AI-Survey for Self-Flying Vehicles: Exploring the Challenges of Deep Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Everyone is talking about intuitive and automated transportation. An important and very challenging part of this research field are autonomous unmanned aerial vehicles (UAV) such as automated air taxis with a vertical take-off and landing (VTOL) capability. On one hand autonomous VTOLs will redesign our personal understanding of urban mobility, on the other hand automated UAVs will drastically change any kind of delivery or transportation services and much more. However, when studying computer vision and machine learning problems for UAVs or VTOLs it becomes increasingly difficult to stay up-to-date. We provide a survey for the topic of automated flights focusing on challenging Deep Learning problems with a state-of-the-art overview. We give an outline of possible sensor set-ups and AI based pipelines with leading results on established data sets. Finally we point out currently missing investigations.

1 Introduction

Autonomous flying is a rapidly advancing application area with a lot of opportunities for Deep Learning or Machine Learning based approaches. In common, two different pipelines can be distinguished:

1. The mediated perception approach which semantically reasons the scene [12, 11, 24] and determines the flight control decision based on it.
2. The end-to-end approach that learns the flying controls based on human behavior in and end-to-end manner [16, 2, 28].

Fig. 1 gives an overview of both pipelines where exemplary possible applications are shown. (a) SLAM is crucial for the local map and the vehicle pose within the environmental model. (b) Scene Understanding is essential to interpret the environment, e.g. to detect static and dynamic objects and their locations such as point wise classifications. (c) Sensor-Fusion is important to exploit the strengths of the different sensor types like classification for cameras, reconstruction for Lidar or dynamics for Radar. (d) End-2-End flying learns all decisions within a single network and can be treated as alternative approach. Compared to other kinds of automated vehicles, Autonomous Flying (AF) has specific challenges that characterize the use cases for Deep Learning:

- Scale Ambiguity: The 6DoF viewpoint ability for aerial vehicles impedes basic geometrical tasks like visual depth estimation or visual reconstruction in comparison to 3DoF use cases for ground robots and cars.
- Data Availability: Public data sets are rare compared to other computer vision tasks.
- Constraint Hardware: Applications have to run on a limited hardware with low energy consumption.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

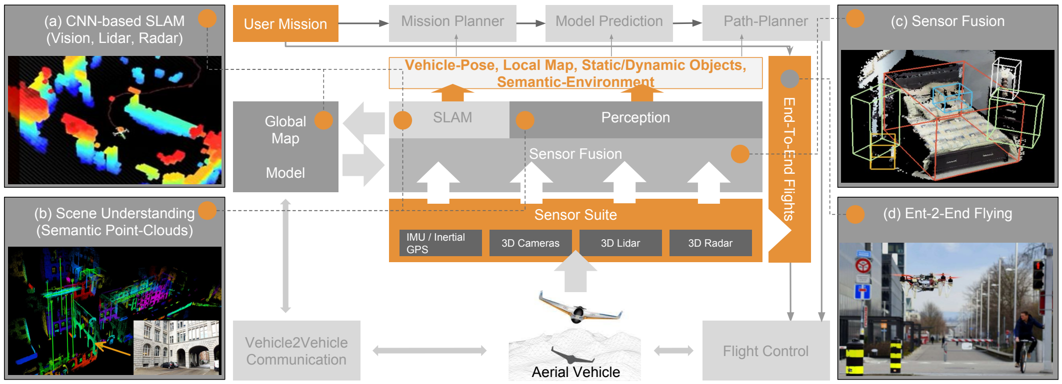

Figure 1: The principles of automated flying. The diagram outlines the state-of-the-art workflow. It all starts with the dedicated user mission. However, vehicle sensor data is essential to develop an environmental model for decision making and path planning. Several sensors like cameras, Lidars or Radars are crucial. In general two different paths are distinguished: 1. The mediated perception approach; 2. End-2-End Flying; Due to the complexity of the different tasks, leading approaches are mainly based on Machine Learning or Deep Learning, in particular Convolutions Neural Networks (CNN). (a-d) Illustrate example functions based on Deep Learning and their specific role within the pipeline [27, 4, 19, 7, 1].

Due to those challenging circumstances our short survey will cover an overview of public aerial data sets for specific tasks with currently leading applications. We give an overview of possible sensor setups, specific work-flows for sensor fusion and point out there strengths and weaknesses. The main part gives an overview of possible Deep Learning based applications for AF referencing exemplary state-of-the-art developments.

Aerial Data Sets							
↓ *Name/Task* →	Semantics	Objects	Odometry	Vision	Lidar	Radar	Size
Stanford Drone [19]	✗	✓(2D)	✗	✓	✗	✗	∼69GB[1]
DOTA [26]	✗	✓(2D)	✗	✓	✗	✗	2806F[2]
ISPRS [15]	✓(2/3D)	✗	✗	✓	✓	✗	∼20GB[3]
VisDrone2018 [30]	✗	✓(2D)	✗	✓	✗	✗	3190F[4]
Inria Aerial [17]	✓(2D)	✗	✗	✓	✗	✗	360F[5]
Drone Mapper	✗	✗	✗	✓	✓	✗	-[6]
Zurich Micro [18]	✗	✗	✓(6DoF)	✓	✗	✗	∼28GB[7]
EuRoC MAV [3]	✗	✗	✓(6DoF)	✓	✗	✗	∼20GB[8]
`Kitti` [12]	✓(2D)	✓(2/3D)	✓(3DoF)	✓	✓	✗	8110F[9]

[1]StandfordD: Several video sequences with instance tracking containing 7 classes in 8 different scenes

[2]DOTA: 2806 images (scale invariant) with 15 different object classes.

[3]ISPRS: Three different scenes (Toronto, Potsdam and Vaihingen) containing Lidar and RGB images ($\sim$ 40 image pairs per scene) with Semantic Pixel Classification (6 classes)

[4]VisDrone: 3190 frames in video and image footage with object boxes and tracking instances (12 classes).

[5]InriaA: Two pixel-wise classes (building, background) covering around 810 km^2 in 5 different regions.

[6]DMapper: Commercial data from https://dronemapper.com with HD-Lidar with accompanied RGB.

[7]ZurichM: A total of 5’237’298 2D keypoint observations and 1’382’274 3D points in Zurich.

[8]EuRoC: Around 10 indoor scenes with a static laser observer for odometry estimations.

[9]Kitti: Automotive Dataset with 8110 images with 2D and 3D Multiclass (8 classes) Object-boxes using Stereo Vision and Lidar such as 3Dof odometry.

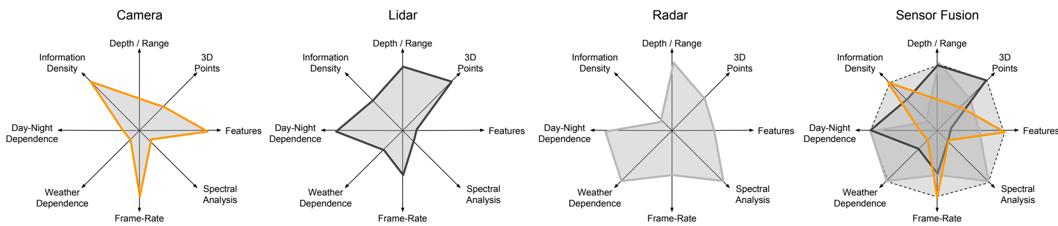

Figure 2: Sensor Fusion for Aerial Machine Learning. The figure shows four Star Plots analyzing the strengths and weaknesses of Camera, Lidar, Radar and Sensor Fusion. Individual strengths differ a lot. To benefit from all strengths Sensor Fusion is necessary. e.g. Object Classification can be trained easily using cameras [6] due to the good information density and the high value of visual features, whereas localization or reconstruction tasks benefit from Lidar sensing. Hence, 3D object detection mainly profits by fusion of cameras and Lidar, what can be proven by the Kitti leaderboard[12]. Radar has it advantages in the spectral analysis (2DFFT), i.e. it can directly measure the velocity of surrounding objects and many more tempo-spatial features. On the other hand Radar is resistent to weather or day/night conditions. Questionable is therefore rare usage of Radar data for ML in the domain of automated Flights.

2 Learning with Aerial Data

2.1 Public data sets

Different kinds of aerial data sets were established as it became important solving aerial computer vision tasks. To the best of our knowledge we summarized the most influential data sets in Tab. 1. At the moment, the main focus of research is aerial perception (e.g. multi-class object detection and tracking) and localization (e.g. odometry prediction) predominantly using camera inputs. All eight mentioned aerial data sets use cameras, only two use Lidar and none of them provide Radar ground truth. For comparison we mention the most comprehensive automotive data set Kitti [12]. Even Kitti does not provide public Radar data. We must conclude missing ground truth 3D boxes for aerial data and any kind of semantic Lidar annotations. Additionally, no one uses cameras with a large Field of View (FoV) or a stitched construction to cover 360 degrees of the vehicle.

2.2 3D Environmental Sensing

Lidar, Camera and Radar have different strengths and weaknesses that are important for solving Aerial Deep Learning Tasks. For a robust solution using Machine Learning Sensor Fusion is inevitable. Fig. 2 points out the advantages of Sensor Fusion. To our surprise, Radar is rarely used in perceptional fusion concepts, although it has standalone properties, like spectral analysis or weather resistance. We recommend a full fusion concept. Since, high quality data is inevitable for any kind of machine learning approach, we summarize the following Deep Learning challenges for our survey:

- Public Radar data (2D, 3D or Semantic ground truth) is missing.
- Additional ground truth for Lidar is (2D, 3D or Semantic) missing.
- Cameras are mainly used with a small FoV not covering 360 degrees.
- Highly redundant (minimum 3 senors types) data sets are missing

3 Deep Learning based Autonomous Flying

Fig. 1 shows the basic principle of AF. We point out opportunities using DL in four different algorithm groups in the field of DL, whereas basic function (e.g. Semantic Segmentation) can be part of several groups (e.g. Semantic Maps):

3.1 Localization, Mapping and Reconstruction

3.1.1 Visual Odometry

Dense Tracking and Mapping (DTAM) [21] was the first published method estimating odometry with simultaneous mapping. Here, a key frame based minimization of the photo-metric error was introduced. The following cost function was used:

$$\mathbf{C}_r = \frac{1}{\|I(r)\|} \sum_{m \in I(r)} \|\mathbf{I}_r(\mathbf{u}) - \mathbf{I}_m(\mathbf{v})\|. \tag{1}$$

Currently, still traditional cost minimization is state-of-the-art. Recently, Direct Sparse Odometry (DSO) was published by Engel et al. [9] with leading results on Kitti [12]. The global cost takes geometric attributes (lens distortion, exposure time) is designed as:

$$\mathbf{C}_r = \frac{1}{\|I(r)\|} \sum_{m \in I(r)} \|\mathbf{I}_r(\mathbf{u}) - b_r - \frac{t_r e^{a_r}}{t_m e^{a_m}} \mathbf{I}_m(\mathbf{v}) - b_m\|. \tag{2}$$

Recently, Delmerico et al. [5] published a comprehensive UAV benchmark for traditional visual odometry estimation using the EuRoC [3] (6Dof, see section 2.2). The ablation study focuses on real-time capacity and accuracy. Most accurate method ODROID is based on key frame based optimization like DTAM (1).

3.1.2 Unsupervised Odometry and Depth Estimation

To our surprise, Deep Learning is currently not dominating odometry challenges. However, promising results are recently published. GeoNet by Yin et al. [29] minimizes an additive cost function that is completely consisting of geometric unsupervised terms, i.e. a joint estimation of monocular depth, optical flow and egomotion. The overall cost is used to train a combination of CNNs. The full pipeline can be devided into a Rigid-Structure-Decoder such as a Non-Rigid-Motion Localizer. The loss is composed by:

$$\mathcal{L} = \sum \sum [\mathcal{L}_{rw} + \mathcal{L}_{ds} + \mathcal{L}_{fw} + \mathcal{L}_{fs} + \mathcal{L}_{gc}] \tag{3}$$

$\mathcal{L}_{rw}$ (warping loss) and $\mathcal{L}_{ds}$ (depth smoothness) define the rigid decoder. $\mathcal{L}_{fw}$, $\mathcal{L}_{fs}$ and $\mathcal{L}_{gc}$ describe the non-rigid motion localizer. The method outperforms significantly ORB-Slam on single Kitti Traces for trajectory accuracy (RMSE) and demonstrates the power of unsupervised Deep Learning.

3.1.3 Competitive Learning of Odometry and Depth

Recently, generative adversarial networks (GAN) outperformed lots of generative computer vision tasks. Milz et al. [20] used a cGAN doing Image-to-Image translation, i.e. Pix2Pix by Isola et al. [14], performing aerial depth estimation using Lidar ground truth. The overall loss minimizes the following term:

$$\mathcal{L} = \mathbb{E}_{x,y}\{\log(D(x,y))\} + \mathbb{E}_{x,z}\{\log(1 - D(x, G(x,z))\} + \lambda \cdot \mathbb{E}_{x,y,z}\{||y - G(x,z)||_1\} \tag{4}$$

The method is composed by a generative G and a descriptive network D (see Fig. 3) In order to create more and more accurate data, the loss of G is reduced, whereas a training step of D results in an increase of the partial loss $(1 - D)$ ideally. Hence, a competitive loss is the result. The advantage of the approach is, that the overall loss design is learned by the network itself.

Ranjan et al. goes a step further and combines a competetive such as a colaborative loss to an overall cost, which is composed by camera motion, monocular depth, optical flow such as motion estimation

[9] See reference [21] for detailed explanation of (1) and [9] for a detailed explanation of (2)

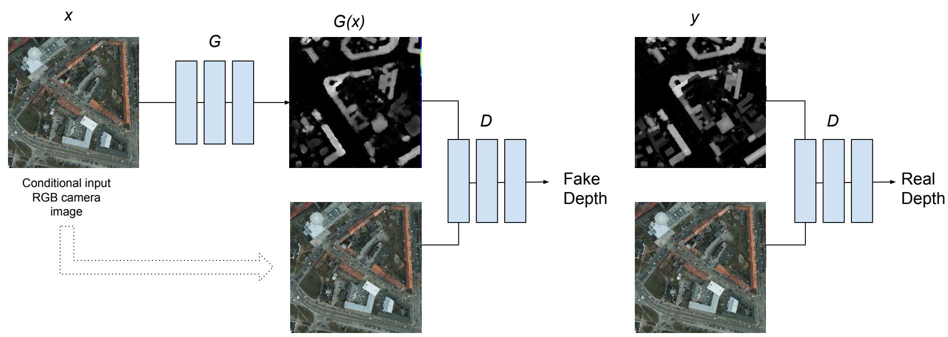

Figure 3: Competetive monocular Depth estimation using conditional GANs. The figure shows Milz et a. [25] implementation of the cGAN playing the minimax game. A generator G is used to create a fake image $G(x)$ (Depth reconstruction) based on the conditional input camera image x. The discriminator D tries to distinguish between a real Depth map $D(y)$ and fake image $D(G(y))$. The method shows promising results on the ISPRS data set[15].

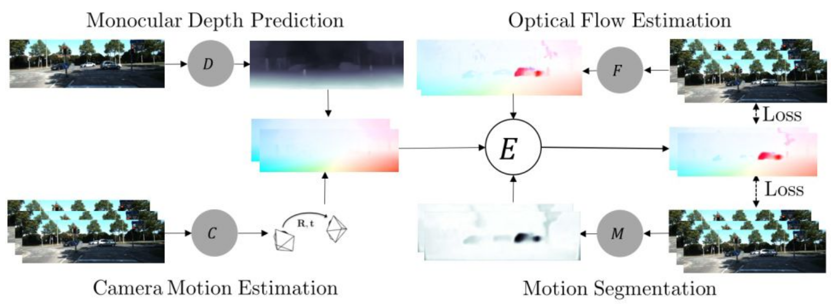

Figure 4: Collaborative and competitive odometry estimation and reconstruction. The figure is taken from [23] outlining the basic idea of the overall loss with promising results on Kitti.

3.1.4 Point-Cloud based SLAM using CNN based Semantic Points

SegMap by Dube et al. [8] uses Lidar based Point-Clouds to perform overall SLAM. The clue is an feature based global optimization function that is performed on semantic point clouds. The semantic point cloud classifcation is performed by a CNN. The model reduces drastically the number of tracked features and improves accuracy. The approach yields competitive results on Kitti (see. Fig.)

3.2 Perception and Scene Understanding

3.2.1 Visual Object Detection

The DOTA leader board [26] is good signpost for modeling visual object detectors. The currently leading approach is a mask R-CNN by He et al. [13]. The mask R-CNN performs instance object segmentation on DOTA with an overall mAP of 0.762.

Statistic	KITTI	Powerplant	Foundry
Duration (s)	114	850	1086
Number of robots	5	3	2
Number of segmented local cloud	557	758	672
Average number of segments per cloud	42.9	37.0	45.4
Bandwidth for transmitting local clouds (kB/s)	4814.7	1269.2	738.1
Bandwidth for transmitting segments (kB/s)	2626.6	219.4	172.2
Bandwidth for transmitting descriptors (kB/s)	60.4	9.5	8.1
Final map size with the *SegMap* descriptor (kB)	386.2	181.3	121.2
Number of successful localizations	113	27	85

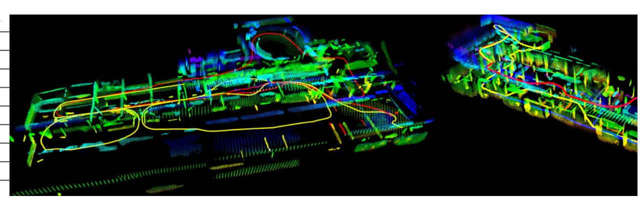

Figure 5: SegMap by Deube et al. performs localization and mapping based on Semantic Point Clouds sensed by Lidars. Results in the left table are promising ([7, 12]). The right area outlines qualitative odemetry and mapping predictions by Dube et al.

Classes	IoU Aerial GAN
Impervious surfaces	79.4%
Building	87.1%
Low vegetation	67.3%
Tree	70.3%
Car	24.1%
Clutter/background	30.7%
Mean IoU	**59.8%**

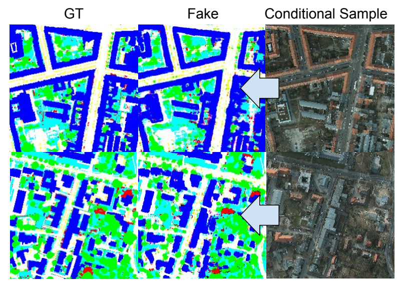

Figure 6: IoU for the aerial GANeration approach by Milz et al. [20] in the domain of image to semantic segmentation translation (ISPRS dataset[15]. The right part shows qualitative results.

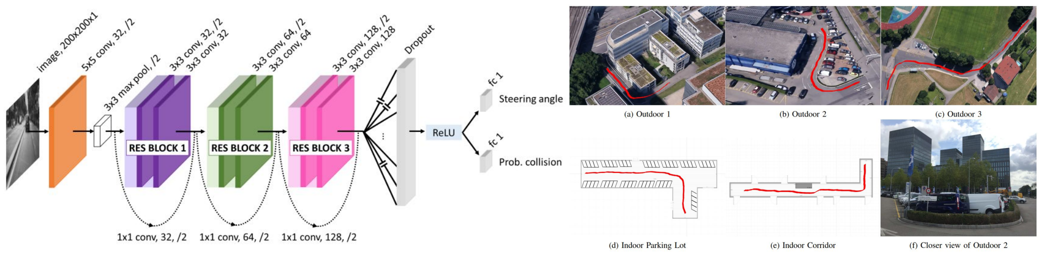

Figure 7: Dronet by Loquercio et al. (Parts of the Figure are taken from [16]). The left part shows the Resnet [10] architecture which is directly trained by the movement of observed agents in the urban area (cars, bicycles). The righ part (a-e) shows qualitative movement results in different scenes.

3.2.2 Semantic Segmentation

Aerial Semantic Segmentation was recently performed by Milz et al. using the ISPRS data set. Similar to section 3.1.3, the approach uses a cGAN to model the task as Image-to-Image translation problem. The results on the ISPRS are state-of-the-art. In Section 3.1.4 we have already referenced to semantic point cloud classification, which could be implicitly used for SLAM. As shown by Qi et al. [22] the overall idea is to approximate a symmetric f function on the point-set $x_{1..n}$ by applying local function h to get transformed elements of the data (5). This approximation is directly used in the overall loss to get a geometrical assessment and therefore a semantic segmentation of the points.

$$f(\{x_1, ...x_n\}) \approx g(h(x_1), ...h(x_n)) \quad (5)$$

3.3 Prediction, Planning and End-to-End flying

Prediction and planning for Aerial Vehicles are currently rarely solved using Deep Learning. Loquercio et al.[16] proposed an End-to-End approach imitating the movement of cars and bicycles using UAVs in Urban areas. The concept uses the ground truth motion of real cars/bicycles to train a CNN directly. The models architecture and qualitative results are shown in Fig.7

4 Conclusion

We have shown a compressed survey for AI based Autonomous Flights using Deep Learning for solving modular Tasks. We note, that DL has arrived in many parts like SLAM, perception, prediction or End-2-End flying. However, currently the main challenge is a comprehensive sensor redundant data set with three-dimensional ground truth (e.g. point semantics). To benefit from the strength of several sensor types. To our surprise, the main research focuses on cameras. Consequently, complex and comprehensive visual models are developed to perform tasks like reconstruction or depth estimation, e.g. competitive learning (section 3), which could be taken directly from Lidar or Radar. Hence, we highly recommend the usage of Lidar, Cameras and Radar.

References

[1] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *International Conference on Computer Vision (ICCV)*, 2017.

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[3] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. 35, 01 2016.

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. *CoRR*, abs/1611.07759, 2016.

[5] Jeffrey A. Delmerico and Davide Scaramuzza. A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2502–2509, 2018.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[7] Renaud Dubé, Andrei Cramariuc, Daniel Dugas, Juan Nieto, Roland Siegwart, and Cesar Cadena. SegMap: 3d segment mapping using data-driven descriptors. In *Robotics: Science and Systems (RSS)*, 2018.

[8] Renaud Dubé, Daniel Dugas, Elena Stumm, Juan Nieto, Roland Siegwart, and Cesar Cadena. Segmatch: Segment based place recognition in 3d point clouds. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5266–5272. IEEE, 2017.

[9] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. In *arXiv:1607.02565*, July 2016.

[10] Clément Farabet, NYU EDU, Camille Couprie, Laurent Najman, and Yann LeCun. Scene parsing with multiscale feature learning, purity trees, and optimal covers.

[11] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. *CoRR*, abs/1807.09713, 2018.

[12] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3d traffic scene understanding from movable platforms. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):1012–1025, 2014.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

[15] K. Khoshelham, L. Díaz Vilariño, M. Peter, Z. Kang, and D. Acharya. The isprs benchmark on indoor modelling. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W7:367–372, 2017.

[16] Antonio Loquercio, Ana I. Maqueda, Carlos R. del-Blanco, and Davide Scaramuzza. Dronet: Learning to fly by driving. *IEEE Robotics and Automation Letters*, 3(2):1088–1095, 2018.

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

[18] András L Majdik, Charles Till, and Davide Scaramuzza. The zurich urban micro aerial vehicle dataset. *Int. J. Rob. Res.*, 36(3):269–273, March 2017.

[19] Huynh Manh and Gita Alaghband. Scene-lstm: A model for human trajectory prediction. *CoRR*, abs/1808.04018, 2018.

[20] Stefan Milz. Aerial ganeration: Towards realistic data augmentation using conditional gan. In *2nd International Workshop on Computer Vision for UAVs*, Sept. 2018.

[21] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. Dtam: Dense tracking and mapping in real-time. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2320–2327, Washington, DC, USA, 2011. IEEE Computer Society.

[22] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016.

[23] Anurag Ranjan, Varun Jampani, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Adversarial collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. *CoRR*, abs/1805.09806, 2018.

[24] Davide Scaramuzza and Roland Siegwart. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE transactions on robotics*, 24(5):1015–1026, 2008.

[25] Min Wang, Baoyuan Liu, and Hassan Foroosh. Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial "bottleneck" structure.

[26] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge J. Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. *CoRR*, abs/1711.10398, 2017.

[27] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. *CoRR*, abs/1711.10871, 2017.

[28] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. *arXiv preprint arXiv:1612.01079*, 2016.

[29] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose, 2018.

[30] Pengfei Zhu, Longyin Wen, Xiao Bian, Ling Haibin, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.