An overview of robot vision

Beatrice van Eden and Benjamin Rosman
Mobile Intelligent Autonomous Systems
Council for Scientific and Industrial Research
and
School of Computer Science and Applied Mathematics
University of the Witwatersrand
South Africa

***Abstract*—Robot vision is an interdisciplinary field that deals with how robots can be made to gain high-level understanding from digital images or videos. Understanding an image at the pixel level often does not provide enough information for decision making and action taking. In this case, higher level semantic information that describes the image is required. This helps the robot to accomplish complex tasks that require visual understanding.**

For robots to add value they need to be sufficiently effective at executing tasks in different settings. Despite many impressive advances in robot vision, robots still lack the ability to function as humans do in complex environments. Importantly, this includes being able to interpret and understand the perceptual complexities of the world.

Robot vision is dependant on ideas from both computer vision and machine learning. In this paper we provide a overview of the advances in these disciplines and how they contribute to robot vision.

I. Introduction

Although robots have gained widespread adoption in well-curated factory environments, there remain many issues when moving to unstructured settings, featuring cluttered scenes or unanticipated events. The perceptual challenges are not only at the level of recognising objects, but often for useful action further contextualisation is required.

For example, when a wind gust blows leaves through an open door, a human will know to close it. In order to interpret the scene in the same way, a robot will first have to recognise and identify the images of the leaves, and then that they are contextually out of place indoors.

In a complex human environment it is still a challenge for robots to perform complex tasks autonomously. This is typically handled through either controlling the environment or augmenting the sensors. For example, a robot to help paralysed army veterans around the house [2] works indoors and can operate in a cluttered environment but relies on customised QR code-like symbols to help the robot identify common objects. On the other hand, an autonomous robotic laundry folding system is restricted to only using certain clothing and operations [3]. Although these robots function in tightly controlled environments, the potential to extend their impact is great [4].

Robot vision (RV) is the field concerned with how a robot gains perceptual information about its surroundings. This draws on techniques from a number of areas, most notably computer vision (CV) and machine learning (ML), as illustrated in figure 1 [5].

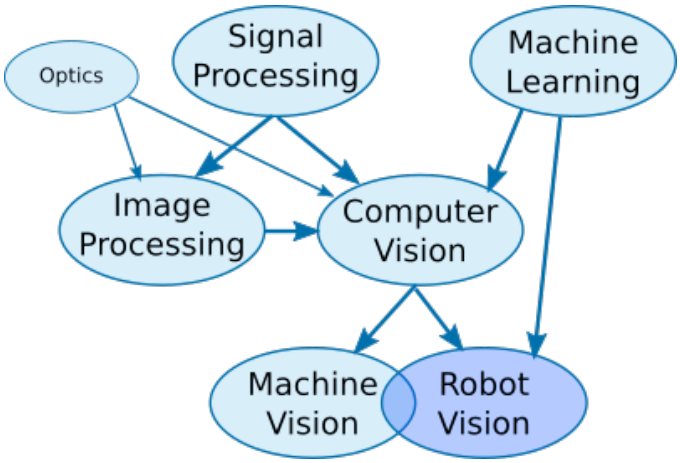

Fig. 1. High level diagram illustrating relationships between disciplines [5]

Humans rely heavily on visual input to interpret and react to their surroundings. CV can be defined as the science that aims to give a similar, or analogous, capability to a machine or computer [5]. CV is concerned with the automatic extraction, analysis and understanding of useful information from a single image or a sequence of images. CV is a combination of signal processing, optics, image processing and ML disciplines. Recent developments in CV rely significantly on deep learning (DL) methods.

As opposed to the conventional settings of CV, where the data being processed is derived from an image or video, RV also considers a video stream that can be controlled. In this way, the interpretation of the data is often intrinsically linked to the behaviour of the robot. This allows a robot to observe, interpret and then act. The robot can either observe the environment further or perform desired actions e.g. open the door, or look at the scene from another angle.

Table I lists the inputs and outputs as considered by these disciplines.

Having laid this groundwork, the rest of the paper is structured as follows. Section II considers recent developments in CV, and examines the tools therein from DL. Section III then proceeds to explore problems in RV and how these

TABLE I
SUMMARY OF INPUTS AND OUTPUTS OF EACH TECHNIQUE

Technique	Input	Output
Signal Processing	Electrical signals	Electrical signals
Image Processing	Images	Images
Computer Vision	Images	Information/features
Pattern Recognition/ Machine Learning	Information/features	Information
Machine Vision	Images	Information
Robot Vision	Images	Physical Action

problems can be addressed with CV and ML techniques. We then round out this work with a discussion on current and future trends in RV, in section IV.

II. COMPUTER VISION

Computer vision is largely concerned with three problems: image classification (object recognition), image classification with localization, and object detection.

A. Image classification

In image classification, a supervised machine learning algorithm considers an input image and predicts a class label of that image, indicating whether or not that class (often referring to an object of interest) is present in the image.

ImageNet [7] is a large visual database that is commonly used as a performance benchmark for classification problems.

The major breakthrough in performance on this dataset came with a deep learning model for image classification: AlexNet [9]. This was built on convolutional neural networks (CNNs), leveraging the idea of automatically learning useful features. This network dramatically outperformed previous image classification results, which relied on hand-crafted features such as the SIFT model, with an error rate of 15.3% compared to 26.2%.

AlexNet has subsequently been outperformed by other models, such as VGG16 in 2015 [10], which chained multiple convolutional layers with the rectified linear unit (ReLU) activation functions together. Further progress has been made with the introduction of yet deeper networks, such as the inception models [11], [12].

The difficulty with deeper networks for image classification is that the error rate increases as the number of layers increases. To address this problem residual learning was introduced. Residual learning creates a connection between the output of one or multiple convolutional layers and their original input with an identity mapping. The model learns a residual function that keeps most of the information and produces only slight changes. ResNet, for example, is able to train 152 convolutional layers using residual learning. Combining inception models with residual learning allowed the Inception-ResNet model to outperform other models [14]. Combinations of layers and models allowed for even faster training and better performance [15].

Self-supervised learning (SSL) is a reliable learning method that allows robots to adapt to their environment. It enables the robot to generate its own training data combined with prior information to improve its performance. Autonomous learning stem from SSL and involves deep learning methods.

The latest trend in image classification involves Neural Architecture Search (NAS) and automatically learning the right network architecture for a task. It is used as a cell in a Recurrent Neural Network to learn its own architecture using reinforcement learning. For a given range of operations and hyperparameters, multiple sequences are realized to maximize the accuracy as a signal reward for a given dataset. The objective is to learn the best sequence of operations (given a maximal depth) to get an optimized architecture [16]. Following the NASNet model is the Progressive Neural Architecture Search (PNAS) model [17]. This has replaced the Reinforcement Learning with a progressive search. This is when a single function describes all possible structures where each structure is stacked with an operator to another. A learned function decides the importance of learning a structure. The highest ranked structures are selected and stacked together.

B. Image classification with localization

In image classification with localization, the model is required to identify both the predicted class as well as a bounding box around the object in the image to indicate where the single object was found. The PASCAL Visual Object Classes (VOC) challenge was another benchmark in object classification and localisation, organised annually from 2005 to 2012 [18]. Traditionally, a common strategy in object localization problems was to use a sliding window. This method involves moving a window over an image to select a sub-region and classify each image region covered by the window using the object recognition model. These methods incurred a high computational cost, leading to other methods such as the sub-window search [19] and branch-and-bound frameworks for object localization [20]. In [21] the two stage sliding window was used with a more novel approach in [22] moving away from operating on a pixel level search.

Subsequent approaches using convolutional neural networks [23] demonstrates object localization without the need for sliding windows.

C. Object detection

Object detection not only tells you which objects are present in the image, it also outputs bounding boxes indicating where multiple objects are. At the heart of object detection is an object classification/recognition algorithm. To localize the object, we have to select sub-regions of the image and then apply the object recognition to these image sub-regions. The straightforward way to generate smaller sub-regions is with the sliding window method. It is an exhaustive search method. All possible locations with different scales need to be searched over the entire image making this method very computational intensive. These limitations are overcome by region-based search methods. These methods take an image as the input and create bounding boxes around all sub-regions in an image that are most likely to be objects. Then we can

classify these bounding boxes using the object recognition model. The region with the highest probability scores are considered to be the locations of the object. Selective search is one of the most popular region search methods. It is based on computing a hierarchical grouping of similar regions based on color, texture, size and shape compatibility [24]. The object detection regional-base convolutional neural network (R-CNN) uses selective search to find the proposed locations and deep learning to perform object recognition [25]. With a Fast Region-based Convolutional Neural Network (Fast R-CNN) [26], a CNN with multiple convolutional layers is used to take the entire image as input instead of using a CNN for each region proposal (R-CNN), thereby reducing the time required to search each proposed region.

A Region Proposal Network (RPN) directly proposes sub-regions, predict bounding boxes and detect objects. RPN quickly and efficiently scans every location in order to assess whether further processing needs to be carried out in a given region. Faster Region-based Convolutional Neural Network (Faster R-CNN) [27] uses a combination of an R-CNN and RPN. RPN avoids the time used on the selective search method, and it allows faster training and testing while improving performance. Region-based Fully Convolutional Network (R-FCN) [28] does not do any regional proposal or selective search but combines it into the CNN to do object detection together with its location [29]. Neural Architecture Search Net (NASNet) [30] discussed in image classification is used with the Faster R-CNN model seen above for better performance [14] with another combination of faster R-CNN seen in [31].

When it comes to object detection, where different objects in an image are classified with localization information, the use of sliding windows has been replaced with techniques such as the You Only Look Once (YOLO) method [32]. YOLO is faster and much more accurate. YOLO uses a single CNN network for both classification and localising. You Only Look Once (YOLO) model takes an image as input, then it divides it into a grid where each cell of this grid predicts bounding boxes with a confidence score. Where previous models usually contained an object in the predicted bounding box, in YOLO there are usually be a high number of bounding boxes without objects. At the end of the network, the highly-overlapping bounding boxes are merged into single one. YOLO9000 [33] increase performance without impacting real-time application speed.

Where the Non-Maximum Suppression (NMS) method is applied at the end of the network in YOLO dealing with empty bounding boxes. Single-Shot Detector (SSD) [34] uses the Hard Negative Mining (HNM) to deal with this issue. SSD is also an end-to-end CNN like YOLO that uses a single CNN network for both classification and localising.

A few different visual datasets have been made available. The availability of these data sets makes different methods of object classification, object classification and localisation and object detection more comparable. As seen in previous paragraphs the three most common databases are the PASCAL Visual Object Classes (VOC), [18], the Common Objects in COntext (COCO) dataset [35] and the ImageNet dataset [7].

III. Robot vision

Over the past decade RV has emerged as a subject area with its own identity. This is due to the advances in hardware and the capability to process and store large quantities of data. RV allows a robot to process visual data from the environment, as seen figure 2. RV has the added benefit of a video stream that can be controlled, making this setting an interactive one.

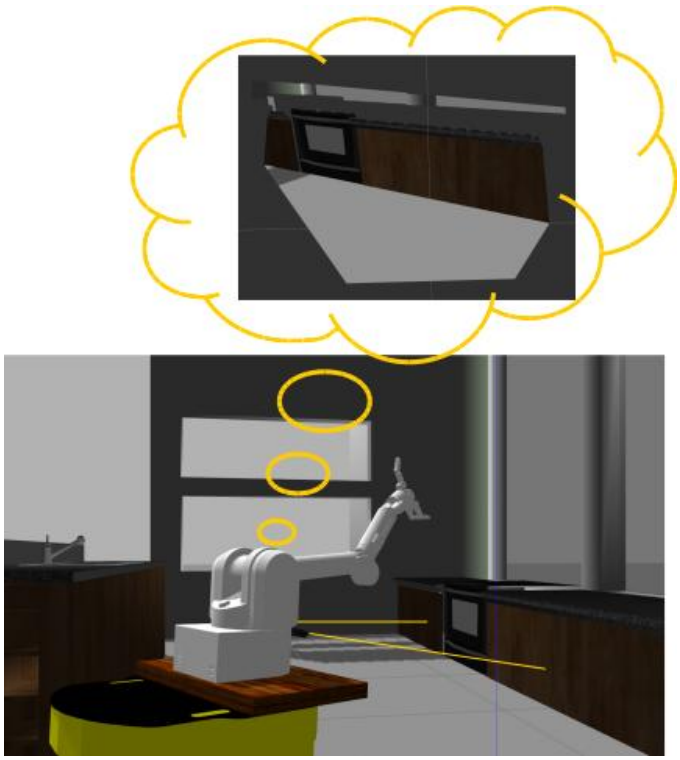

Fig. 2. Robot perceiving its environment using an on-board camera

RV presents a set of unique challenges. Many of these are due to different forms of partial observability in how the robot perceives its environment. These arrive often because the data is collected on-the-fly from a mobile sensor, rather than through a well-curated dataset or a specifically placed camera. The exact position and orientation of the robot or sensor may not be explicitly known. In addition, the sensors would be susceptible to different lighting conditions, or motion of the sensor may induce unexpected blur on the images. The fact that a robot can move around its environment means that many objects may be observed at different scales or orientations to what was seen during training. The location of the robot may also result in objects being partially or completely occluded, and objects of interest may thus not even be visible. This all results in complicating issues such as object recognition or localisation.

RV has to address these issues, and to deal with a variety of conditions and environments through development of robust and generalizable methods.

RV continues to address unique problems; active vision in section III-A, anomaly detection in section III-B, interest detection in section III-C, semantic scene understanding in section III-D, place recognition in section III-E, simultaneous localization and mapping in section III-F, vision based control in section III-G, development of real-time and efficient solutions for relevant application areas in section III-H.

The following sections highlight some of the latest work in RV addressing these challenges.

A. Active vision

Using the ability to manipulate the viewpoint of a camera, which is a very natural advantage of a camera on a robot, the environment can be investigated and better information can be obtained. This is active vision. Active vision was introduced to improve the perceptual quality of tracking results [36]. This method addresses the issues of object occlusion, limited field of view or limited camera resolution. It has also been suggested that visual attention and the selective aspect of active camera control can help in other tasks like learning more robust models of objects and environments with less labeled samples or autonomously. For more information see the survey of active vision applications in robotics [37].

B. Anomaly detection

Anomaly detection is an unsupervised learning task where the goal is to identify abnormal patterns in data. Anomaly detection in the manufacturing industry is commonly used on robots to detect manufacturing flaws [**?**]. Anomaly detection can be extended to images of a different nature, for example to detect anomalous faces using autoencoders [38]. Work on anomaly detection using one-class neural networks [39] is another approach to the problem. Unsupervised and semi-supervised video anomaly detection is summarised in [40] using deep learning methods.

C. Interest detection

When a robot is deployed semi-autonomously it might be useful if interesting information could be summarised and communicated to the operator. This might be something as simple as highlighting something that is out-of-place or unusual. The difficult question is to define what is interesting. This depends on the situation and environment. To ask the robot to send or store only relevant images and information makes more sense than saving a video of the duration of the vacuum operation for the operator to watch. Recent work focused on interest detection in robot vision can be seen in [41].

D. Semantic scene understanding

Semantic scene understanding, in contrast to object recognition, attempts to analyse objects in context with respect to the 3D structure of the scene, its layout, and the spatial, functional, and semantic relationships between objects, [42], [43], [44]. Semantic scene understanding is an important for many robot applications. A self driving car is one example that can benefit from traffic scene understanding [45]. In general, these techniques imbue a robot with the ability to conduct higher-order reasoning in its environment.

E. Place recognition

Slightly different to object detection and recognition is place recognition, which is more concerned with classifying a scene or area in an environment based on a variety of different visual cues. Place recognition has a number of applications ranging from autonomous driving, robot navigation to augmented reality, geo-localizing archival imagery. Vision is the primary sensor for many localization and place recognition algorithms. The difficulty with place recognition is the factors like illumination and season that change the appearance of an image of the same place significantly. There is an overview [46] on place recognition and a visual place recognition survey [47] that provide insight into this topic.

F. Simultaneous Localization and Mapping (SLAM)

Simultaneous Localization and Mapping (SLAM) allows the robot to understand where it is in a map of an environment while also updating the map of this environment. SLAM is a method of simultaneously determining the position of the camera and the structure of the environment in real-time. The work mentioned here is all reliant on camera data, visual SLAM, although earlier research was mostly reliant on laser scanners for SLAM. Visual sensors offer advantages over traditional robotic mapping sensors, including low cost, small size, passive sensing and low power consumption [48]. Vision-based mapping includes FAB-MAP [49], MonoSLAM [50], FrameSLAM [51], V-GPS [52], Mini-SLAM [53], and SeqSLAM methods among others. When SLAM with image data began to give successful results [54], visual SLAM became more widely used. Recent work demonstrates the creation of large scale maps [55], [56]. These maps can be automatically enhanced with meaningful 3D structures [57], and recover shapes [58] all in real time. Place recognition has been used in visual multi-robot simultaneous localisation and mapping [59]. RGBD sensors have also been shown to successfully support SLAM [60].

G. Vision-based control

Vision-based control involves executing local behaviours on a robot, based on perception. An example of this is obstacle avoidance using RGB-D cameras [61], [62], [63].

Robots grasping and interacting with the environment can also be done using RGB-D data. Robot vision allows interacting with the environment in this scenario. In [64] they proposed Generative Grasping Convolutional Neural Network (GG-CNN) to predict the quality and pose of grasps at every pixel. This one-to-one mapping from a depth image overcomes limitations of current deep-learning grasping techniques by avoiding discrete sampling of grasp candidates and long computation times.

H. 3D object detection

In work introduced in the previous sections, the information of the images as well as the 3D structure of the scene plays an important role. There is not a hard line between computer vision and machine learning techniques being implemented on RGB data and on RGB-D data. Steadily, depth information is being incorporated into these techniques that is tested to gain the benefit of more information that can influence decision making and high level functioning. With robot vision the use of depth data can be seen more often. A robot has the ability to perform actions in its environment and requires information about objects around it.

Since robots need to operate in the real world 3D information becomes important for understanding and interacting with the environment. In [66] they propose a holistic approach that exploits 2D segmentation, 3D geometry, as well as contextual relations between scenes and objects. In more recent work [67] a single-stage detector that outputs oriented 3D object estimates decoded from pixel-wise neural network predictions, was proposed for autonomous driving. Implementing YOLO for 3D object detection [68] is another recent application.

IV. Future trends

The sub-fields of robot vision, image classification and object recognition have seen great strides in real time application. This can be largely attributed to deep learning. Convolutional Neural Networks have made training on large data sets possible, enabling exceptionally good classification and recognition of images. These methods were further advanced by transfer learning where a model trained for one task is re-used on another related task. Transfer learning also allows building powerful image classification models using smaller amounts of data.

We have covered the crux of computer vision and robot vision in previous sections. This forms the basis of how to address higher level problems. Methods providing a more holistic understanding of the environment will be the focus for future robotic vision research. This will improve a robot's interpretations of its surroundings.

The following sections highlight the leading developments in machine learning towards these problems. Work focusing on these methods have been used in recent publications and combinations of methods to address RV problems can be seen.

A. Deep learning

Deep networks have demonstrated their ability to learn from images but there is still much work to be done in understanding aspects of the learning dynamics and training mechanisms to enable its greater development and use. The work on DL will only continue to grow with one of the emerging research areas focusing on the inner workings of these methods. This will further enable the deployment of these methods into applications where there is a requirement to understand the decisions made by the robot.

B. Generative models

Generative models are a class of unsupervised learning models that use data to train a model to generate more data matching the original distribution. These models aim to learn the underlying probability distribution of the training data so that it could easily sample new data from that learned distribution. The most efficient approaches are Variational Autoencoders (VAE) [69] and Generative Adversarial Networks (GAN) [70].

GAN's consists of two neural networks: a generator and a discriminator. During training, the generator tries to generate realistic samples, while the discriminator needs to determine whether they are fake or real. At the end, the generator is capable of generating data that looks like the real thing. GANs can be used for image to image translation, or improving the quality of low-resolution images.

C. Deep reinforcement learning

The latest work in reinforcement learning (RL) is image captioning [71]. Being able to describe an image in natural language definitely contributes to understanding of the image. DRL is a method that learns by interacting with the environment through observations, actions, and rewards. This method does not require labelled data and use less data than other methods.

D. Lean and augmented data learning

Since most methods require a lot of data creating data or using transfer learning, using a model trained for one task or domain for another application, is an important.

E. Probabilistic graphical models

In [72] they propose a hierarchical generative model that classifies the overall scene, recognizes and segments each object component, as well as annotates the image with a list of tags.

V. Conclusion

Robot vision draws from many other disciplines. The most prominent fields contributing to recent developments in this area are computer vision and machine learning. The combined progress in these fields of study is being implemented in the field of robot vision allowing robots to get closer to being able to interpret and understand complexities of the world as humans do.

Machine learning techniques greatly contribute to how visual sensor data get interpreted by a robot. This enables a robot to interpret and understand complexities of the world better and also enhance interactions between the robot and its surroundings.

The future of robot vision will be creative implementations of machine learning methods to allow robots to represent their environments in ways that facilitate optimally interacting in those settings.

REFERENCES

[1] S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski, and T. Kircher, "Can machines think? interaction and perspective taking with robots investigated via fmri," *PloS one*, vol. 3, no. 7, p. e2597, 2008.

[2] A. J. Hawkins, "Toyota built a robot to help a paralyzed Army vet around the house," https://www.theverge.com/2017/6/30/15900634/toyota-robot-ai-paralyzed-army-veteran, 2017.

[3] S. Miller, J. Van Den Berg, M. Fritz, T. Darrell, K. Goldberg, and P. Abbeel, "A geometric approach to robotic laundry folding," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 249–267, 2012.

[4] T. L. Chen, M. Ciocarlie, S. Cousins, P. M. Grice, K. Hawkins, K. Hsiao, C. C. Kemp, C.-H. King, D. A. Lazewatsky, A. E. Leeper *et al.*, "Robots for humanity: using assistive robotics to empower people with disabilities," *IEEE Robotics & Automation Magazine*, vol. 20, no. 1, pp. 30–39, 2013.

[5] A. Owen-Hill, "Robot Vision vs Computer Vision: What's the Difference?" https://blog.robotiq.com/robot-vision-vs-computer-vision-whats-the-difference, 2016.

[6] T. Ray, "Demystifying Neural Networks, Deep Learning, Machine Learning, and Artificial Intelligence," https://www.stoodnt.com/blog/ann-neural-networks-deep-learning-machine-learning-artificial-intelligence-differences/, 2018.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[8] Y. LeCun, L. eon Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *PROC. OF THE IEEE*, p. 1, 1998.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, vol. 4, 2017, p. 12.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[16] A. Bernstein and E. Burnaev, "Reinforcement learning in computer vision," in *Tenth International Conference on Machine Vision (ICMV 2017)*, vol. 10696. International Society for Optics and Photonics, 2018, p. 106961S.

[17] C. Liu, B. Zoph, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," *arXiv preprint arXiv:1712.00559*, 2017.

[18] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[19] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[20] ——, "Efficient subwindow search: A branch and bound framework for object localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, p. 2129, 2009.

[21] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 237–244.

[22] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 670–677.

[23] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[24] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[26] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[30] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *arXiv preprint arXiv:1707.07012*, vol. 2, no. 6, 2017.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[33] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint*, 2017.

[34] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[36] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *International journal of computer vision*, vol. 1, no. 4, pp. 333–356, 1988.

[37] S. Chen, Y. Li, and N. M. Kwok, "Active vision in robotic systems: A survey of recent developments," *International Journal of Robotics Research*, vol. 30, no. 11, pp. 1343–1377, 2011.

[38] A. Bhattad, J. Rock, and D. Forsyth, "Detecting anomalous faces with'no peeking'autoencoders," *arXiv preprint arXiv:1802.05798*, 2018.

[39] R. Chalapathy, A. K. Menon, and S. Chawla, "Anomaly detection using one-class neural networks," *arXiv preprint arXiv:1802.06360*, 2018.

[40] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[41] M. Burke, "User-driven mobile robot storyboarding: Learning image interest and saliency from pairwise image comparisons," *arXiv preprint arXiv:1706.05850*, 2017.

[42] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, "Optical flow with semantic segmentation and localized layers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3889–3898.

[43] R. Gadde, V. Jampani, R. Marlet, and P. V. Gehler, "Efficient 2d and 3d facade segmentation using auto-context," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1273–1280, 2018.

[44] A. Geiger and C. Wang, "Joint 3d object and layout inference from a single rgb-d image," in *German Conference on Pattern Recognition*. Springer, 2015, pp. 183–195.

[45] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3d traffic scene understanding from movable platforms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 1012–1025, 2014.

[46] C. Zhu, "Place recognition: An overview of vision perspective," *arXiv preprint arXiv:1707.03470*, 2017.

[47] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[48] M. Milford, W. Scheirer, E. Vig, A. Glover, O. Baumann, J. Mattingley, and D. Cox, "Condition-invariant, top-down visual place recognition," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5571–5577.

[49] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[50] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 1052–1067, 2007.

[51] K. Konolige and M. Agrawal, "Frameslam: From bundle adjustment to real-time visual mapping," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1066–1077, 2008.

[52] D. Burschka and G. D. Hager, "V-gps (slam): Vision-based inertial system for mobile robots," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 409–415.

[53] H. Andreasson, T. Duckett, and A. Lilienthal, "Mini-slam: Minimalistic visual slam in large-scale environments based on a new interpretation of image similarity," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 4096–4101.

[54] A. J. Davison, I. D. Reid, N. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1052–1067, 2007.

[55] E. Fernández-Moral, J. G. Jiménez, and V. Arévalo, "Creating metric-topological maps for large-scale monocular slam." in *ICINCO (2)*, 2013, pp. 39–47.

[56] H. A. Daoud, A. Q. M. Sabri, C. K. Loo, and A. M. Mansoor, "Slamm: Visual monocular slam with continuous mapping using multiple maps," *PloS one*, vol. 13, no. 4, p. e0195878, 2018.

[57] O. Haines, J. Martínez-Carranza, and A. Calway, "Visual mapping using learned structural priors," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2227–2232.

[58] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2320–2327.

[59] Z. Li, R. Chellali *et al.*, "Visual place recognition for multi-robots maps merging," in *Safety, Security, and Rescue Robotics (SSRR), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 1–6.

[60] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the rgb-d slam system," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1691–1696.

[61] V. Thapa, S. Capoor, P. Sharma, and A. K. Mondal, "Obstacle avoidance for mobile robot using rgb-d camera," in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2017, pp. 1082–1087.

[62] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 922–928.

[63] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an rgb-d camera," in *Robotics Research*. Springer, 2017, pp. 235–252.

[64] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.

[65] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.

[66] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3d object detection with rgbd cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1417–1424.

[67] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.

[68] M. Simon, S. Milz, K. Amende, and H.-M. Gross, "Complex-yolo: Real-time 3d object detection on point clouds," *arXiv preprint arXiv:1803.06199*, 2018.

[69] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[70] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[71] H. Shi, P. Li, B. Wang, and Z. Wang, "Image captioning based on deep reinforcement learning," in *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service*. ACM, 2018, p. 45.

[72] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2036–2043.