In: AIAA SciTech 2023 Forum, National Harbor, Maryland, 23-27 January 2023
DOI:10.2514/6.2023-1148

# Development of Vision Guided Real-Time Trajectory Planning System for Autonomous Ground Refuelling Operations using Hybrid Dataset

Suleyman Yildirim*
*Cranfield University, Bedford, MK43 0AL, UK*

Zeeshan A. Rana†
*Cranfield University, Bedford, MK43 0AL, UK*

Gilbert Tang‡
*Cranfield University, Bedford, MK43 0AL, UK*

**Accurate and rapid object localisation and pose estimation are playing key roles during some of the real-time robotic operations such as object grasping and object manipulating. To do so, high-level robotic vision solutions need to be adopted. Computer vision approaches require a large amount of data to be able to create a perception pipeline robustly. Preparing such dataset to train the deep neural network could be challenging as the collection and manual annotation of huge amounts of data can take long hours and the development of the dataset needs to cover different conditions in weather and lighting. To ease this process, generating a synthetic dataset could be used. Due to the limitations of the synthetic dataset which will be described further down, instead of using a sole synthetic dataset, a hybrid dataset can be developed with the real dataset to overcome the limitations of both datasets. Even though the main objective of this study is to fulfil an autonomous nozzle insertion process for the ground refuelling operation of civil aircraft, the proposed approach is generic and can be adapted to any 3D visual robotic manipulation operation. This study is also offered to be the first visual trajectory planning control mechanism depending on the hybrid dataset to this date.**

## I. Introduction

Object location and pose are required for object grasping and object manipulating operations in robotics. As the objects differ in size, shape, location and orientation there is a need for a general solution for any problem. The optical sensor appears to be the most suitable choice to obtain and analyse the data as they have the ability to generate depth map and RGB images. There are many studies that take optical sensors into consideration to solve object-grasping problems. As investigated during Amazon's Robotics Challenge[1, 2], instance segmentation has been used to solve object manipulation and object detection problems.

Object localisation and pose estimation in 3D space has drawn attention in recent years and there are several studies that have adopted deep neural networks as an approach to such a problem. By Wong et al.[3] and Marion et al.[4] fitting a known 3D model into the point cloud data to find the 6D pose of the desired object has been adopted. To make this approach cover every object, Pavlakos et al.[5] have fit a generic 3D model using its key points onto the objects. As the fitting process depends on exact 3D models and estimation needs to be done for each object, the accuracy has been inadequate and the whole process has been time-consuming.

In this study, pre-defined key points have been used to locate the object in 3D point cloud data and to estimate the object's pose after successful detection in the RGB stream. As this study does not require masking the point cloud data with appropriate masking information, it is not necessary to process the whole point cloud data to locate the object. To find the key points of the object, computer vision and deep learning algorithms have been employed. As the synthetic

*PhD Researcher, Digital Aviation Research and Technology Centre, suleyman.yildirim@cranfield.ac.uk
†Senior Lecturer, Aerodynamics, Centre for Aeronautics, zeeshan.rana@cranfield.ac.uk
‡Lecturer, Robotics, Centre for Robotics and Assembly, g.tang@cranfield.ac.uk

data generation approach has been adopted, correctly and pixel-perfectly annotated unlimited data can be generated to train the custom-designed deep neural network. To overcome the reality gap problem with synthetic data which forms the limitation of the synthetic data and will be discussed further down, high-quality meshes have been applied to the 3D refuelling adaptor model, and different weather and lighting conditions have been simulated to make the deep neural network model robust to real-world conditions.

## II. Related Work

The demand for synthetic data and its popularity is rising as developing a dataset for custom objects takes long hours and effort. One of the biggest, richly-annotated, large-scale synthetic datasets of 3D shapes is ShapeNet, developed by Chang et al.[6], and consists of more than three million models. Tremblay et al.[7] have also used a synthetic dataset to train the neural network for their robotic manipulation study. By combining photo-realistic data with domain-randomised RGB images, they have reduced the reality gap regarding the 6D pose estimation of the objects. Their study has resulted in satisfactory accuracy for pose estimation of household objects in real-world robotics grasping. To add a single class to the neural network, it has been needing at least 120,000 images. So this was the main drawback of the study and resulted in a great problem.

Another study that uses a synthetic dataset has been presented by Hinterstoisser et al.[8]. They stated that training the neural network solely on synthetic datasets could have resulted in below-average if not the difference between real images and rendered images is reduced. Therefore, they have decided to freeze the lower layers of the pre-trained neural network and only train the top layers of the neural network with the synthetic dataset. By applying this method to the neural network, while the basic features of the real image domain remain suitable for training, the classification layer could be tuned for other classes.

In some cases, depth calculation or 3D pose estimation could be hard and laborious. To find the corresponding regions for the extraction of the depth information from the scene, Nikolenko et al.[9] have done the alignment of the 2D stream and 3D model of the object. In Gupta et al.'s study[10], a neural network has been trained using 3D model alignment of synthetic data and render of the synthetic object for detection and instant segmentation. Using a large-scale synthetic object dataset, Xiang et al.[11] proposed a method called PoseCNN to estimate 6D pose for only 21 known objects. Adopting a single-shot detection method, Kehl et al.[12] introduced a novel approach for detection and 6D pose estimation of 3D objects using RGB stream.

Hu et al.[13] developed a segmentation base method where every visible face of the object has been used as a dedicated form of 2D key points to estimate each face's local pose. To obtain the generic pose of the object, these local pose estimation has been combined to form the corresponding 3D pose. Even though their approach works decent in clutter scenes, the PnP algorithm they used for generic pose estimation, the solution became limited to be used for specific areas as this algorithm depends on the rigidity and 3D model of the model.

Zakharov et al. presented the Dense Pose Object Detector[14] i.e. associates object's masks with their corresponding 3D models. To train the Dense Pose Object Detector model, they have used both real and synthetic data. To eliminate the domain adaptation problem as mentioned above, they freeze the first layers after training the model with real data then they trained the last layers of the model with synthetic data. Calculating dense correspondences using the Dense Pose Object Detector lets the pose estimation be more accurate and robust, unlike the other studies that use regression of object's projections bounding boxes[15, 16] or describe the problem as a discrete classification[12].

By utilising a hybrid representation that uses geometric information to express the input image, edge vectors, key points and symmetry correspondence, HybridPose has been presented by Song et al.[17] to offer a different approach to 6D pose estimation problems. HybridPose differs from recent 6D pose estimation studies[16, 18] as they use leverage key points for intermediate representation. To output the edge vectors which can be found between neighbouring key points, the HybridPose method employs a prediction network. HybridPose employs dense pixel-wise correspondence i.e. reflects symmetric correspondences as most of the objects have reflection symmetry. As this study adopted a class-based approach to estimate the 6D pose of the object, it needs the intrinsic matrix of the camera to function properly.

DeepIM, pose matching deep neural network is proposed by Li et al.[19]. DeepIM estimates the next stage of the pose

by iterating the refined pose to match the rendered image with the observed image, as it has been given the initially estimated pose. The neural network has been trained using the 3D orientation and 3D location information to estimate the transformation in the relative pose. Due to being relied on the refinement of the previous pose information and lacking the key points between images and objects, this method tends to fail in pose approximation.

## III. Methodology

### A. Dataset Development

As shown in Figure 1 pressurised fuel adaptor also referred to as a bottom loading adaptor[20] is a connection adaptor that is used to deliver the pressurised fuel to the aircraft. The construction and design of the pressurised refuelling adaptor must conform to both MS24484-5 and MIL-A-25896 standards. "MIL-STD" short for Military Standard is a United States defence standard and helps to satisfy the standardisation objectives of the United States Department of Defence. Standardisation is beneficial in achieving commonality, interoperability, reliability, the total cost of ownership, compatibility with logistics systems and defence-related objectives, and ensuring products meet certain requirements[21]. To ensure maximum strength and durability, the pressurised refuelling adaptor is constructed of high-strength stainless steel and aluminium.

The dataset quality matters therefore the bad data in the dataset needs to be eliminated as they do not have any use. "quality" is a vague term. Considering that picking an empirical option and adopting an approach that provides the best result, explains the quality in a broader meaning. It can be said that the quality dataset enables us to overcome the problem by achieving the best solution. During the dataset development stage, it is helpful to have a concrete description of quality data. By corresponding to better performance in terms of feature representation, minimising skew and reliability, the quality dataset has particular aspects[22].

In the early stages, articulation of the problem forms the most critical part of the dataset development before generating or collecting quality data. Knowing how and what data to collect, and what to predict are the key questions to articulate the problem. The category of the problem such as clustering, regression or classification needs to be determined before formulating the solution and conducting any data exploration[23].

As the manual data collection process overwhelms and burdens people, establishing the data collection mechanisms to eliminate this repetitive and boring task is an important step during dataset preparation. Due to the limitations of both approaches, the dataset collection step has been divided into two stages. Establishing a camera rig system to collect real data from Boeing 737-400 forms the initial stage. Later on, by drawing a 3D design of the object and adding different materials to it and by simulating different weather and lighting conditions, the synthetic dataset has been generated[23]. In Figure 3, the elements that have been used in developing real and synthetic datasets can be seen.

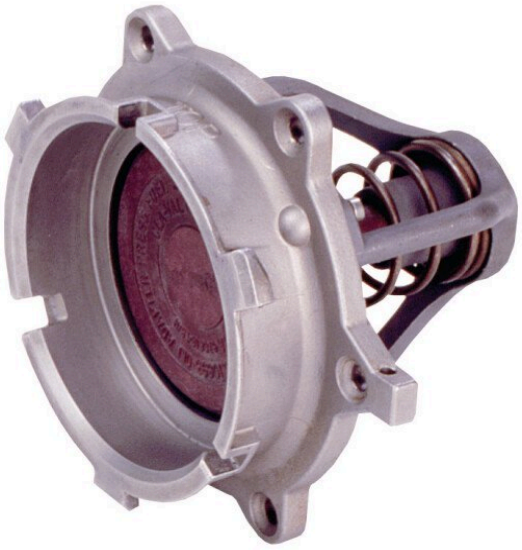

**Fig. 1 Pressurised Refuelling Adaptor[24]**

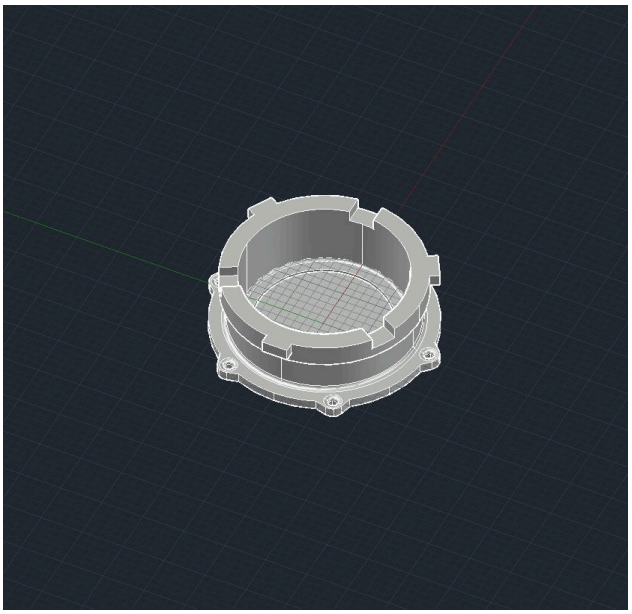

**Fig. 2 3D Design of Refuelling Adaptor**

**Fig. 3 Real Data vs Synthetic Data**

The hybrid dataset contains 770 training, 74 validation and 37 test images[25]. The image collection of the refuelling adaptor from Airbus 737-400 forms the real part of the dataset. The following augmentation methods have been applied to the dataset; clockwise and counter-clockwise rotations, horizontal flip, vertical flip and random Gaussian blur. Annotations have been stored in COCO format. The COCO format has been selected due to its widespread usage and simplicity. The COCO format stores the annotations as a bounding box, object classes etc. and image metadata[26].

In Figure 4, sample images from the collection of the refuelling adaptor can be seen. To improve the quality of the dataset, the 3D CAD model has been used to generate synthetic images. In these images, different textures, HDRIs, weather and lighting conditions have been applied to diversify and extend the coverage of the dataset. In Figure 5 some of the generated images are shown.

**Fig. 4 Real Dataset**

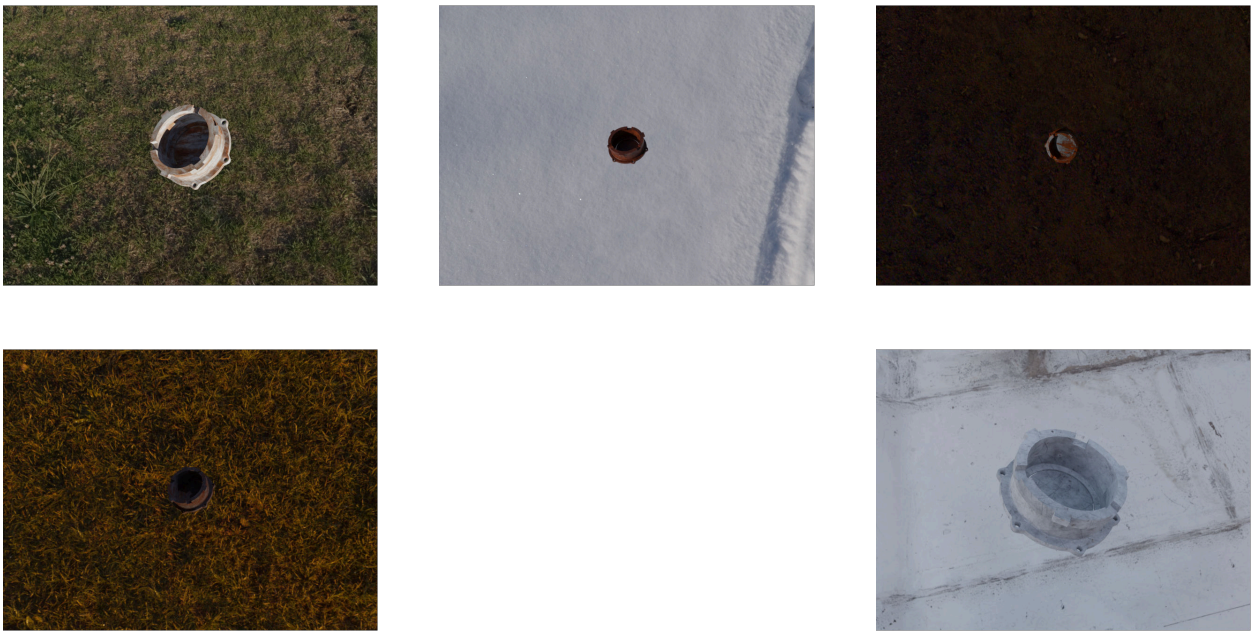

**Fig. 5 Synthetic Dataset**

Domain randomisation employs hundreds of variations of an object and the environment to enable the machine learning model to comprehend the general pattern easily. One of the biggest problems that are being encountered with synthetic data generation is the domain gap. Domain gap can be defined as the difference in the prediction space that the machine learning model would have if the model was trained on the real dataset. Domain randomisation technique helps us to deliver the synthetic dataset at its highest capability by improving the accuracy of the machine learning model[27]. In this regard, domain randomisation stands out as really crucial for the synthetic dataset generation stage of the research.

The limitations of both datasets are listed down below. The main moral of this research is to overcome these limitations by merging these two datasets together to exploit the advantage of both techniques.

Limitations of the real dataset can be listed as:

- Collecting and annotating thousands of images take long hours
- Even though there are tons of free datasets, the dataset needs to be collected and annotated for custom objects
- Annotations are generally created by humans and humans tend to make mistakes
- The content of the dataset might be involving the wrong classes of images
- Real dataset can only have basic annotations such as bounding box, segmentation or label

Limitations of the synthetic dataset can be listed as:

- Although it can copy most of the authentic properties of real data, some of the original content might not be copied so this might be critical for the case and affect the accuracy negatively
- The quality of the generated data is highly dependent on the quality of the 3D model

**B. Developing Custom Neural Network**

The custom neural network has been designed based on three principles to achieve the best accuracy and reduce the computational cost as much as possible while still having higher fps numbers in real-time operation. These principals are "Compound Scaling, Neural Architecture Search and Inverted Residual Block". These techniques have been introduced to develop different neural networks and improve their accuracy. To use their advantages, these three methods have been combined to develop the custom neural network. To have a better understanding, how they improved the accuracy is explained down below.

The most common way of scaling up a neural network was either by its dimensions- height, width, depth- or its image size before the compound scaling was introduced with EfficientNet[28]. EfficientNet's compound scaling approach scales the network's depth, width and resolution uniformly with a set of fixed scaling coefficients, unlike the conventional practice which is arbitrary scaling the factors. To use $2^N$ times more compute power, constant coefficients $\alpha$, $\beta$, $\gamma$ determined by grid search on the original model where they are used as the increase of the depth by $\alpha^N$, the width by $\beta^N$ and the image size by $\gamma^N$. Instead of using a different number of coefficients, EfficientNet employs a compound coefficient $\phi$ to scale the network uniformly. As the convolutional neural network is going to need more layers to capture fine-grained patterns as the input image gets bigger, balancing all the dimensions of the network gives better overall performance on the contrary scaling depth, width and resolution by different coefficients.

Neural architecture search is a reinforcement learning-based method where a baseline neural architecture has been developed by leveraging a multi-objective search that optimises for accuracy and FLOPS as they are the optimisation goal rather than latency. The objective function has been defined as a control mechanism to find the best-performing model in accuracy as well as FLOPS. The controller defines the model architecture and the model architecture has been used to train. After each training sequence, a reward function calculates and sends feedback to the controller to define another model architecture. This process repeats until the best-performing architecture is achieved according to the given accuracy and latency goals. The objective function is described as $ACC(m) \times [FLOPS(m)/T]^w$ in MNasNet[29].

Inverted residual blocks are placed in MBConv in Figure 6 below. They are introduced first in the Inception-V2[30]. Instead of reducing the number of channels, the inverted residual block increases the number of channels to 3 times. A standard convolution operation is computationally expensive. Therefore, a depthwise convolution operation is used to obtain the output feature map. The second convolution layer down-samples the number of channels in the final stage.

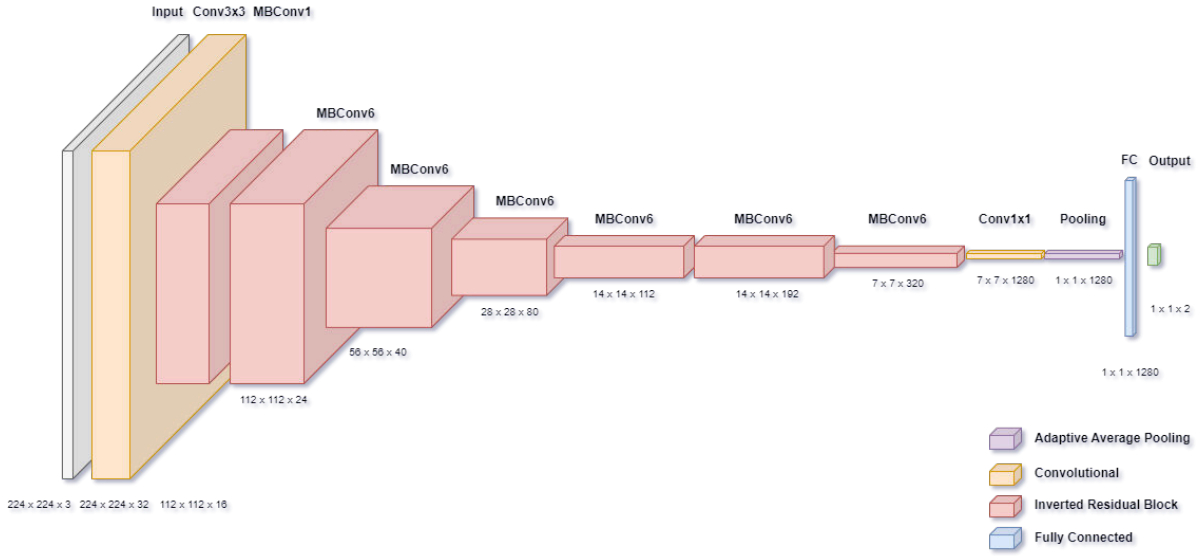


**Fig. 6 Custom Neural Network**

### C. Experimental Setup

The real dataset has been collected from Cranfield University's Boeing 737-400 aircraft. The synthetic dataset has been generated using Blender and zpy[31] open source computer vision toolkit for Blender.

The custom neural network has been trained on Cranfield University's HILDA high-performance computer. The allocated space for this operation had 112 Intel Xeon Gold 6258R CPU, 4 NVIDIA A100 80GB GPU, 377GB DDR4-2933 RAM and 330Tb storage capacity.

The Intel® RealSense™ D435 depth camera[32] has been used in the detection and localisation stages. The depth camera has Intel® RealSense™ D4 Vision Processor. It can stream both Active Stereo Depth up to 1280x720 resolution and RGB up to 1920x1080 resolution. Dual global shutter sensors allow up to 90 FPS and their Field of View is over 90°.

The structure of vision guided trajectory planning system can be seen in Figure 7.

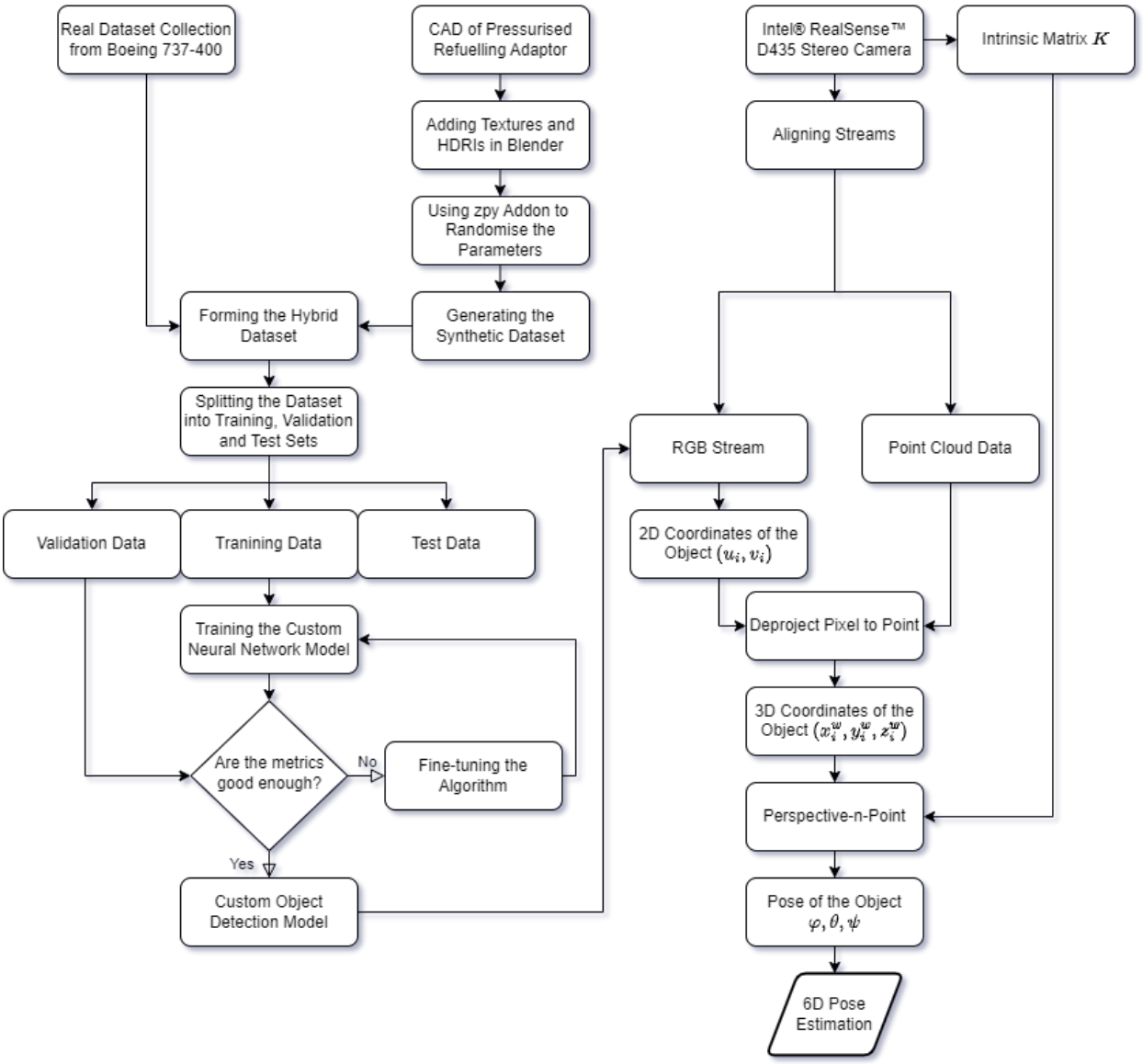


**Fig. 7 6D Pose Estimation Flowchart**

## IV. Results

The most uncomplicated performance metric to understand is accuracy, which is just the proportion of correctly predicted observations to all observations. One would assume that if the neural network model is accurate, it represents the best. Accuracy is an excellent indicator, but only when the values of the false positive and false negative rates are nearly equal in the datasets. As a result, the other metrics must be considered while determining the effectiveness of the neural network model. In terms of positive observations, precision is the proportion of accurately predicted observations to all predicted positive observations. The recall is the percentage of accurately predicted positive observations to all of the actual class's observations. The weighted average of Precision and Recall is the F1 score. Therefore, both false positives and false negatives are considered while calculating this score. Although F1-score is generally more beneficial than accuracy, especially if you have an uneven class distribution, it is not intuitively as simple to understand as accuracy. When false positives and false negatives cost about the same, accuracy performs best. It is preferable to include both Precision and Recall if the costs of false positives and false negatives are significantly different.

In the literature[33, 34], the distinguishing characteristics of the F1-score have been examined. F1 differs from MCC primarily in two ways. First, if the positive class is renamed negative and vice versa, F1 fluctuates, whereas MCC is invariant. The macro/micro averaging process may be extended to the binary situation itself[35], the F1 score is defined for both the positive and negative classes, and the two values are then averaged, and the average sensitivity and average precision values are used to address this issue. The behaviour of the micro/macro averaged F1 is more comparable to MCC and it is class-swapping invariant. This approach[36] is prejudiced and the community is still far from adopting it as a norm. Second, the F1 score is unrelated to the number of samples that were accurately categorised as negative. A number of researchers have recently drawn attention to the F1 measure's inadequacies[37, 38]. In fact, Hand and Peter[39] indicate that alternative measures should be considered due to their significant conceptual weaknesses.

The Matthews correlation coefficient is a more dependable statistical measure that only generates a high score if the prediction performed well in each of the four categories of the confusion matrix i.e. true positives, false negatives, true negatives, and false positives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - score = \frac{(1 + \rho) \cdot Precision \cdot Recall}{\rho \cdot Precision + Recall} \tag{4}$$

$$Matthews\ Correlation\ Coefficient = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{5}$$

Equations (1)–(5) refer to accuracy, precision, recall, F-score and Matthews Correlation Coefficient respectively[40]. The subscripts *TP*, *TN*, *TN*, and *FN* refer to true positives, true negatives, false positives, and false negatives categories respectively.

A precision-recall curve, which illustrates the balance between the precision and recall values at certain thresholds, exists due to the significance of both precision and recall. The optimum threshold to maximise both metrics can be chosen using this curve. The precision-recall curve can be condensed into a single value that represents the average of all precisions using the average precision (AP). The difference between the current and future recalls is determined using a cycle that iterates through all precisions and recalls, and it is then multiplied by the present precision. The AP is, in other words, the weighted sum of precisions at each threshold, where the weight equals the increase in recall. The mean average precision (mAP) is a measurement used to assess object detection models like R-CNN and YOLO. The mAP calculates a score by comparing the detected bounding box to the ground-truth bounding box. The more precise the neural network model's detections are, the higher the score is.

**Table 1 Metrics Table**

| | |
|---|---|
| Accuracy | 0.9662 |
| Precision | 0.9615 |
| Recall | 0.9709 |
| F1-score | 0.9662 |
| Matthews Correlation Coefficient | 0.9324 |
| $mAP@0.5$ | 0.996 |
| $mAP@0.95$ | 0.951 |

In Table 1, the scores for different neural network metrics can be seen. The following metrics are calculated as; Accuracy 0.9662, Precision 0.9615, Recall 0.9709, F1-score 0.9662, Matthews Correlation Coefficient 0.9324, $mAP@0.5$ 0.996 and finally $mAP@0.95$ 0.951.

The Intersection over Union is a metric that quantifies how well the predicted and ground truth bounding boxes match. The Intersection over Union helps to figure out if a region contains an object or not. In Figure 14, the ground truth and the predictions can be seen side by side. The custom neural network's predictions are very close to ground truth. This also can be observed from $mAP@0.5$ and $mAP@0.95$ scores. In Figure 15, the vision-guided trajectory planning system can successfully estimate the 6D pose of the aircraft refuelling adaptor. The necessary coordinate information in the X, Y, and Z axis and their angles are measured for precise nozzle insertion sequence.

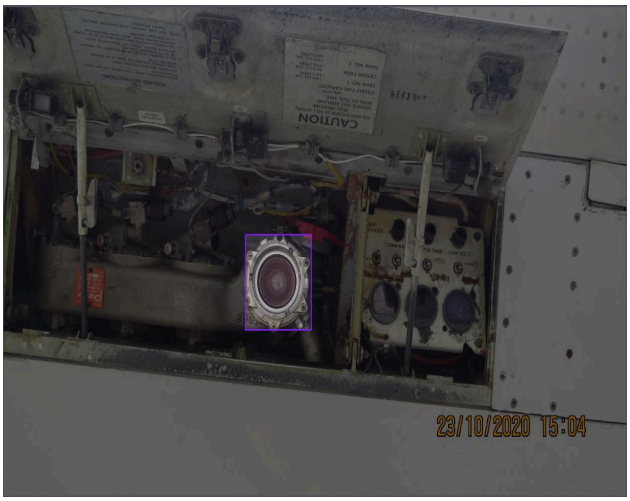


Fig. 8 Prediction 1

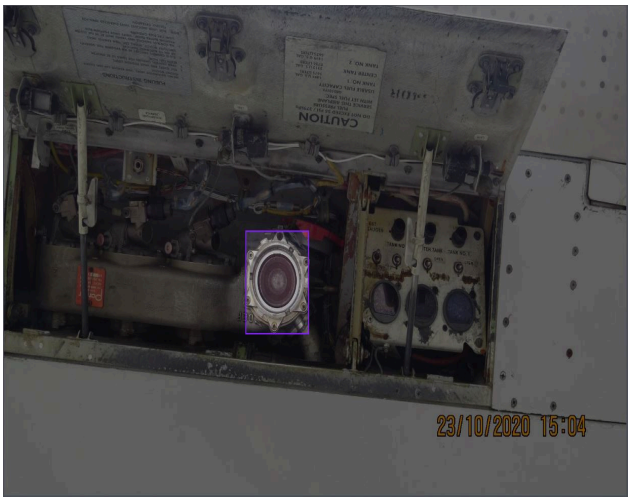


Fig. 9 Ground Truth 1

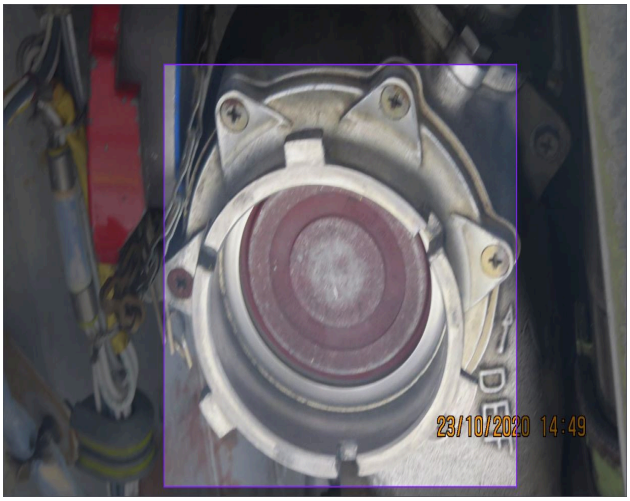


Fig. 10 Prediction 2

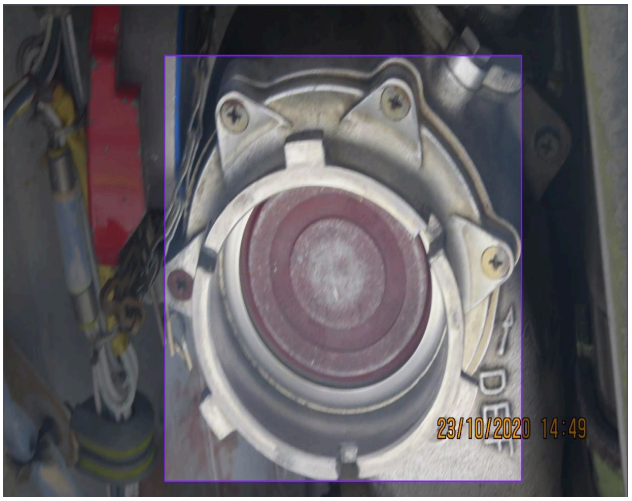


Fig. 11 Ground Truth 2

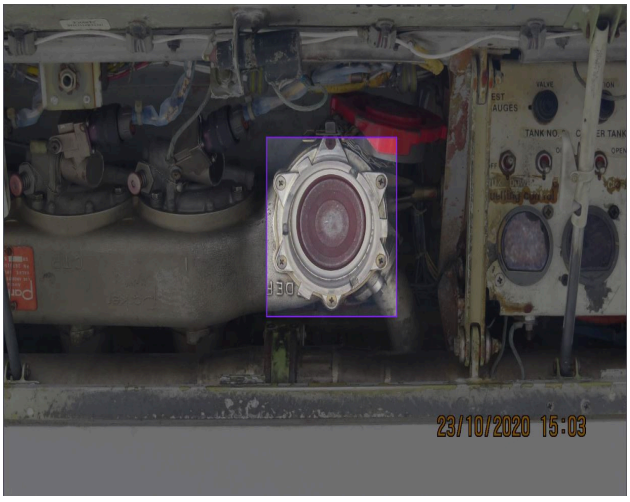


Fig. 12 Prediction 3

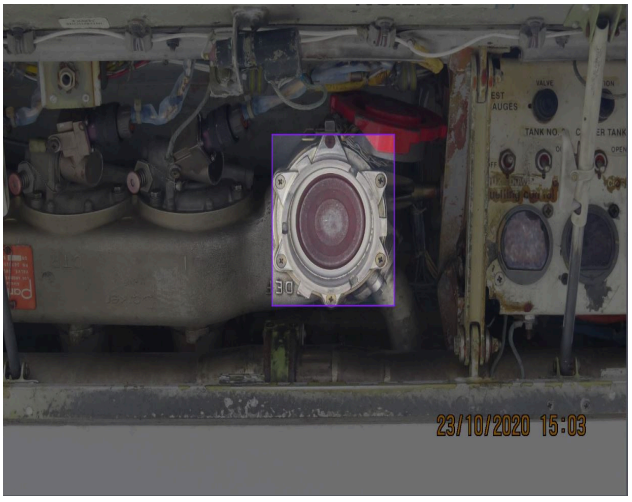


Fig. 13 Ground Truth 3

Fig. 14 Predictions vs Ground Truths


**Fig. 15 Experimental Result**

## V. Conclusion

In this research, a 6-DoF vision-guided real-time trajectory planning system for autonomous ground refuelling operation has been presented. The main contribution of this research is to show real data and synthetic data can be used together to eliminate the disadvantages of both approaches. Compared to traditional approaches which use either real data or synthetic data only, the custom neural network has been trained on a hybrid dataset. To keep the computational cost low and enable high fps real-time operation, the custom neural network has been designed fairly light. Instead of scaling the neural network arbitrarily, a compound coefficient $\phi$ has been used to scale the network uniformly.

To validate the results, an alternative measuring method such as a laser tracker needs to be employed. After the validation of the results, the developed system needs to be tested with a passenger aircraft to evaluate in real-life conditions. Future work will be addressed with these two stages.

## VI. Declarations

**Contributions:** Conceptualisation S.Y., Z.A.R., G.T.; Methodology S.Y.; Software S.Y.; Formal Analysis S.Y.; Draft Preparation S.Y.; Review and Editing S.Y., Z.A.R., G.T; Supervision Z.A.R., G.T; All authors have read and agreed to the published version of the manuscript.

**Acknowledgements:** This project was supported by the Republic of Turkey, Ministry of National Education through a grant awarded to the first author. The dataset has been collected from Cranfield University's Boeing 737-400 aircraft. The neural network has been trained on Cranfield University's HILDA high-performance computer.

**Availability of Data and Materials:** The data that support the findings of this study are openly available in Cranfield Online Research Data at `https://cord.cranfield.ac.uk/articles/dataset/AircraftRefuellingAdaptorLocalisation_v3i_coco/20445579`.

**Code Availability:** Due to the nature of this research, participants of this study did not agree for their code to be shared publicly, so supporting code is not available.

**Conflicts of Interest:** None of the authors reports any conflict of interest for this research.

**Ethical Approval:** This paper does not report research that requires ethical approval.

**Consent to Participate:** Consent to participate statement is not required.

**Consent for Publication:** Consent to publish statement is not required.

## References


[1] Anton Milan et al. "Semantic segmentation from limited training data". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 1908–1915.

[2] Max Schwarz et al. "Fast object learning and dual-arm coordination for cluttered stowing, picking, and packing". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 3347–3354.

[3] Jay M Wong et al. "Segicp: Integrated deep semantic segmentation and pose estimation". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 5784–5789.

[4] Pat Marion et al. "Label fusion: A pipeline for generating ground truth labels for real rgbd data of cluttered scenes". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 3235–3242.

[5] Georgios Pavlakos et al. "6-dof object pose from semantic keypoints". In: *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2017, pp. 2011–2018.

[6] Angel X Chang et al. "Shapenet: An information-rich 3d model repository". In: *arXiv preprint arXiv:1512.03012* (2015).

[7] Jonathan Tremblay et al. "Deep object pose estimation for semantic robotic grasping of household objects". In: *arXiv preprint arXiv:1809.10790* (2018).

[8] Stefan Hinterstoisser et al. "On pre-trained image features and synthetic images for deep learning". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.

[9] Sergey I Nikolenko et al. *Synthetic data for deep learning*. Springer, 2021.

[10] Saurabh Gupta et al. "Aligning 3D models to RGB-D images of cluttered scenes". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4731–4740.

[11] Yu Xiang et al. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes". In: *arXiv preprint arXiv:1711.00199* (2017).

[12] Wadim Kehl et al. "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1521–1529.

[13] Yinlin Hu et al. "Segmentation-driven 6d object pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3385–3394.

[14] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. "Dpod: 6d pose object detector and refiner". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1941–1950.

[15] Mahdi Rad and Vincent Lepetit. "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3828–3836.

[16] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. "Real-time seamless single shot 6d object pose prediction". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 292–301.

[17] Chen Song, Jiaru Song, and Qixing Huang. "Hybridpose: 6d object pose estimation under hybrid representations". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 431–440.

[18] Sida Peng et al. "Pvnet: Pixel-wise voting network for 6dof pose estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4561–4570.

[19] Yi Li et al. "Deepim: Deep iterative matching for 6d pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 683–698.

[20] URL: https://cla-val.ch/product/cla-val-340af-pressure-fuel-servicing-adapter/.

[21] 2014. URL: https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/412024m.pdf.

[22] URL: https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality?hl=tr.

[23] URL: https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/.

[24] URL: https://www.liquip.com/products/aviation/aviation-bottom-loading-accessories/bottom-loading-adaptor/claval-bottom-loading-adaptor.

[25] Suleyman Yildirim. "AircraftRefuellingAdaptorLocalisation.v3i.coco". In: (Aug. 2022). DOI: 10.17862/cranfield.rd.20445579.v1. URL: https://cord.cranfield.ac.uk/articles/dataset/AircraftRefuellingAdaptor_v3i_coco/20445579.

[26] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[27] Gerard Andrews. *What is synthetic data?* en-US. June 2021. URL: https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data/.

[28] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: (2019). DOI: 10.48550/ARXIV.1905.11946. URL: https://arxiv.org/abs/1905.11946.

[29] Mingxing Tan et al. "MnasNet: Platform-Aware Neural Architecture Search for Mobile". In: (2018). DOI: 10.48550/ARXIV.1807.11626. URL: https://arxiv.org/abs/1807.11626.

[30] Christian Szegedy et al. *Rethinking the Inception Architecture for Computer Vision*. 2015. DOI: 10.48550/ARXIV.1512.00567. URL: https://arxiv.org/abs/1512.00567.

[31] H. Ponte, N. Ponte, and S. Crowder. *zpy: Synthetic data for Blender.* 2021.

[32] Leonid Keselman et al. *Intel RealSense Stereoscopic Depth Cameras*. 2017. DOI: 10.48550/ARXIV.1705.05548. URL: https://arxiv.org/abs/1705.05548.

[33] Yutaka Sasaki et al. "The truth of the F-measure". In: *Teach tutor mater* 1.5 (2007), pp. 1–5.

[34] David MW Powers. "What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes". In: *arXiv preprint arXiv:1503.06410* (2015).

[35] Marina Sokolova and Guy Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information processing & management* 45.4 (2009), pp. 427–437.

[36] Vincent Van Asch. "Macro-and micro-averaged evaluation measures [[basic draft]]". In: *Belgium: clips* 49 (2013), pp. 1–27.

[37] Peter Flach and Meelis Kull. "Precision-recall-gain curves: PR analysis done right". In: *Advances in neural information processing systems* 28 (2015).

[38] Adam Yedidia. *Against the F-score*. 2016.

[39] David Hand and Peter Christen. "A note on using the F-measure for evaluating record linkage algorithms". In: *Statistics and Computing* 28.3 (2018), pp. 539–547.

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.