as *DTVRed*. A typical city scene with roads and build-

ings is denoted as *DTVCity*. Fig. 1 represent samples for a location from the first flight path gallery (e.g. DTVA) of the dataset. Left-top image is UAV image (e.g. DTV-Reference) whereas right-top is the corresponding SAT image from manually aligned best-match gallery (e.g. SAT-BM). The next two images are retrieved from SAT-Year-wise galleries for the same location for two years. Historical and atmospheric effects can be noticed from second row SAT images (e.g. left and right images) respectively. SAT-Year09 image has more greenery and non-existence of white building (next to blue roof top). Whereas, SAT-Year12 image has mild clouds. SAT-Year-wise images are aligned manually for initial locations. These SAT-year-wise variability lead to an added challenge in designing image matching algorithms in this cross-platform setting. Similarly, UAV images of the same location in the second and third flight paths are shown in Fig. 2.

We generate *Target-bin*-profile curve for query images manually(e.g. DTV frames). For each query image, first find out *Target-bin*(e.g. start index to end index) visually in each SAT gallery. *Target-bin* is set of frames containing at least 50% scene common with query image. For each query image, we find overlap using manual marking of corresponding points. This overlap score for a query image against each SAT image in *Target-bin* region of SAT gallery formulate *Target-bin*-profile. It helps to find best match index(e.g., position) for a query image in SAT gallery. Fig.3 illustrate our manual point correspondences pictorially. The first row displays input images (e.g. *DTVRed*) and SAT image (from SAT-Year12 gallery) shown in Figs. 3a and 3b respectively. We mark corresponding points manually and estimate homography. Using this homegraphy matrix find overlap. The second row shows the corresponding overlap area (concerning image size) in Figs. 3c and 3d respectively. Images with higher overlap are indicative of a better match.

Figure 1: Representative image from the 1st flight path DTVA (top-left) and corresponding manually matched satellite images. Top-right is the best matching satellite image, bottom-left and bottom-right are satellite image matches over different years (best viewed in color).

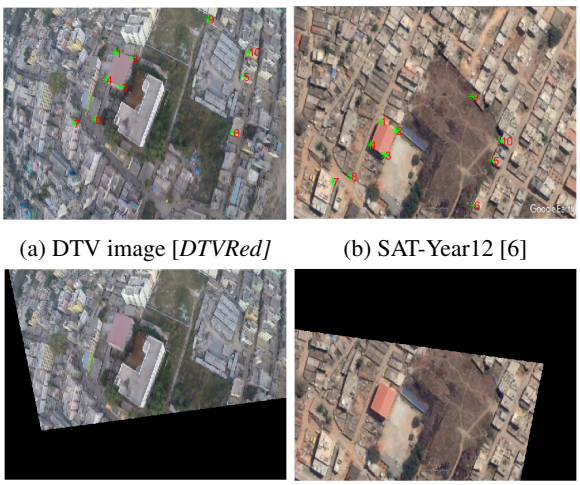
Figure 2: Representative images from 2nd and 3rd flight paths.

3.3. Semantic-segments Label Transfer

Unlike other aerial image datasets such as the INRIA dataset [7] which have two labeled regions (*Building/non-building*), we provide twenty labels that allow for the design of fine-grained segmentation algorithms. The labels include *Avenue, Building, Building Side, Construction Shed, Dry Field, Double Road, Mud Road, Pedestrian-path, Prominent Building, Runway, Runway Strips, Shed, Sheet, Sky, Street, Trees, Train, Vehicle, Vegetation Misc, Void, Water*. The semi-automatic region labeling methodology is described next. We first label images manually at regular intervals, with the interval depending on scene activity (content variation over the video). Labels for in-between frames are predicted (transferred) automatically, thus making our approach semi-automatic. Specifically, we propose a two-pass approach for label transfer that includes a forward pass and a backward pass. The labels transferred (predicted) using the forward and backward passes are merged appropriately to create a single semantic label for the in-between video frames. A few sample images and their corresponding labels are shown in Fig. 4.

3.4. Cross-platform Dataset from Open-source Data

We also construct a cross-platform dataset using the publicly available KAUST airborne dataset [9]. We select a building complex and designate it as the reference template. This reference template resembles our DTV (live camera of UAV) template. We manually search for this building complex using GEE for potential matching regions. While it was a difficult manual exercise due to urban development and drastic mismatch in resolution, we were able to find matching regions from GEE. We observed that SAT images are of very poor quality (perhaps be due to access restriction



(a) DTV image [*DTVRed]* (b) SAT-Year12 [6]

(c) DTV overlap image (d) SAT overlap image

Figure 3: Corresponding points illustration

of regulatory authorities). To simulate a fly-by scenario, we set up forward and backward way-points. These way-points formulate virtual paths of an aircraft, and thereby its trajectory. We collect SAT images for the forward and backward paths and designate them as *KAUSTComplexForward* and *KAUSTComplexBackward* galleries respectively. A sample reference template (e.g. *BuildingComplex*) from the KAUST dataset [9] and the manually retrieved SAT image from GEE can be visualized in Fig. 5.

3.5. Summary of the Proposed Dataset

The proposed dataset is composed of three reference DTV paths (viz., DTVA, DTVB and DTVC galleries) and corresponding SAT image galleries. Each gallery serves a different purpose as summarized in Table 1. We have identified 'best match' manually and provided it as part of the dataset. The efficacy of manual interventions (SAT-BM) can be visualized from Table 2. This table shows the average of the 2D correlation coefficient of corresponding SAT galleries with a DTVA-Reference gallery. SAT-Year-wise provides historical data over the period of time for DTVA. DTV-Test and SAT-Test contain few query images with target bin profile over SAT-Year-wise galleries. DTV-BC contains images from path B and C, along-with SAT-BC with equivalent SAT images. SAT-BM-seg consists of segmented maps. Additionally, we have provided manually marked corresponding frames for query images (of DTV-Test gallery) forming *target-bin region*.

4. Performance Analysis

We analyze the complexity of the proposed dataset using state-of-arts relating to image matching and semantic segmentation. We motivate the proposed cross-platform aerial dataset for image retrieval. To do so, we applied a variety of image matching methods over our dataset gallery for one query image (e.g. reference image from DTV-Test gallery). These methods include traditional matching methods like SSIM index [29], 2D correlation coefficient and contemporary deep methods [23, 30, 31]. The matching results are presented in Fig. 6. It is evident from the figure that all these methods have multiple local maxima and minima, making it difficult to indicate the best match image/region. This provides coarse evidence that the proposed dataset can potentially be used to improve the performance of aerial image matching algorithms.

4.1. Image Matching

Image retrieval and dense correspondences are the two parts of image matching. Image retrieval is applicable for a query image in full SAT gallery while dense correspondence in a specific *Target-bin* region. Content based image retrieval(CBIR) is a methodology of retrieving similar images for a query image. We have followed the same approach and experimented with BoG Spatial [32], pretrained VGG16 [30], pretrained ResNet50 [31], SimNet [22] and CNN-registration [23] for feature extraction. These extracted features are matched for DTV query and SAT image galleries. Lowest euclidean distance is considered best match. Searching over full gallery is considered a valid match over *Target-bin*-region. Top@N and mean Precision (mAP) for the state-of-art methods are summarized in Table 3 for all DTV-Test images over all SAT-Year-wise galleries and visually represented for (e.g. a single query image in SAT Year-wise galleries) in Fig. 7.

Performance of matching or outlier removal algorithms can be evaluated quantitatively using this points correspondence in *Target-bin*-region in terms of Percentage of Correct Keypoints (PCK), Mean Error (ME) and ratio-metric (r). PCK5 implies mismatch within euclidean distance of 5 pixels. PCK and ME for a query image over a SAT gallery (in *Target-bin*-region) are shown graphically in Fig. 8a and 8b respectively. As evident from Fig. 8a, ME values are low near best match index and increase either sides for all state-of-arts methods. Additionally, PCK curve in Fig. 8b verifies ME values of respective state-of-arts of Fig. 8a. To further validate quantitatively, we tested DTV test folder gallery images over SAT-Year-wise galleries and tabulated performance of standard evaluation metrices in Table 4. This quantitative and qualitative analysis have consistent findings for state-of-art methods.

To generalise the findings for any scene or target (e.g. *BuildingComplex*), we have used the open-source KAUST database [9] and generated two satellite galleries.We apply state-of-the-art matching methods to match the query image of Fig. 5a over the SAT gallery (e.g. *KAUSTComplexForward*) and evaluate with standard metrices (eg. PCK, ME,

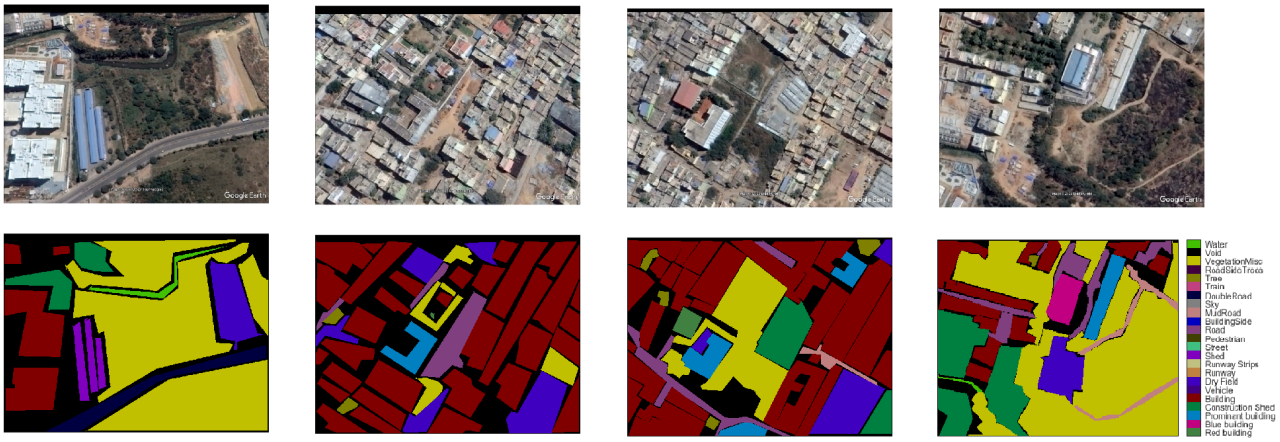


Figure 4: Samples from semantically labeled image gallery. Top row shows DTV images and bottom row shows corresponding semantic labels.

Table 1: A summary of the various galleries in the proposed dataset.

Name	Resolution	# Images/Galleries	Description
DTV-Reference	640×480	2500	Resized Reference DTV images of first path (DTVA)
SAT-BM	640×480	2500	Best manual match GEE imagery for DTVA
SAT-Year-wise	640×480	2500 per year	Aligned matching GEE satellite images from 2009–2020
SAT-Drift	640×480	2000 (per case)	Simulating flight path drift (left and right) in GEE
DTV-Test	640×480	9×2	Real DTV test images with target-profile curve
SAT-Test	640×480	5×2	Five SAT sub-galleries with various perturbations
DTV-BC	640×480	2500×2	DTV images of second (DTVB) and third (DTVC) paths
SAT-BC	640×480	2500×2 per year	Matching GEE images (2009–2020) for DTVB and DTVC
SAT-BM-Seg	640×480	2500	Semi-automatically generated semantic segments
SAT-Test-Seg	640×480	100	Manually labeled semantic segments for test
KAUST-cross-platform [9]	640×480	450×2	Forward and backward SAT sequence for KAUST

(a) DTV image [9] (b) SAT image [6]

Figure 5: KAUST DTV [9] and SAT sample [6] images.

Ratio-metric). Results are tabulated in Table 4 alongwith previous findings over SAT-Year-wise galleries.

4.2. Semantic labeling

We now evaluate semantic labeling part of dataset with standard/state-of-art methods. As described in section 3.3, we manually labeled frames at periodic intervals and used label transfer for generating labels for intermediate frames calling semi-automatic labels. To counter the argument that generated semiautomatic labels performance can be achieved with augmentation, we train the state-of-art models (e.g. FCN [24], SegNet [25] and DeeplabV3+ [27]) with manual labels and the generated semiautomatic labels. These trained models are tested with images and labels of SAT-Test-Seg gallery. The efficacy of semi-automatic labels is clear from mean Intersection of Union(mIoU) in Table 5.

In order to further evaluate proposed semantic dataset with contemporary aerial dataset [7]. The INRIA dataset [7] has two labels viz., *building, non-building*. The authors claimed generalisation by training FCN network over few cities and testing over other cities. Since we could not find the trained model, we trained the FCN over a few cities of INRIA dataset [7] images and tested over other cities. The same trained model (due to generalisation claim), we applied over our proposed dataset images. Per-

Table 2: Correlation of DTV-Reference with SAT-year-wise data

Metric	SAT-Year09	SAT-Year12	SAT-Year16	SAT-Year20	SAT-BM	Remarks
Corr2 (gray)	0.100	0.139	0.145	0.083	0.192	Grey scale images
Corr2 (color)	0.103	0.134	0.149	0.088	0.196	Sum of correlation of channels

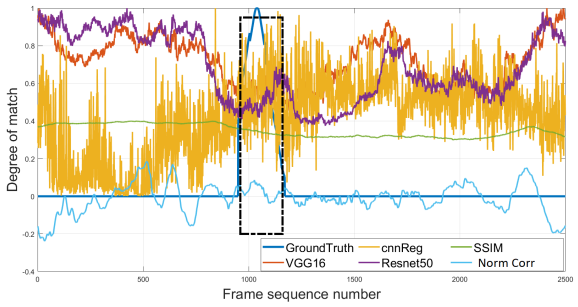


Figure 6: The performance of contemporary image matching methods on a DTV query image over a SAT gallery. The dashed box represents the expected target-bin region. The challenge with off-the-shelf methods is that all of them have multiple minima/maxima (Best viewed with zoom and color).

Table 3: Match performance on the entire gallery

No.	Method	Top@1	Top@5	Top@10	mAP
1.	BoG Spatial [32]	22.2	38.8	41.6	23.9
2.	CNN-registration [23]	7	50	50	15
3.	SimNet [22]	27.5	38.4	49.5	14.4
4.	Pretrained VGG16 [30]	27.7	36.1	41.6	23.1
5.	Pretrained ResNet50 [31]	16.6	19.4	22.2	18

Table 4: Matching performance over "*Target-bin*-region" for Proposed/KAUST [9]. Except last column higher value indicate better performance.

Method (Our/KAUST)	PCK5 (%)	PCK10 (%)	$r_{PCK \geq 50\%}$	Mean Error
SuperGlue [19]	2.17/ 3.5	8.03/ 11.2	3.39/ 0.38	159.7/ 99
RIFT [20]	1.1/ 0.2	3.52/ 0.48	1.24/ 0.2	146.8/ 201
Deep Matching [21]	18.8/ 22.9	35.9/ 49.2	18.42/ 22.5	79.3/ 57

formance can be visualised in Fig. 9 for inria [7] test city and proposed dataset image. As evident from Fig. 9, gen-

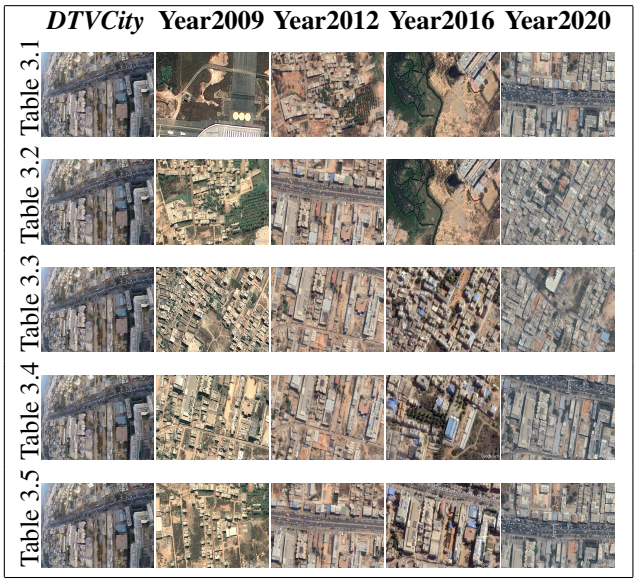
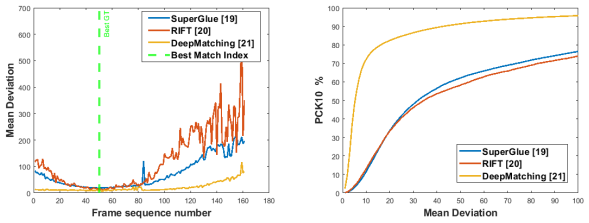


Figure 7: Top@1 searches in dataset for *DTVCity* query (round 2 or DTVB path). Rows are in the same order as table 3.



(a) Mean error curve (b) PCK curves

Figure 8: ME and PCK curves for a query DTV image in a SAT gallery. **Best viewed with zoom and color display.**

Table 5: Performance of semi-automatic semantic labels

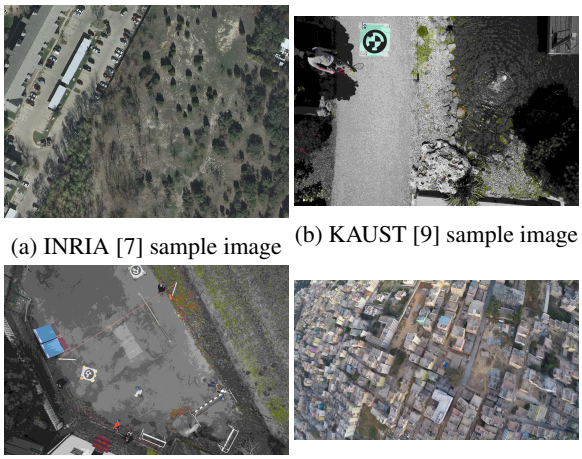
No.	Label type	FCN [24]	SegNet [25]	DeepLab V3+ [27]
1.	Manual	0.13	0.22	0.26
2.	Semiautomatic	0.17	0.26	0.38

eralisation is reasonable for INRIA [7] test image but poor for proposed dataset test image from SAT-Test-Seg gallery. From Fig. 9 it is clear that there is a need of relevant datasets applicable to other parts of the world.

Table 6: Dataset comparison

Dataset	Images	Resolution	Platform	Video	Classes	Application	Altitude	Seg. Maps
DOTA [2]	2086	4K × 4K	GEE	No	15	Detection	NA	No
HRSC [3]	436	512 × 800	GEE	No	20	Classification	NA	No
VHR [4]	800	600 × 600	GEE	No	10	Classification	NA	No
KAUST [9]	123 videos	1920 × 1080	MAV	Yes	NA	Tracking	5-25 m	No
INRIA [7]	360	20K ×20 K	GEE	No	2	Segmentation	NA	Yes
EuroSat [8]	27K	64 × 64	Sinetal-2	No	10	Classification	NA	No
Drone [10]	400	$4K \times 6K$	Aerial	No	20	Segmentation	5-30 m	Yes
AIDER [18]	8545	Multiple	Web	No	6	Emergency	NA	No
Proposed	25028	640 × 480	GEE, Aerial	Yes	20	Recognition	5000'	Yes

Figure 9: Predicted semantic labels (FCN fine-tuned over INRIA [7]) train images. Tested over images from other city of inria [7] and SAT-Test-Seg gallery of proposed dataset.



(a) INRIA [7] sample image (b) KAUST [9] sample image

(c) Aerial drone [10] image (d) Proposed dataset image

Figure 10: A comparison with related datasets shows that our dataset adds to the variety of content

We now present a brief discussion of our contributions in this work. To the best of our knowledge, the proposed aerial cross-platform path planning dataset is the first of its kind. Our proposed dataset is compared with contemporary datasets in Table 6. As points of comparison, we have considered the number of images, resolution, acquisition platform, disturbances, applications, etc. We find that contemporary datasets acquire images or videos (sequential frames) either from GEE or aircraft primarily for object detection, tracking, semantic segmentation. The proposed dataset is cross-platform containing carefully curated data acquired from GEE. The proposed dataset covers historical variation (i.e., urbanization over time), attitude distortions (e.g. altitude, zoom, look-angle etc.) and realistic environmental distortions (for e.g., small clouds). We qualitatively compare the proposed dataset with a few other datasets in Fig. 10 using representative image samples. As can be seen, the proposed dataset adds to the variety of aerial content thereby providing more data points for improving matching and segmentation algorithm performance.

5. Conclusions

The proposed dataset is useful for the development of aerial cross-platform path planning and evaluation of image matching/segmentation algorithms. It is aimed to bridge the gap between aerial navigation and deep methods. Additionally, we have demonstrated with a case study of using a query image from open source dataset and developing corresponding cross-platform galleries. We compared various aerial datasets from the literature and showed the utility of our dataset. We have evaluated several matching and segmentation algorithms over the proposed dataset. We have qualitatively and quantitatively demonstrated the usefulness of our dataset for CBIR, outlier rejection and semantic segmentation algorithms. We believe that this is a timely contribution given the increased use of unmanned aerial vehicles for a variety of applications ranging from emergency response to commercial to military.

References

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

[2] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.

[3] Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping Yang. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geoscience and Remote Sensing Letters*, 13(8):1074–1078, 2016.

[4] Hao Su, Shunjun Wei, Min Yan, Chen Wang, Jun Shi, and Xiaoling Zhang. Object detection and instance segmentation in remote sensing imagery based on precise mask R-CNN. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 1454–1457. IEEE, 2019.

[5] Tianwen Zhang, Xiaoling Zhang, Xiao Ke, Xu Zhan, Jun Shi, Shunjun Wei, Dece Pan, Jianwei Li, Hao Su, Yue Zhou, et al. LS-SSDD-v1. 0: A deep learning dataset dedicated to small ship detection from large-scale sentinel-1 sar images. *Remote Sensing*, 12(18):2997, 2020.

[6] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.

[7] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, 2017.

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[9] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European Conference on Computer Vision*, pages 445–461. Springer, 2016.

[10] http://dronedataset.icg.tugraz.at.

[11] András L Majdik, Charles Till, and Davide Scaramuzza. The zurich urban micro aerial vehicle dataset. *The International Journal of Robotics Research*, 36(3):269–273, 2017.

[12] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. *arXiv preprint arXiv:1612.00423*, 2016.

[13] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3616, 2017.

[14] Numan Khurshid, Talha Hanif, Mohbat Tharani, and Murtaza Taj. Cross-view image retrieval-ground to aerial image retrieval through deep learning. In *International Conference on Neural Information Processing*, pages 210–221. Springer, 2019.

[15] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2015.

[16] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838, 2020.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[18] Christos Kyrkou and Theocharis Theocharides. Emergencynet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1687–1699, 2020.

[19] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020.

[20] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Rift: Multi-modal image matching based on radiation-invariant feature transform. *arXiv preprint arXiv:1804.09493*, 2018.

[21] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision*, 120(3):300–323, 2016.

[22] Noa Garcia and George Vogiatzis. Learning non-metric visual similarity for image retrieval. *Image and Vision Computing*, 82:18–25, 2019.

[23] Zhuoqian Yang, Tingting Dan, and Yang Yang. Multi-temporal remote sensing image registration using deep convolutional features. *IEEE Access*, 6:38544–38555, 2018.

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[25] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[26] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[27] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[28] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] 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.

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