Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen
International Center for
Information Technology

R&D Project Proposal

Multi-input Conditional Object Detection

Gokul Krishna Gandhi Chenchani

Supervised by

Prof. Dr.-Ing. Sebastian Houben
M.Sc. Deebul Sivarajan Nair

November 2023

1 Introduction

Object detection is one of the important functionality in computer vision and deep learning for a system to understand and analyze the visual environment[15]. Object detection in specific involves the identification and location coordinates of an object within a given input. Detection of objects and their coordinate details in a given input can be formulated by various techniques such as 1. Bounding box regression[13] in which the model generates four points (Eg. 2D Image input) surrounding the corners of the object enclosed as a rectangular box, 2. Single key-point[17] of the object which locates the center of the object and 3. Center key-point and the orientation of the pose (6-DoF)[16] for the object in which the first a center point is determined and later using that feature a 3D bounding box along with the pose of the object is obtained.

During object detection when an input such as an image with multiple objects that belong to different classes is passed to the trained model, the model generates a huge output vector of all the detected objects in the input image. In several scenarios, where the end use-case involves manipulation by a robotic arm, the total number of target objects to be manipulated in a single action is usually one. In this case, where only one object needs to be manipulated, the detection of all other additional objects in the input image seems unnecessary especially with the huge output vector containing data of all objects. One of the example of this usecase is the tasks involved in RoboCup@Work[6] where the mobile robot equipped with a manipulator needs to transfer certain objects from one workstation to another workstation. During the execution of this task manipulator can only pick or place a single object in a single action which is already determined by the task planner and after the objects are detected by the trained model, the the robot base aligns with the object to be manipulated before the action is executed. Thus, the detection of all other objects during this manipulation action simply increases the computational power in terms of tensor.

Conditional object detection is one of the possible approach to detect only a single object from the given input image and reduce the amount of computational power in terms of tensor. Conditional object detection and object detection are different. They aim to do different things. Conditional object detection looks for

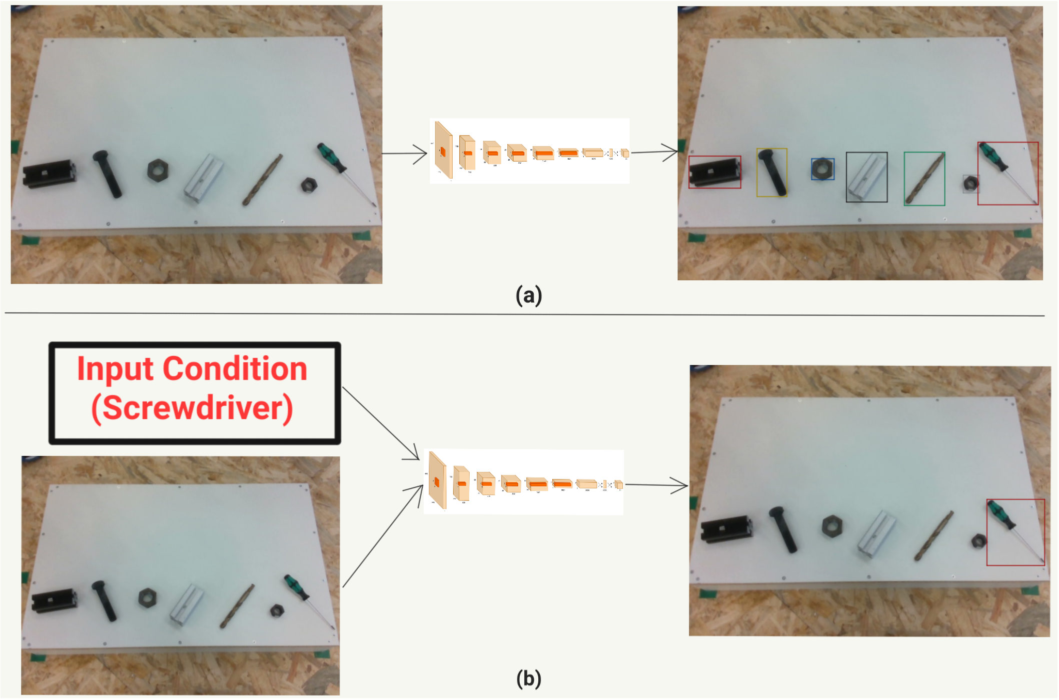


Figure 1: (a) Basic architectural representation of object detection. (b)Basic architectural representation of conditional object detection.

objects in test images that are similar to an object in a given support image. It can find objects even from categories it hasn't seen before. On the other hand, object detection finds all objects in categories it has been trained on, but can't identify objects from new categories. The way they are trained is also different. Conditional object detection is trained using pairs of support and query images, while object detection is trained in a typical way with lots of examples. They are evaluated in different ways. The effectiveness of conditional object detection is tested using a variety of pairs of support and query images, while the effectiveness of object detection is tested using many testing images[7].

In this R&D project, we propose an approach to use conditional object detection where an input condition such as an encoded label is passed along with the image to the model and expected output is the detection of the object based on the input condition. A model block diagram of both object detection and conditional object detection is shown in the figure 1.

1.1 Relevance of This R&D Project

The primary objective of this R&D project is to focus on the object detection of RoboCup@Work tasks and try to reduce the computational power to detect the object more efficiently and faster than the current used model. The results obtained in this project will be tested thoroughly on the RoboCup@Work dataset[6][19] and also by capturing images of the objects in real-time and evaluating the performance of the system by testing the research on youBot[1] at b-it-bots@Work[18] using RealSense cameras. This study will also be tested on the secondary dataset of HOPE[27].

The benefits and study done for this project will be helpful in implementing the proposed approach in youBot at b-it-bots@Work for the RoboCup@Work 2024 competition by improving the system performance and efficiency of the object detection model to accurately identify and then manipulate the object based on the object coordinates and pose from the model. Also, one of the main challenges to be addressed using this approach is to determine the accuracy of detection of objects on arbitrary surfaces which has always been a challenge due to multiple factors during the competition.

2 Related Work

J Gauthier[10] discusses the development of the conditional generative adversarial network (CGAN) on a face image dataset used for convolutional face generation. This development works on generative adversarial network (GAN) with an arbitrary conditioning y, which is an embedding space. The architectural framework of this approach is shown in figure 2.

Mathematical representation of GAN is shown in equation 1 and the representation of CGAN is shown in equation 2

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z\sim p_z(z)} [\log(1 - D(G(z))] \tag{1}$$

[1] `http://www.youBot-store.com/developers/kuka-youBot-kinematics-dynamics-and-3d-model`

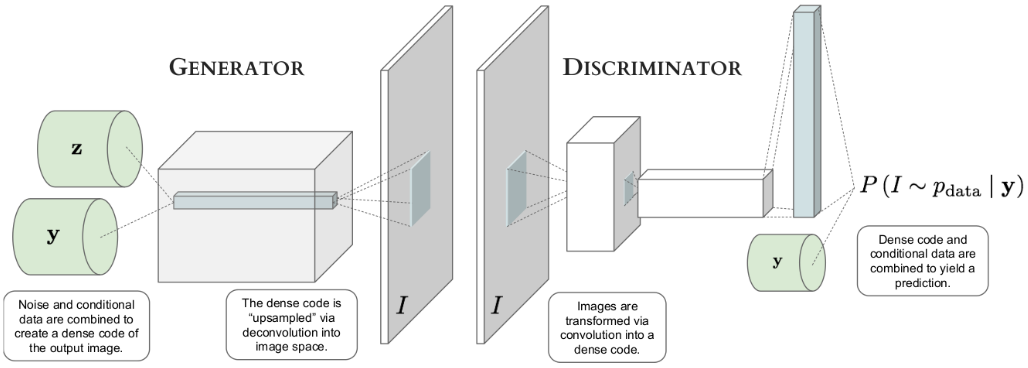

GENERATOR
DISCRIMINATOR
z
y
I
I
$P\,(I \sim p_{\text{data}} \mid \mathbf{y})$
y
Noise and conditional data are combined to create a dense code of the output image.
The dense code is "upsampled" via deconvolution into image space.
Images are transformed via convolution into a dense code.
Dense code and conditional data are combined to yield a prediction.

Figure 2: Graphical overview of the conditional generative adversarial network (CGAN) framework[10]

$$\min_G \max_D \mathbb{E}_{x,y\sim p_{\text{data}}(x,y)}\left[\log D(x,y)\right] + \mathbb{E}_{y\sim p_y, z\sim p_z(z)}\left[\log(1 - D(G(z,y),y))\right] \quad (2)$$

Ezeme et al.[5] discuss on the design and implementation of CGAN for anomaly detection in non-parametric multivariate data. This work also studies the realistic distributions of a given dataset and solving the issue of imbalance in data in anomaly detection tasks. The study uses a single class CGAN and entails the process of learning the pattern of the minority class data samples. This knowledge can be used to detect the minority class samples, allowing a binary class CGAN to train with a balanced dataset on both normal and harmful characteristics. From the results observed in this study, AD-CGAN performs better than most algorithms in conventional measures such as Precision, Recall, and F-1 Score.

Hyojin Park et al.[24] analyze another novel approach of using multiple conditions to detect an object using CGAN. In this approach, the model generates a new image by analyzing the backdrop of an original base picture and a new object defined by the text description in a specific location. The approach is implemented by a series of synthesis blocks in which the inputs constitute the seed feature map from a fully connected layers. Then the combination of image features from the backdrop image generates an output image and a segmentation mask. This study was carried out on Caltech-200 bird[28] and Oxford-102 flower[21] datasets.

Several other studies of conditional object detection were carried out in various domains such as abnormal vibration of aero engine analysis by Yang, Lu[29], fuse sar and multi-spectral optical data for cloud removal from sentinel-2 images by Grohnfeldt et al.[12] and one-shot conditional object detection by Kun Fu et al.[8]

3 Problem Statement

The current object recognition in youBot (RoboCup@Work) is implemented using YOLOv8[25][26] model, where the input image is fed to the model and the output received is the list of all the classes identified along with the bounding boxes in the given image on which the model is trained upon which the object that needs to be manipulated is queried from the list and the co-ordinate details are sent to the manipulator to grasp the object.

As per the existing implementation, the system knows which object to manipulate from the planner even before passing the image as input to the model. This sometimes can increase the tensor computation while identifying all the possible classes in the input image.

Another challenge while using the current YOLOv8 model on the RoboCup tasks is that the identification of objects on arbitrary surfaces is not very reliable and sometimes also identifies an object with an incorrect class.

The intended objective of this R&D project is to study and use the concept of conditional object detection and train a model for which there are multiple inputs such as image I_{hxwx3} and encoded label l (using an encoding approach). Any fusion methods such as concatenating the latent vectors of input image and input label to the model can be studied through which the output of the image contains a one-point coordinate and pose of the object (x, y, z) of the input label class.

Another objective is the reduction of size of the output vector which also reduces the space and time taken for the processing and detection of the object in the given image.

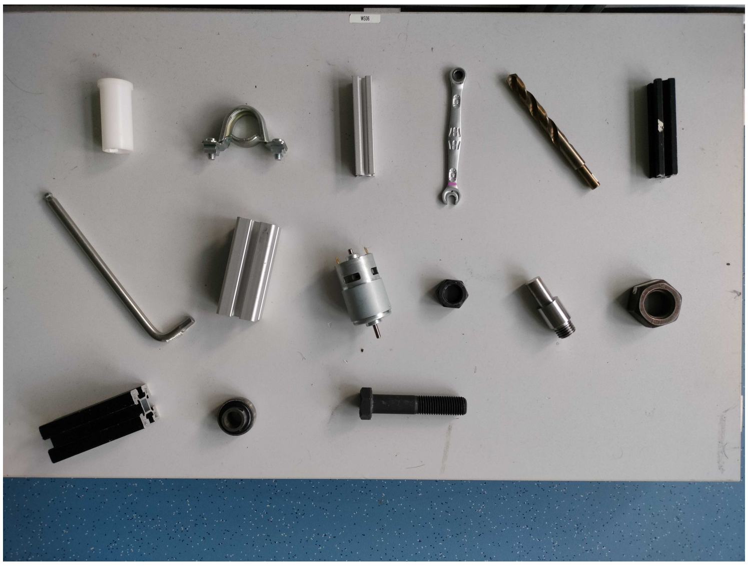


Figure 3: Image showing objects of RoboCup@Work dataset

3.1 Datasets

In this R&D project, we intend to study and analyse the use of conditional object detection on two datasets one being a primary dataset of RoboCup@Work(industrial objects) and the other being a secondary dataset of HOPE(Household Objects for Pose Estimation with 6-DoF).

3.1.1 RoboCup@Work Dataset

RoboCup@Work contains a set of 18 objects including basic and advanced objects and two containers. The dataset contains around 1000 images. The objects used in this dataset are shown in figure 3.

3.1.2 HOPE Dataset

HOPE(Household Objects for Pose Estimation) contains 28 toy grocery objects in 50 scenes from 10 household/office environments [22]. The dataset contains around 200 labeled images in different scenes and environments. The objects used in this dataset are shown in figure 4

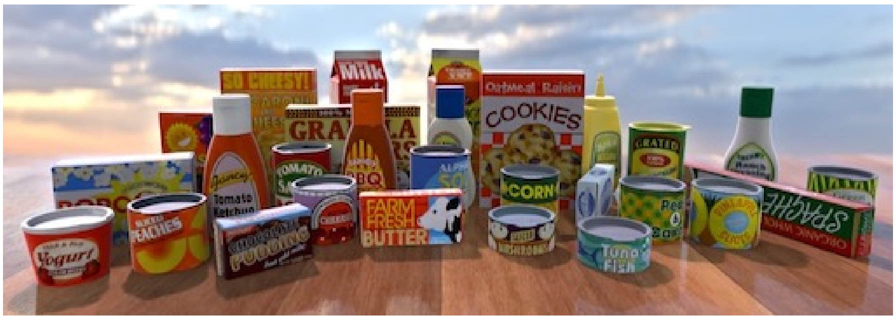


Figure 4: Image showing objects of HOPE dataset

3.2 Encoding Approaches

The process of converting a categorical label data into numerical data such that the converted data can be easily understood by the model to process is known as encoding approach. In this R&D project, we intend to apply encoding approaches to the input label before it is fused with the image input vector. Some of the possible encoding approaches are mentioned below.

3.2.1 One-Hot Encoding

One-hot encoding is a method of converting data of all the class representations of the trained model into one new binary variable by denoting an integer 1 to the class that needs to be detected and 0's to all the other classes [4][2] .

3.2.2 Error-Correcting Encoding

Error-correcting encoding is a technique for breaking down a multi-way classification problem into several binary classification tasks. This method assigns a unique n-bit vector to each label of class size m (where n > log2 m). Each bit vector is considered a unique coding for a label, forming a code matrix, denoted by C. Each row, C_i, and the value of the jth bit in this row, C_{ij}, represent specific codings.[3]

3.3 Fusion Approaches

Fusion approach in this project is used to fuse both input image and encoded input label into a single vector and pass it to the model to detect the object in the image based on the inputs provided. In this case the encoded input label vector is reshaped into the size of the input image vector and then used for fusing the these inputs into a single input vector.

3.3.1 Early Fusion

Early fusion is the process of concatenating features from many modalities at an early stage before they are transferred to the modeling process units.

- Summation of latent vectors of image and label text into a single vector.
- Text to visual features transformation for a CNN model[1][9].

The proposed approach for the problem statement described above will be performed on both primary and secondary datasets and the results will be compared with the existing YOLOv8 model used in youBot at b-it-bots@Work.

4 Project Plan

4.1 Work Packages

The R&D project will contain the following work packages

WP1 **Literature study**

- Conduct a comprehensive literature study on conditional object detection approach and its various use cases.
- Conduct a comprehensive literature study on encoding approaches.
- Conduct a comprehensive literature study on fusion techniques for image and encoded label.

WP2 **Experimental setup and analysis**

- Collect annotated dataset of RoboCup@Work and HOPE.
- Train a model for both the datasets.

- Perform basic object detection on the trained models.

WP3 **Implementation of encoding and fusion approaches**

- Analyse the performance of one-hot encoding for input label on the trained model.
- Analyse the performance of error-correcting encoding approach for input label on the trained model.
- Analyse the performance of early fusion technique of input image and encoded input label.

WP4 **Real-time testing of implemented approaches**

- Integrate the implemented approach with the youBot object detection module.
- Testing of the implemented approaches using real-time images from the youBot.
- Observing the performance results on the trained model for regular images and real-time images.

WP5 **Evaluation of approach and comparison with existing approach**

- Evaluate the performance of the system using proposed approach and compare with existing YOLOv8 model.
- Evaluate the output vector size,space and time using proposed approach and compare with existing YOLOv8 model.

WP6 **Project Report**

- Documentation of conditional object detection approach with relevance to the proposed R&D project.
- Documentation of state-of-the-art of the use of conditional object detection techniques.
- Documentation of state-of-the-art of the use of encoding approaches in deep learning.

- Documentation of state-of-the-art of the use of fusion techniques for image and encoded label in deep learning.
- Documentation of the proposed approach and the methodology.
- Documentation of the use of encoding approaches and fusion techniques used in the project.
- Documentation of the results obtained using the proposed approach.
- Documentation of the comparative results of the encoding approaches and fusion techniques used in the project.
- Documentation of the comparative results of the implemented approach with the existing YOLOv8 model on youBot.
- Draft R&D report explaining the findings of the research.
- Final R&D report explaining the findings of the research.

4.2 Milestones

M1 Literature review completed and best practice identified.

M2 Implementation of the encoding and fusion techniques on the trained model.

M3 Mid term report.

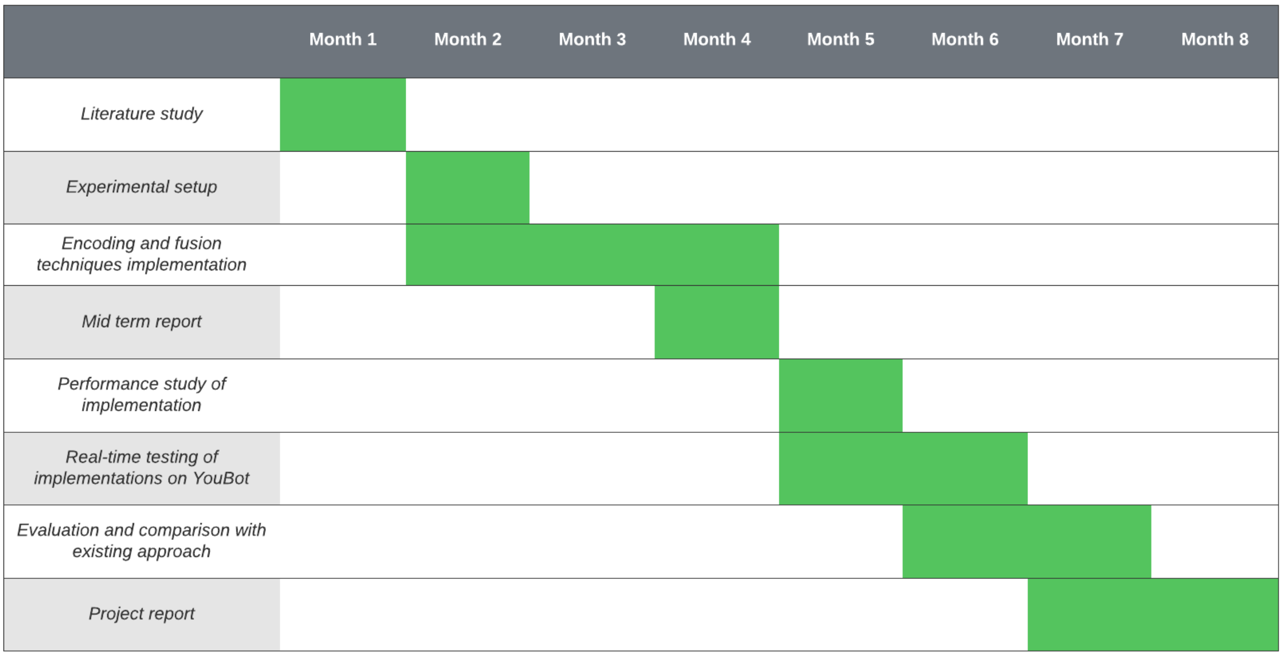
M4 Performance study and testing on real-time of the implemented approaches.

M5 Evaluation and comparison of the implemented approaches with existing approach.

M4 Report submission.

4.3 Project Schedule

The timeline for the project can be seen in the figure 5



	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8
Literature study	🟩							
Experimental setup		🟩						
Encoding and fusion techniques implementation		🟩	🟩	🟩				
Mid term report				🟩				
Performance study of implementation					🟩			
Real-time testing of implementations on YouBot					🟩	🟩		
Evaluation and comparison with existing approach						🟩	🟩	
Project report							🟩	🟩

Figure 5: Project schedule

4.4 Deliverables

Minimum Viable

- Literature review insights.
- Training a model on the RoboCup@Work dataset.
- Implement encoding and fusion techniques.
- R&D final report documentation.

Expected

- Implementation of all the mentioned encoding and fusion techniques.

- Performance and computational analysis of the implemented approach.

Desired

- Real-time testing and integration of implemented approach on the youBot.
- Comparative analysis of existing YOLOv8 approach vs Implemented approaches.
- Paper publication on the study conducted.

References

[1] George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning. *arXiv*, 2020. URL `https://arxiv.org/abs/2011.07191`.

[2] Dan Becker. Using categorical data with one hot encoding, 2018. URL `https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding`.

[3] Adam Berger. Error-correcting output coding for text classification. *In Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, 07 2001.

[4] Mwamba Kasongo Dahouda and Inwhee Joe. A deep-learned embedding technique for categorical features encoding. *IEEE Access*, 9:114381–114391, 2021. doi: 10.1109/ACCESS.2021.3104357.

[5] Okwudili M Ezeme, Qusay H. Mahmoud, and Akramul Azim. Design and development of ad-cgan: Conditional generative adversarial networks for anomaly detection. *IEEE Access*, 8:177667–177681, 2020. doi: 10.1109/ACCESS.2020.3025530.

[6] RoboCup Federation. Robocup@work. URL `https://atwork.robocup.org/`.

[7] Kun Fu, Tengfei Zhang, Yue Zhang, and Xian Sun. Oscd: A one-shot conditional object detection framework. *Neurocomputing*, 425:243–255, 2020. URL `https://api.semanticscholar.org/CorpusID:219004774`.

[8] Kun Fu, Tengfei Zhang, Yue Zhang, and Xian Sun. Oscd: A one-shot conditional object detection framework. *Neurocomputing*, 425:243–255, 2021. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.04.092. URL `https://www.sciencedirect.com/science/article/pii/S0925231220306779`.

[9] Ignazio Gallo, Alessandro Calefati, Shah Nawaz, and Muhammad Kamran Janjua. Image and encoded text fusion for multi-modal classification. *arXiv*, 2018. URL `https://arxiv.org/abs/1810.02001`.

[10] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester*, 2014(5):2, 2014.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv*, 2014. URL `https://arxiv.org/abs/1406.2661`.

[12] Claas Grohnfeldt, Michael Schmitt, and Xiaoxiang Zhu. A conditional generative adversarial network to fuse sar and multispectral optical data for cloud removal from sentinel-2 images. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1726–1729, 2018. doi: 10.1109/IGARSS.2018.8519215.

[13] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. *arXiv*, 2019. URL `https://arxiv.org/abs/1809.08545`.

[14] Jonathan Hui. Gan — cgan infogan (using labels to improve gan), 2018. URL `https://jonathan-hui.medium.com/gan-cgan-infogan-using-labels-to-improve-gan-8ba4de5f9c3d`.

[15] Steven Lang, Fabrizio Ventola, and Kristian Kersting. Dafne: A one-stage anchor-free approach for oriented object detection. *arXiv*, 2022. URL `https://arxiv.org/abs/2109.06148`.

[16] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, and Stan Birchfield. Single-stage keypoint-based category-level object pose estimation from an rgb image. *arXiv*, 2022. URL `https://arxiv.org/abs/2109.06161`.

[17] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. *arXiv*, 2020. URL `https://arxiv.org/abs/2002.10111`.

[18] MAS-group. b-it-bots@work, . URL `https://www.h-brs.de/en/a2s/b-it-bots`.

[19] MAS-group. Robocup@work dataset, . URL `https://github.com/b-it-bots/mas_industrial_robotics`.

[20] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *arXiv*, 2022. URL `https://arxiv.org/abs/2205.06230`.

[21] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

[22] NVIDIA. Hope dataset. URL `https://github.com/swtyree/hope-dataset`.

[23] Priyadarshini Panda, Abhronil Sengupta, and Kaushik Roy. Conditional deep learning for energy-efficient and enhanced pattern recognition. *arXiv*, 2016. URL `https://arxiv.org/abs/1509.08971`.

[24] Hyojin Park, YoungJoon Yoo, and Nojun Kwak. Mc-gan: Multi-conditional generative adversarial network for image synthesis. *arXiv*, 2018. URL `https://arxiv.org/abs/1805.01123`.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *arXiv*, 2016. URL `https://arxiv.org/abs/1506.02640`.

[26] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv*, 2023. URL `https://arxiv.org/abs/2305.09972`.

[27] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[28] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

[29] Lu Yang. Conditional generative adversarial networks (cgan) for abnormal vibration of aero engine analysis. In *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT*, pages 724–728, 2020. doi: 10.1109/ICCASIT50869.2020.9368622.