PhD Research Proposal

Vision Based Navigation for Aerial and Ground Vehicles

M.H.G.D. Tissera

188013F

Supervisors:
Dr. Ranga Rodrigo
Dr. Beshan Kulapala

September2018

Department Electronic and Telecommunication Engineering
Faculty of Engineering
UNIVERSITY OF MORATUWA - SRI LANKA

Declaration

I declare that this is my own research proposal and this proposal does not incorporate without acknowledgment any material previously published submitted for a Degree or Diploma in any other university or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Signature: Date:

....

M.H.G.D. Tissera

I have read the proposal and it is in accordance with the approved university proposal outline. I am willing to supervise the research work of the above candidate on the proposed area.

Signature of the Supervisor(s): Date:

....

Dr.Ranga Rodrigo

....

Dr.Beshan Kulapala

Abstract

Autonomous navigation is one of the most interesting problems among computer vision community. Although a considerable research has already been taken place in this area, vision based full autonomy is still not fully addressed and there are a number of issues to be addresses before full autonomy can be deployed in large scale around the globe. In this research we try to overcome the issue of deploying a single model which is dynamically adaptive to the varying driving context without forgetting the experience learned from other similar but slightly different contexts. In addition to that training such a system in real world is a challenging task. This research aims at producing an end-to-end vision based system for navigation based on conditionally generated feature maps which are dynamically adaptive to the context. Also we focus on efficiently training such systems from pre-recorded real world large scale diverse data as well as practically training in real world.

Keywords:

Autonomous Navigation, Computer Vision, Deep Learning, Neural Networks, Reinforcement Learning, Supervised Learning

TABLE OF CONTENTS

Abstract ii

Table of Contents iii

1 Introduction 1

2 Literature Review 5

2.1 Computer Vision and Deep Learning 5

2.1.1 Convolutional Neural Networks 5

2.1.2 Sequential Neural Networks 6

2.1.3 Training a network 7

2.1.4 Learning Algorithms 8

2.2 Vision Based Navigation 9

2.2.1 Individual Tools 9

2.2.2 End-to-end architectures 10

3 Research Objectives 13

4 Research Methodology 14

5 Work Plan and Resource Requirements 16

5.1 Work Plan 16

5.2 Resource Requirement 16

6 Conclusion 17

References 28

Section 1
INTRODUCTION

Autonomous navigation is a fast growing area in both industry and academia. Deep learning is a promising approach to model complex functions which are structured in high dimensional data which is why it is applicable to many domains [1]. With the recent growth of computer vision applications which utilize deep learning autonomous navigation problem has been tried to address in different approaches such as focusing on individual components [2–4] as well as end-to-end architectures [5, 6]. This research tries to address two main problems existing in the literature which is adaptability of learning models to slightly different environments while preserving the ability to generalize to all the data and training such a system from diverse data in real world.

Regarding ground vehicles the level of autonomy there are two basic categories, namely semi-autonomy and full autonomy. Semi-autonomous vehicles are partially involved with decision making where navigation is being done with the help of human driver. Full autonomy involves vehicle being fully responsible for the whole navigation task. Autonomous navigation is being conducted with the help of several types of sensors. These include Image sensors, LIDAR and radar sensors. These sensors are mounted on self-driving cars in multiple numbers and put together to form the whole system which is complex and involves more processing power where human being driving a vehicle seems much simpler task where most of the input being only vision. Human seemingly take vision input to build the corresponding image in their minds which includes particular information relevant to the scene and task on which they try to navigate.

Regarding navigation of ground vehicles we see different road types such as muddy roads, roads under construction, roads with no lane division where vehicles driving in both directions have to share the road, roads with double lanes whose are opposite directions, highways and expressways. In each type of road, when driving a vehicle a human driver looks for different information and navigation is different. For example traveling in a highway is totally different from traveling in an expressway. Also we see different conditions with regard to time of the day and weather.

To cater these different scenarios coming up with a single fixed large model is inefficient. [5, 6] The number of trainable parameters will be huge and training the model with data in different contexts is difficult. although the system might be able to generalize to different contexts in each context the system might not perform as a single smaller model trained specifically to that context. The evolution of neural networks mostly contained collecting large diverse datasets [7, 8] and to learn from those datasets making the models deeper [9]. Although these approaches tried to make the learning models generalized to diverse data these approaches seem inefficient. Training a system in autonomous navigation is also a challenging task. Supervised learning involves learning from pre-annotated labels in order to generalize to the validation set which is coming from a similar or the same distribution where training data came from. Reinforcement learning involves lot of trial and errors which is impractical in driving in real world. Also imposing a reward function is difficult unlike training agents for video games [10]. Because the reward should include whether collision takes place. Training agent in a simulated environment does not address this issue fully [11]. The trial based approach is questionable for training an autonomous driving agent. Supervised learning approach is also not fully suitable since this approach hardly motivates the model to generalize to data from unseen distributions.

We see much sophisticated sensors which are deployed in vehicles and robots for the purpose of navigation. Different kinds of sensors are put together to build complex systems to address the autonomous navigation problem. These systems are complex and need much computational power and energy. But a human driving a vehicle, controlling an aerial vehicle or simply walking is a less complex problem where most of the decisions are made with the help of vision input they receive. It is very important and interesting to understand how humans perform certain tasks so that we can try to mimic those behaviors in AI systems in order to make those systems more intelligent as human.

Humans seemingly build an imaginary picture based on the input which is embedded with relevant information for a particular task(say navigation) where these images are adaptive to the type and condition of the particular task human is working on. It seems that humans take actions not basing on the pure input of their eyes but based on a feature map built in their minds. The most important evidences which highlight this facts are, people missing easily visible objects and human being unable to distinguish dreams from real scenarios. We can argue that even if the vision from eyes is not available, blind people extensively use touch from sticks and sound to build this kind of a picture in their minds. Existing navigation systems try to navigate base on the inputs received from cameras, not something which was generated, conditioned on the input [5, 6, 12].

Our objective is to address the autonomous navigation problem based on conditionally generated dynamically adaptive intermediate feature maps created from pure vision input for low flying aerial vehicles and ground vehicles. In addition to that we intent to use a hybrid approach of existing learning methods to train such a system the system in real world. Also it will be interesting to conduct a study of what will these intermediate feature maps look like. Advantages include Reduce complexity of the main network. Can understand what components play a major role in vision based navigation from the input. So that we can work more towards highlighting these components to make the prediction faster and reduce the complexity of the models.

We carry out a critical literature review with regard to computer vision, deep learning, training deep neural networks and vision based navigation in the following section. In section 3 we state our research objectives. Section 4 includes research methodology to achieve the stated objectives and section 5 includes work plan and resource requirement. Finally section 6 concludes the content of this research proposal.

Section 2

LITERATURE REVIEW

2.1 Computer Vision and Deep Learning

Computer vision has come a long way with the support of deep learning techniques. Deep learning solves the problem of extracting high-level abstract features from raw data [13]. The computer vision community has already deployed deep learning in a wider range thanks to convolutional neural networks such as image classification, image retrieval, object detection, verification, semantic segmentation, autonomous driving, playing games, pose estimation, image captioning, activity recognition etc. [6, 10, 14–21] Autonomous driving can benefit from deep learning compromising different areas together such as object detection, pedestrian detection, localization and mapping, perception control, scene understanding etc. In this section we review the evolution of convolutional neural networks and recurrent neural networks as sequence models.

2.1.1 Convolutional Neural Networks

Conventional neural networks consists of neurons in each layer which are fully connected with the neurons in the previous and subsequent layers. Having this kind of neural network is inefficient in case where the input is pixel values of a raw image. The number of parameters will drastically increase and there will be tons of useless connections between neurons.

Convolutional Neural networks(CNN) [22] are a great solution to learn from images where small filters are employed on an entire image to output feature maps associated with each filter. CNNs are advantageous in terms of weight sharing which reduces the number of connections and parameters in a network and the fact that each filter can look for a particular entity or pattern within the entire image which adds ability to translate about weights learned from one area of image to other areas. Convolutional neural networks fastened the usage of deep learning in almost all the areas among computer vision community. LeNet [14] was among the first attempts to use CNN where the application was hand digit recognition. AlexNet [15] was the first CNN to win the famous ImageNet Challenge(ILSVRC) [23] in 2012 which is a large scale visual recognition challenge for image classification. Since then all subsequent years have been won by CNN based networks which improves the CNN based architectures and made the networks deeper and deeper. [9, 24–27].

Architectures won image-net evolution show that in order to achieve a small improvement in error models get deeper and more complex. For example GoogLenet(2014) [26] with 22 layers had an error rate of 6.7% on ImageNet while the ResNet(2015) [9] improves that error to 3.57% with a total of 152 layers. Since making networks deeper and deeper is not the optimum solution for making neural networks more intelligent and adaptive, subsequent researches tried to either harvest more information within a layer such as pose and orientation [28, 29] or make the neural network adjustable in some manner, making fixed parameters also learnable and weights adjustable [27, 30]. Including the attention component in neural networks is also an interesting finding where networks are trained to pay attention to particular parts of the inputs [31, 32].

2.1.2 Sequential Neural Networks

CNNs are great for harvesting information from images i.e spatial information but lacks the ability to extract temporal information. Recurrent Neural Networks are an extension to conventional neural networks which is able to handle sequential data with variable length [33]. However Bengio et al(1994) [34] observed that training a conventional RNN is challenging because RNNs are subjective to inability of capturing long term dependencies [34] and the gradient vanishing problem. One solution was to device better learning algorithms than gradient descent [35,36]. The other approach of designing better activation functions resulted Long Short Term Memory(LSTM) [37] and Gated Recurrent Unit(GRU) [38]. These were able to capture long term dependencies in sequential data which enhanced the usage of RNNs.

Sequence models have been widely adopted covering number of areas such as language modeling [39], character recognition and generation [31,40], sequence to sequence translation [41], speech recognition [42], image captioning [43,44] and video analysis [45].

2.1.3 Training a network

Training a deep neural network is a challenging task merely due to the large number of parameters. Apart from the learning algorithm used in a deep neural network, adjusting the weights of connections in a network is also an important task. Despite of the learning algorithm most widely used approach to train the weights of a neural network is back-propagation introduced by Rumelhart et al [46]. Back-propagation repeatedly adjusts the weights of a neural network in order to minimize the distance between the actual output and the expected output of the network. Several improvements have been proposed to improve the training of large neural networks trying to address the problem of overfitting such as dropout [47], regularization [48], data augmentation and early stopping. Gradient vanishing is the problem where the weight updates become very small in a neural network mainly due to the size of the network and several approaches to overcome this issue have been proposed [37] [9]

2.1.4 Learning Algorithms

In literature we observe three main types of learning algorithms to train a neural network which are supervised learning, unsupervised learning and reinforcement learning. Here we intend to discuss in detail about supervised and reinforcement learning as those two are the most commonly used approaches to learning of neural networks associated with computer vision. In supervised learning a model is given a dataset containing inputs and the associated outputs which are namely labels. Models are supposed to learn from these data and generalize to the unseen data in the validation and test sets. Unsupervised learning involves letting the model to learn from the data with no labels itself. Reinforcement learning imposes an objective function on an agent in an environment where the agent is supposed to take the optimal action based on the current state of the environment and possible future states which is then rewarded.

Current applications of deep learning mainly focus on supervised learning, however the use of additional unsupervised learning to facilitate supervised learning is also there [49]. Learning from demonstration [50] is a branch of supervised learning where the autonomous agent is supposed to learn from the human instructor's demonstration. There are navigation architectures which are trained in a supervised way [5, 12]

Reinforcement learning agent is supposed to learn without a teacher from occasional real-valued positive or negative rewards. They should discover and learn how to interact with a dynamic unknown environment in order to maximize expected cumulative future rewards. [49]. The recent development in reinforcement learning influenced in training agents for various tasks such as play video games [10, 51, 52], detection [53] and autonomous driving in simulated environments [54, 55]. It is exciting to see RL agents being able to perform at human level [56] as well as surpassing human level [51, 52].

Since the introduction of generative adversarial networks (GAN) [57] extensive research has already been taken place with regard to GAN based image inpainting. Unlike conventional methods of generative modelling such as Variational Auto Encoders [58], Boltzmann Machines [59, 60], GANs are able to model implicit intractable probability density functions and the training approach includes converging to a nash equilibrium of a game which is more difficult than optimizing an objective function [61]. The deep convolutional generative adversarial networks (DCGAN) introduced in Radford et al. [62] which proposed several improvements to the architectural topology of convolutional GAN which stabilized the training of GANs. This work inspired subsequent research on GANs which involved deep convolutional networks. Conventional GANs are using merely noise as input to the generator and the goal is to learn a distribution. Conditional GANs adds informative conditioning variable in addition to noise which lets the model to learn a conditional distribution. Conditional GANs have been deployed in generating images conditioned on class label [63], conditioned on text [64], conditioned on image [65] etc.

2.2 Vision Based Navigation

Vision based navigation is not a recent topic but a regularly attempted problem among computer vision community. The autonomous navigation problem can be sub-categorized to three problems namely sensing the environment, localizing and mapping, and taking optimum policy to navigate. In the following subsection we review each of these areas in literature as well as end-to-end architectures for vision based autonomous navigation

2.2.1 Individual Tools

Autonomous navigation problem can be categorized into 3 basic areas which are sensing the environment, localization and mapping and optimum control policy.

Sensing the environment and understanding the scene is the first thing an autonomous vehicle should do. Applications of deep neural networks in this area involves pedestrian detection [3, 66], road sign detection [67, 68], road/lane detection [2,4,69–71] and segmentation [17,72,73] Although an extensive research has already been taken place in this area, still autonomous systems are far behind than human level performance of perception learning.

Localizing and mapping answers the question "Where am I?". Visual odometry [74] and Simultaneous Localization and Mapping(SLAM) [75–77] are problem domains which address the requirement of localizing the navigation agent in an environment and map the current position in the global map. Visual SLAM [78] refers to using images as the only external source in order to establish the position of the agent while reconstructing the explored zone.

Based on the intelligent information acquired from the surrounding environment and autonomous agent should come up with an optimal decision on what to do in the current context. The output of the system depends on the task system is supposed to learn which is either a specific part of navigation [79–81] or the full navigation problem [5, 6]

2.2.2 End-to-end architectures

ALVINN [12] was among the first attempts to use a neural network for autonomous driving. ALVINN is a shallow neural network with only one hidden layer, designed for road following based on the raw images taken from a camera as well as a laser range finder. The system was trained using back-propagation [46] and in order to make the system adaptive to varying conditions authors have used a simulated road generator which provided more road images. This simple system was surprisingly performing well and inspired the subsequent research.

Among recent developments Bojarski et al [5] proposed a raw pixel to steering output system with the use of CNNs. This work by NVIDIA was somewhat similar to ALVINN where the raw images were from a single front-facing camera mounted on a vehicle. The algorithms are fairly simple and the cars were able to navigate in simple environments such as highway, following lanes and obstacle free roads. Xu et al [6] proposed an approach to learning a generic model from large scale diverse video data, training an end-to-end architecture for autonomous driving. Authors have deployed a FCN-LSTM architecture and used scene segmentation as a side task. The model addresses autonomous driving as a future ego motion prediction problem and were able to formulate as a generic model. Santana & Hotz [11] also addressed the autonomous navigation problem as a visual prediction task where future video frames are predicted using previous frames. The system was trained in a simulator and the learning approach is a combination of GANs and Variational Auto Encoders [58]. We also can observe end-to-end architectures focused on limited contexts such as highway driving [82],lane keeping [83] and steering [81].

Training these models in real world is challenging. We can observe that this issue has been tried to address in several ways. One is training the system in a simulated environment [11, 54, 55]. But still there exists a huge gap between a modeled environment and real world. Most of the existing end-to-end architectures have been trained in large scale datasets in a supervised manner [5, 6, 12]. The prediction based driving model proposed by Santana & Hotz [11] is a combination of generative models which shows more realistic in terms of training in real world.

All the above discussed approaches do not fully solve the problem of an adaptive single architecture for autonomous driving. Researchers have tried to gather large scale diverse visual data either by fully recording [6] [8] or generating varying samples from the original distribution. Trying to train a generic model with fixed parameters is inefficient because a much simpler model trained only in one context might outperform the large generic model in that particular context. The intelligence of switching between similar models or varying the content of the model is not yet fully addressed in the literature. Also training such a system in real world is still a challenging task where supervised learning involves lot of data and making models memorize while reinforcement learning involves trial based approach which is impractical for an autonomous driving agent. Training an autonomous agent in a simulated environments [11,54,55] is hardly a practical solution since there exists a considerable gap between real world and simulated environments. In fact modeling the varying contexts in real world navigation in a simulated environment is itself an impossible task.

Section 3

RESEARCH OBJECTIVES

The main objective of this research is to develop an end-to-end pure vision based system for autonomous navigation based on conditionally generated contextualized feature maps which are adaptive to the environment and to efficiently train such system from real world diverse data.

Objectives are:

1. To develop a pure vision based end-to-end architecture for autonomous navigation
2. To develop novel network architectures to generate intermediate contextualized feature maps conditioned on the input received from cameras
3. To develop models which are dynamically adaptive to the environment of input while not forgetting what has already been learned from other environments
4. To develop models which extract both spatial and temporal information from input videos to output more realistic commands according to the particular scenario.
5. To develop efficient and promising approaches to train such models from large scale diverse data recorded from real world navigation as well as to train in real time.

Section 4

RESEARCH METHODOLOGY

1. Critically review the literature with regards to following topics in order to identify and justify the problem statement.

    - Computer Vision and Deep Learning
    - Training deep neural networks
    - Vision Based Navigation

2. Develop individual tools to generate contextualized feature maps from input images where the models are dynamically adaptive to the nature of input. Initially these models will be trained with pixel wise human annotated outputs such as road/lane detection, drivable area detection etc. with the help of supervised learning

3. Extend the system to extract temporal information using Recurrent Neural Networks(RNN). Input frames are fed to the models developed in the previous stage which outputs feature maps to be fed in to RNN in each time step.

4. Handover to the model, the responsibility of training feature map extractors from merely input frame sequence and corresponding expected navigation command, say steering direction and desired velocity. Remove pixel-wise annotated expected output maps where the only output system can observe is the navigation commands. Models will be trained with pre-recorded videos along with navigation commands as output in a supervised manner.

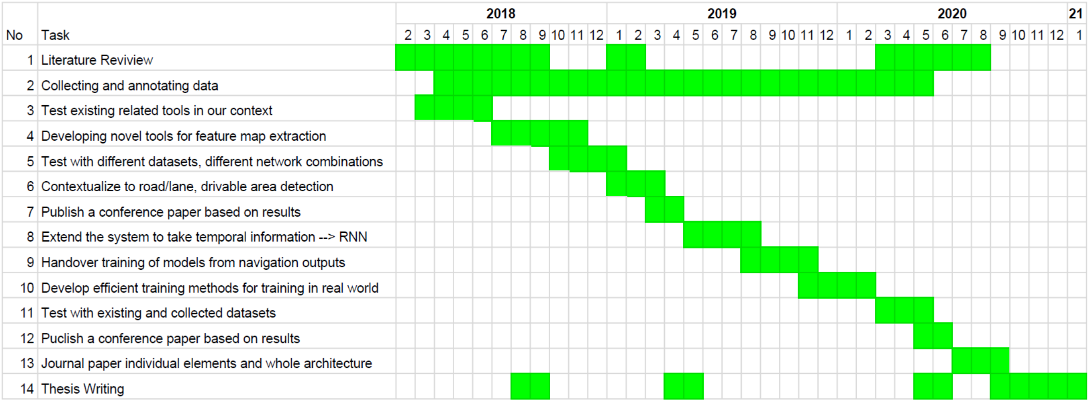
5. Study the conditionally generated feature maps in order to identify the features triggered by the input to make sure relevant details are extracted

from the content of input to the model. Make necessary changes to the models by adjusting the network architectures to make sure that intended objectives are achieved by the networks

6. Impose reinforcement learning and adversarial training on top of existing supervised learning in order to learn from real world navigation in real time.

Section 5
WORK PLAN AND RESOURCE REQUIREMENTS

5.1 Work Plan



		2018											2019												2020												21
No	Task	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1
1	Literature Reviview	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩				🟩	🟩													🟩	🟩	🟩	🟩	🟩	🟩					
2	Collecting and annotating data			🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩								
3	Test existing related tools in our context		🟩	🟩	🟩	🟩																															
4	Developing novel tools for feature map extraction						🟩	🟩	🟩	🟩	🟩																										
5	Test with different datasets, different network combinations									🟩	🟩	🟩	🟩																								
6	Contextualize to road/lane, drivable area detection												🟩	🟩	🟩																						
7	Publish a conference paper based on results														🟩	🟩																					
8	Extend the system to take temporal information --> RNN																🟩	🟩	🟩	🟩																	
9	Handover training of models from navigation outputs																			🟩	🟩	🟩	🟩														
10	Develop efficient training methods for training in real world																						🟩	🟩	🟩	🟩											
11	Test with existing and collected datasets																										🟩	🟩	🟩								
12	Puclish a conference paper based on results																												🟩	🟩							
13	Journal paper individual elements and whole architecture																														🟩	🟩	🟩				
14	Thesis Writing							🟩	🟩							🟩	🟩												🟩	🟩			🟩	🟩	🟩	🟩	🟩

Figure 5.1: Work Plan

5.2 Resource Requirement

1. GPU server
2. High performance single board computers
3. Large scale dataset for ground vehicles in Sri Lankan context
4. Large scale dataset for low flying aerial vehicles following roads

Section 6
CONCLUSION

The intention of this research is twofold. Firstly overcoming the issue of deploying a single adaptive model to similar but slightly different navigation environments which learns from varying environments under varying conditions without affecting its existing knowledge and experience obtained from other environments. Having a single fixed deeper model is inefficient and even though the model after learned from diverse dataset will be able to generalize to all contexts might not producing best results in each context where a smaller model trained from that particular context only might perform better. Secondly training such system from real world diverse data as well as training in real time.

We highlighted that the above mentioned needs are not yet fully addressed in the literature where existing researches either try to make pre-recorded dataset larger, or make the neural network deeper. Existing approaches to training does not fully solve the problem of training a navigation system in real world. We believe that training an agent in a simulated environment is not the solution since there is a considerable gap between real world and a simulated environment.

The research objectives include building a pure vision based end-to-end architecture for autonomous navigation which is based on conditionally generated contextualized feature maps which are dynamically adaptive to the environment and taking a promising approach to train such system in real world large scale diverse data. The models we develop are required to learn from slightly different environments while maintaining the experience learned from previous environments. Our research is inspired by the biological behavior of these particular tasks.

To achieve these objectives a comprehensive research methodology was proposed which starts from a critical review of literature and gradually building required tools towards an end-to-end architecture which extracts both spatial and temporal information from the input in order to output navigation commands. Finally these individual tools will be assembled together in order to build the final system. Specific attention will be given to train the system from real world data taking a hybrid approach of existing learning methods.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] H. Kong, J.-Y. Audibert, and J. Ponce, "General road detection from a single image," *IEEE Transactions on Image Processing*, vol. 19, no. 8, pp. 2211–2220, 2010.

[3] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on.* IEEE, 1997, pp. 193–199.

[4] M. Bertozzi and A. Broggi, "Gold: A parallel real-time stereo vision system for generic obstacle and lane detection," *IEEE transactions on image processing*, vol. 7, no. 1, pp. 62–81, 1998.

[5] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[6] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," *arXiv preprint arXiv:1612.01079*, 2016.

[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] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, 2018.

[9] 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.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[11] E. Santana and G. Hotz, "Learning a driving simulator," *CoRR*, vol. abs/1608.01230, 2016. [Online]. Available: http://arxiv.org/abs/1608.01230

[12] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Advances in neural information processing systems*, 1989, pp. 305–313.

[13] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning.* MIT press Cambridge, 2016, vol. 1.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[15] 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.

[16] S. Ren, K. , 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.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[18] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[19] V. Mnih and G. E. Hinton, “Learning to detect roads in high-resolution aerial images,” in *European Conference on Computer Vision.* Springer, 2010, pp. 210–223.

[20] A. Toshev and C. Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[21] P. Sermanet and Y. LeCun, “Traffic sign recognition with multi-scale convolutional networks,” in *Neural Networks (IJCNN), The 2011 International Joint Conference on.* IEEE, 2011, pp. 2809–2813.

[22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[24] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision.* Springer, 2014, pp. 818–833.

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

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

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

[28] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Advances in Neural Information Processing Systems*, 2017, pp. 3856–3866.

[29] G. E. Hinton, S. Sabour, and N. Frosst, “Matrix capsules with EM routing,” in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=HJWLfGWRb

[30] D. Ha, A. M. Dai, and Q. V. Le, “Hypernetworks,” *CoRR*, vol. abs/1609.09106, 2016. [Online]. Available: http://arxiv.org/abs/1609.09106

[31] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra, “DRAW: A recurrent neural network for image generation,” *CoRR*, vol. abs/1502.04623, 2015. [Online]. Available: http://arxiv.org/abs/1502.04623

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, 2015, pp. 2048–2057.

[33] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.

[34] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[35] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, “Advances in optimizing recurrent networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8624–8628.

[36] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Citeseer, 2011, pp. 1033–1040.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[39] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[40] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[41] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[42] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[44] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[45] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional

networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[46] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, p. 533, 1986.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Advances in neural information processing systems*, 1992, pp. 950–957.

[49] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[50] A. Stafylopatis and K. Blekas, "Autonomous vehicle navigation using evolutionary reinforcement learning," *European Journal of Operational Research*, vol. 108, no. 2, pp. 306–318, 1998.

[51] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[52] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[53] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[54] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 1.

[55] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[56] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[57] 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.

[58] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[59] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning.* ACM, 2007, pp. 791–798.

[60] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, D. van Dyk and M. Welling, Eds., vol. 5, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009, pp. 448–455.

[61] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[62] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[63] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[64] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[65] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[66] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[67] S. Maldonado-Bascón, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gómez-Moreno, and F. López-Ferreras, "Road-sign detection and recognition based on support vector machines," *IEEE transactions on intelligent transportation systems*, vol. 8, no. 2, pp. 264–278, 2007.

[68] C.-Y. Fang, S.-W. Chen, and C.-S. Fuh, "Road-sign detection and tracking," *IEEE transactions on vehicular technology*, vol. 52, no. 5, pp. 1329–1341, 2003.

[69] Y. He, H. Wang, and B. Zhang, "Color-based road detection in urban traffic scenes," *IEEE Transactions on intelligent transportation systems*, vol. 5, no. 4, pp. 309–318, 2004.

[70] Y. Wang, E. K. Teoh, and D. Shen, "Lane detection and tracking using b-snake," *Image and Vision computing*, vol. 22, no. 4, pp. 269–280, 2004.

[71] J. Kim and C. Park, "End-to-end ego lane estimation based on sequential transfer learning for self-driving cars," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1194–1202.

[72] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[73] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[74] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 1. Ieee, 2004, pp. I–I.

[75] S. Thrun, "Simultaneous localization and mapping," in *Robotics and cognitive approaches to spatial mapping.* Springer, 2007, pp. 13–41.

[76] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[77] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (slam): Part ii," *IEEE Robotics & Automation Magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[78] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[79] D. Wang and F. Qi, "Trajectory planning for a four-wheel-steering vehicle," in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, vol. 4. IEEE, 2001, pp. 3320–3325.

[80] V. Rausch, A. Hansen, E. Solowjow, C. Liu, E. Kreuzer, and J. K. Hedrick, "Learning a deep neural net policy for end-to-end control of autonomous vehicles," in *American Control Conference (ACC), 2017.* IEEE, 2017, pp. 4914–4919.

[81] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, and U. Muller, "Explaining how a deep neural network trained

with end-to-end learning steers a car," *arXiv preprint arXiv:1704.07911*, 2017.

[82] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue *et al.*, "An empirical evaluation of deep learning on highway driving," *arXiv preprint arXiv:1504.01716*, 2015.

[83] Z. Chen and X. Huang, "End-to-end learning for lane keeping of self-driving cars," in *Intelligent Vehicles Symposium (IV), 2017 IEEE.* IEEE, 2017, pp. 1856–1860.