Applications of Computer Vision for the Improvement of Autonomous Vehicle Design

by

Christian Kauten

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 7, 2021

Keywords: Automated Driving Systems, Computer Vision, Image Restoration

Approved by

Xiao Qin, Chair, Alumni Professor of Computer Science and Software Engineering
Ashish Gupta, Co-chair, Professor of Analytics
Han Li, Associate Professor of Management Science and Information Systems
Gerry Dozier, Charles D. McCrary Eminent Chair Professor of Computer Science and Software Engineering
Anh Nguyen, Assistant Professor of Computer Science and Software Engineering

Abstract

This dissertation approaches methods of improving autonomous vehicle design through two separate lenses. In the first study, we investigate how technological transparency can improve driver trust in artificial intelligence and ultimately encourage the adoption of automated driving systems. Automated driving systems provide a means of reducing the inherent danger of operating a personal motor vehicle. However, barriers to adoption exist due to low trust in the artificial intelligence that powers the systems. To fill this deficit of trust, Chapter 3 proposes a deep learning-based visual alert system that allows passengers to monitor the artificial intelligence performance in real-time. Using a trained object detection model, we design a novel perception augmentation system for conveying information about the driving scene to the passenger through the lens of artificial intelligence. We conduct an empirical study that confirms that the proposed system improves the trust in the underlying artificial intelligence technology. Trust in artificial intelligence is also found to not only positively affect the perceived benefit from– and intention to use an automated driving system, but also negatively influence the perceived risk associated with using the technology. Perceived enjoyment from the autonomous vehicle is also found to have a strong effect on the perceived benefit from– and intention to use the system.

In the second study in this dissertation, we take a close look at methods to improve the quality of sensor data in automated driving systems. Although deep learning methods continue to set new state-of-the-art metrics on deblurring benchmarks, a comprehensive understanding of what losses are effective for the deblurring task is missing from the literature. The study in Chapter 4 provides an empirical foundation for the selection of a loss function when developing image deblurring models. Despite the popularity of mean squared error as a content loss function for image restoration tasks, we demonstrate that mean absolute error produces higher

quality results to the human visual system. Furthermore, we show that deblurring models trained solely using a perceptual content loss produce outputs that are perceptibly better than the same model trained using a plain mean absolute error or mean squared error loss despite validation metrics that would indicate otherwise. Finally, we demonstrate that adversarial losses do not produce generators capable of confidently deblurring images in the absence of auxiliary loss functions; however, the combination of adversarial and content losses in some cases produces higher quality results than either constituent loss when trained in isolation. Compared to state-of-the-art methods, the best model developed in this work produces worse quantitative validation metrics, but visibly better results on real-world blurs in natural images.

Acknowledgments

The completion of this dissertation would not have been possible without an immense amount of external guidance and help from advisors, co-workers, family, and friends.

I would first like to thank my advisor, Dr. Xiao Qin, without whom I may not have embarked on the Ph.D. journey to begin with. Dr. Qin's deep knowledge of computer science and keen skill as a professor was a source of inspiration during my time studying at Auburn.

I am also ineffably grateful for the work of my advisor, Dr. Ashish Gupta, whose guidance and expertise were invaluable during the numerous research projects on which we collaborated. Dr. Gupta worked tirelessly to ensure that I had ample time to devote to research. He also helped me to engage with other labs and contracting agencies on practical real-world problems, such as the first project in this dissertation.

I would like to thank my committee members, Dr. Anh Nguyen and Dr. Gerry Dozier, who reviewed my proposal and dissertation. The guidance they provided allowed me to reorient my second project to solve a higher impact problem using more novel methods. I also have immense gratitude for Dr. Dave Bevly and the GAVLAB, who provided the autonomous vehicle and expertise necessary to embark on my first project. Without Dr. Han Li, I would not have been able to design an effective experiment for my first project; I am eternally grateful for her contribution to my work. I would also like to thank Dr. Stan Reeves, whose expertise in digital signal and image processing was invaluable to the success of my second project.

I am thankful for the funding I received towards my research projects from the Waltosz Fellowship, directed by Mr. Walter Waltosz. The contribution of Mr. Waltosz provided me the hardware that I required for the projects in this dissertation.

I would like to thank my collaborators in the research group, namely, Chaowei Zhang, Xiaopu Peng, Alison Jenkins, Thomas Heckwolf, and Jianzhou Mao. I am thankful for the

Department of Computer Science and Software Engineering, the Department of Systems & Technology, and the Department of Electrical Engineering for providing a wonderful environment for study and research, and promoting interdepartmental collaboration.

Finally, I'd like to thank my friends and family. My mom, Deanna Kauten, and my dad, Jim Kauten, were endless sources of support during my struggles as a graduate student and a human. I am also thankful for my friends, who tolerated my excitement for esoteric computer science topics and helped me to understand things from different perspectives. In particular, I'd like to thank Bryce Langston for proofreading much of this report.

Table of Contents

Abstract ii
Acknowledgments iv
List of Figures ix
List of Tables xii
List of Abbreviations xiii
1 Introduction 1
1.1 Motivation for studying trust in autonomous vehicles 2
1.2 Motivation for comparing loss functions in deep image restoration 5
1.3 Dissertation Organization 7
2 Literature Review 9
2.1 Automated Driving Systems, Artificial Intelligence, and Trust 9
2.1.1 Trust in Artificial Intelligence 11
2.1.2 Artificial Intelligence and Human Interfaces 11
2.2 Image Degradation and Restoration 13
2.2.1 Image Acquisition Model 13
2.2.2 Image Degradation Models 14
2.2.3 Classical Image Restoration Approaches 15
2.2.4 Non-Uniform Image Degradation Dataset Synthesis 16
2.2.5 Metrics 17
2.2.6 Generative Adversarial Networks 17
2.2.7 Training Deep Networks 25
2.2.8 State-of-the-art Image Restoration Approaches 27
3 Does Trust Influence Autonomous Vehicle Adoption? 34

3.1 Perception Augmentation Module Design . . . 35
3.1.1 Framework . . . 35
3.1.2 Perception Augmentation Module . . . 36
3.1.3 Roof-Mounted Camera . . . 39
3.1.4 Vision Model . . . 39
3.1.5 Object Color Map . . . 43
3.1.6 Dashboard Display . . . 44
3.1.7 Windshield LED Strip . . . 45
3.2 Psychological Study . . . 47
3.2.1 Theoretical Development . . . 48
3.2.2 Extending the Social Contract Model of Health IT to AV Adoption . . . 49
3.2.3 Research Model and Hypotheses . . . 50
3.2.4 Research Methodology . . . 54
3.2.5 Data Analysis and Findings . . . 60
3.3 Discussion . . . 65
3.3.1 Contributions to Theory and Research . . . 68
3.3.2 Implications for Practice . . . 70
3.3.3 Limitations and Implications for Future Work . . . 71
4 Choosing a Loss Function for Deep Image Deblurring . . . 73
4.1 Methodology . . . 74
4.1.1 Architecture . . . 74
4.1.2 Loss Functions . . . 78
4.1.3 Training . . . 82
4.1.4 Validation . . . 84
4.2 Results . . . 84
4.2.1 Adversarial Losses . . . 84
4.2.2 Content Losses . . . 88

4.2.3 Combined Losses . . . 93
4.2.4 Comparisons to State-of-the-art . . . 96
4.3 Discussion . . . 102
4.3.1 Contributions to Theory and Research . . . 104
4.3.2 Implications for Practice . . . 105
4.3.3 Limitations and Implications for Future Work . . . 106
5 Conclusion . . . 108
5.1 The influence of trust on autonomous vehicle adoption . . . 108
5.2 The choice of loss function for deep image restoration . . . 109
5.3 Limitations and Future Work . . . 109
Bibliography . . . 111

List of Figures

3.1 A framework for understanding Human-Computer Interaction (HCI) in Automated Driving Systems (ADS). 36

3.2 The high-level architecture and data flow of the proposed perception augmentation module. 38

3.3 The layout of the roof-mounted and cockpit cameras installed on the vehicle. 39

3.4 A hypothetical YOLO output grid with a bounding box output for a given prior. . . . 43

3.5 The color-coding for the generalized object classes. 44

3.6 An example object detection output. 45

3.7 A windshield LED strip divided into seven zones. 46

3.8 The research model describing the constructs and hypothesized paths. 50

3.9 The design of the perception augmentation module simulator. 55

3.10 The survey scenario that participants read before embarking on the study. Here we use the less technical term “Self Driving Car (SDC)” to refer to an ADS or AV for ease of communication to a non-technical audience. 56

3.11 Results of testing hypotheses using Partial Least Squares (PLS) analysis. 65

4.1 A depiction of a simple residual block. Two convolutional layers, shown in gray, are applied to the inputs, shown in light gray, and added back to the inputs before the final activation function f. 75

4.2 The convolutional generator network. The inputs and output of the model are RGB images of size $(M, N, 3)$. Gray blocks denote convolutional layer activation maps after a Leaky ReLU activation function and blue blocks indicate residual sub-network outputs (see Figure 4.1). 77

4.3 The convolutional discriminator network. The inputs to the model are RGB images of size $(M, N, 3)$. Gray blocks denote convolutional layer activation maps after a Leaky ReLU activation function. The pink block describes the outputs of a pyramid pooling layer. The flattened outputs of the pyramid pooling layer pass to a dense network with a single layer and single output unit. 78

4.4 Example average activation outputs from the `block1_conv1`, `block2_conv2`, and `block3_conv3` layers of the VGG-19 network for a sharp-blurry image pair. 82

4.5 Average PSNR and SSIM per metric during the training procedure for models trained with adversarial losses. 85

4.6 Examples restorations of "`face2`" from the dataset of Lai et al. (2016) based on generators trained with different adversarial loss functions. 87

4.7 Average PSNR and SSIM per metric during the training procedure for models trained with content losses. 89

4.8 Examples restorations of "`face2`" from the dataset of Lai et al. (2016) based on generators trained with different content loss functions. 92

4.9 Examples restorations of "`face2`" from the dataset of Lai et al. (2016) based on generators trained with different combinations of adversarial and content losses. . . 95

4.10 Examples restorations of "`385/11_01_003028`" from the dataset of Nah et al. (2017). Pre-trained models are used to evaluate existing methods. 98

4.11 Examples restorations of “face2” from the dataset of Lai et al. (2016). Pre-trained models are used to evaluate existing methods. 99

4.12 Examples restorations of “text10” from the dataset of Lai et al. (2016). Pre-trained models are used to evaluate existing methods. 101

List of Tables

2.1 A comparison of state-of-the-art deep learning-based image restoration models. A mark of "–" indicates that a value was not specified by the authors. 31

2.2 A comparison of training and validation parameters used by state-of-the-art deep learning-based image restoration models. A mark of "–" indicates that a value was not specified by the authors. 33

3.1 The demographic distribution of survey respondents from two pilot studies. 57

3.2 The demographic distribution of survey respondents in the final study. 58

3.3 The survey instrument for collecting data from the participants of the study. 59

3.4 Composite Reliability (CR), Average Variance Extracted (AVE), and loadings and cross-loadings of reflective scales. 62

3.5 Discriminant validity of reflective measurement scales. 63

3.6 Path coefficients of alternative testing models. 66

4.1 Adversarial loss functions used to train models in this study. 80

4.2 Training parameters that are held constant in this study based on associated loss configurations. 83

4.3 PSNR and SSIM metrics on the GoPro and REDS test benchmarks based on generators trained with different adversarial loss functions. The best values are shown in bold and the second-best values are underlined. 86

4.4 PSNR and SSIM metrics on the GoPro and REDS test benchmarks based on generators trained with different content loss functions. The best values are shown in bold and the second-best values are underlined. 90

4.5 PSNR and SSIM metrics on the GoPro and REDS test benchmarks based on generators trained with different combinations of adversarial and content loss functions. The best values are shown in bold and the second-best values are underlined. . . . 94

4.6 PSNR and SSIM metrics on the GoPro test benchmark. Metrics for previous works were derived from the papers. 96

List of Abbreviations

AC-GAN	Auxiliary Classifier GAN
AI	Artificial Intelligence
ADS	Automated Driving System
AEB	Automatic Emergency Braking
AV	Autonomous Vehicle
AVE	Average Variance Extracted
BAM	Blur-Aware Module
BANet	Blur-Aware Network
CCD	Charge-Coupled Device
CGAN	Conditional GAN
CMV	Common Method Variance
COCO	Common Objects in Context
CMOS	Complementary Metal-Oxide-Semiconductor
CNN	Convolutional Neural Network
CRF	Camera Response Function
DCT	Discrete Cosine Transform
DDM	Dense Deformable Module
DDT	Dynamic Driving Task
DFS	Discrete Fourier Series

DFT Discrete Fourier Transform

DMPHN Deep Multi-Patch Hierarchical Network

DVD Deep Video Deblurring

FCWS Forward Collision Warning System

FLOP Floating Point Operation

GAN Generative Adversarial Network

GNSS Global Navigation Satellite System

HCI Human-Computer Interaction

HIDE Human-aware Image DEblurring

HIN Half-Instance Normalization

HINet Half-Instance Normalization Network

HVS Human Visual System

INT Intention

IoU Intersection over Union

JOY Perceived Joy

JPEG Joint Photographic Experts Group

KL KullbackâĂŞLeibler

Leaky ReLU Leaky REctified Linear Unit

LED Light Emitting Diode

LSGAN Least-Squares GAN

MAE Mean Absolute Error

MPRNet Multi-Stage Progressive Image Restoration Network

MSE Mean Squared Error

NFS	Need For Speed
NHTSA	National Highway Traffic and Safety Administration
PFR	Performance Risk
PI	Personal Innovativeness
PSNR	Peak Signal-to-Noise Ratio
PSR	Psychological Risk
RADNet	Region-Adaptive Deblurring Network
RaGAN	Relativistic average GAN
REDS	REalistic and Dynamic Scenes
ReLU	REctified Linear Unit
RGAN	Relativistic GAN
RGB	Red-Green-Blue
SAPHNet	Spatially-Attentive Patch-Hierarchical Network
SCR	Social Risk
SCT	Social Contract Theory
SFR	Safety Risk
SRN	Scale-Recurrent Network
SSIM	Structural Similarity Index Measure
TAI	Trust in AI
UI	User Interface
VGG	Visual Geometry Group
VOC	Visual Object Classes
WGAN	Wasserstein GAN

WGAN-GP WGAN Gradient Penalty

XAI Explainable Artificial Intelligence

YOLO “You Only Look Once”

Chapter 1

Introduction

This dissertation studies two separate but related methods to improve the design of Autonomous Vehicles (AVs). In the first part of this dissertation, we approach the problem from the perspective of the user. By developing a user interface based on the same Artificial Intelligence (AI) technology that powers the Automated Driving System (ADS), we aim to improve the user's trust in the autonomous control system. Beyond the design of the system, the first part of this study also describes the validation of the proposed system on human subjects through a psychological study. We show that the proposed system improves the user's trust in artificial intelligence and that this trust positively influences the user's experience and intention to adopt and use the ADS. Relative to the effect of enjoyment, we show that trust in AI significantly reduces the perceived risk from the system.

Another way of improving the design of AVs is through improving the perception modules that provide data to the control and presentation layers. This has the effect of not only improving the performance of the ADS but also any other AI features, such as user interfaces, that rely on the perception data. A common degradation that occurs in AV sensor data is motion blur due to the motion of the objects in the scene and the motion of the vehicle itself. Although a large amount of contemporary research surrounds developing models for restoring blurred images, few works provide comprehensive comparisons of how the different selections of model components affect optimization. Because authors frequently pack many innovations into a single published model, it can be challenging to unpack what innovations are truly effective. In particular, a variety of loss functions have been applied in image restoration models on many different generators, but there is no empirical study to explore each loss function in a common training environment.

This dissertation consists of two parts. The first part introduces a novel perception augmentation system and a psychological study to validate its effect in terms of human constructs. The second part describes a comparative study of the different loss functions used in deep learning image restoration models to determine an empirically grounded selection of loss functions. Section 1.1 elaborates the motivation for the first part of the dissertation while Section 1.2 proposes the reasoning behind the second part. Section 1.3 goes on to describe the organization of this dissertation.

1.1 Motivation for studying trust in autonomous vehicles

An AV is a motor vehicle equipped with any number of driving automation features to augment or replace the human driver's abilities during the Dynamic Driving Task (DDT) (Various 2018). These systems embody a combination of sensors, decision-making processors, and hardware controllers to provide these ADS to the human driver. As the features of the AVs evolve and become more complex, so do the underlying ADS. Newer vehicles are moving towards a centralized computer architecture capable of handling the high throughput of data from the sensors (Lin et al. 2018). Centralization of onboard compute allows for synchronized processing of the sensor data and drive-by-wire control systems. It also simplifies the task of interfacing with the human driver through IO devices. These IO devices can be used to explain unexpected behavior of the ADS, augment the driver's perception through the sensors of the vehicle, and adjust decision-making policies to fit the driver's preference.

The adoption of AVs could benefit society in a variety of ways, namely, (1) by reducing the number of road deaths, (2) improving traffic patterns, (3) optimizing freight lines, and (4) enabling drivers to participate in other tasks instead of the DDT that they would otherwise be responsible for (Lutin et al. 2013, Bimbraw 2015). The National Highway Traffic and Safety Administration (NHTSA) reports that human error accounts for 90% of road deaths each year (Rosenzweig and Bartl 2015). In theory, AVs can reduce the number of road deaths by replacing sub-optimal human driving policies and reaction times with quicker and more

informed decision-making, thus preventing wrecks, traffic, and the like. AVs can reduce traffic by preventing wrecks and other holdups, and also by planning optimal routes through traffic. In a similar vein, AVs stand to optimize freight lines by removing the failure-prone elements of human drivers, like drowsiness. Finally, AVs can improve the lives of individual people by freeing up time for other activities, such as daydreaming, reading, and talking, to name a few.

The features that compose an AV can have varying degrees of autonomy, ranging from none, where a human is required for all elements of the DDT, to fully autonomous, where no human interaction is required to perform the DDT. The SAE J3194 standard defines a classification system for characterizing the levels of automation in an ADS (Various 2018). Level 0 refers to purely manual vehicles with no automation features. Level 1 introduces the notion of driver assistance, specifically in the lateral *or* longitudinal axis, but not both. This level describes independent ADS features that are responsible for specific sub-tasks along a singular transport axis (e.g., cruise control). Level 2 introduces the combination of multiple lateral and longitudinal components into a more seamless package. The driver is still expected to remain in the loop to monitor both the driving scene and the performance of the ADS. Level 3 ADS are capable of handling most situations autonomously to the point that the driver no longer needs to attentively monitor the system. However, the driver is expected to maintain situation awareness and perform interventions when the ADS reaches an edge case that it cannot handle. Level 4 is where an ADS becomes fully autonomous in that the system no longer expects human interventions within the operational design domain. Level 5 improves upon Level 4 in that the vehicle can handle situations without condition, even those situations that are outside the operational design domain of the system.

ADS can be decomposed into a conceptual framework with three processing phases, namely, perception, planning, and control (Badue et al. 2020). In the perception phase, the system combines data from sensors into a synchronized view. A variety of sensors, like video cameras, lidar, and radar, to name a few, are used depending on the features of the ADS. The perception phase also contains computer vision logic for processing the sensor data to

understand the salient properties of the driving environment, such as people, cars, and road markings. In the planning phase, route data, such as from GPS, are used in conjunction with the perception data to determine the needed hardware controls for the vehicle to maintain the optimal route. This phase can include complex forward models that track instances of salient objects in the scene, like neighboring vehicles. In the control phase, the planned controller updates are mapped to the control surface of the vehicle in terms of accelerator pedal compression, brake pedal compression, and steering angle.

Current ADS primarily address the DDT without explicitly considering interaction with the driver. As a result, ADS can be opaque to the end-user and behave like a "black box". In cases where the ADS performs unexpectedly, the driver could become confused due to lacking information and in turn, begin to distrust the actions of the ADS. This is a problem because if the human driver fails to sufficiently trust the ADS, they may elect to override the ADS features and resume manual control of the vehicle.

Explaining the results of the ADS is a complicated task that can be approached from a variety of directions (Amershi et al. 2019). One method is for the system to react to driver prompts regarding vehicle decision-making and respond with answers that explain what the vehicle is perceiving and/or why the vehicle may have performed a certain action. An alternative method involves pro-actively relaying data from the perception, planning, and control phases to the driver to keep the driver informed about critical information and/or events during the DDT. This information could include perceptual data, like the names and locations of objects surrounding the vehicle; planning data, such as adjustments to the route based on traffic or weather patterns; or control data, like that the vehicle is coming to an abrupt halt.

The study in Chapter 3 follows the second aforementioned approach to design a perception augmentation module that improves the trust of the driver in the underlying ADS. The perception augmentation module utilizes the outputs of a theoretical perception phase in an ADS to describe the visual surroundings of the AV using a video camera, an object detection model, and a screen. The object detection model is a deep learning algorithm that detects people, cars,

etc. in the video camera data represented as still images. The detected objects are shown to the driver using a screen (e.g., with dimensions $6" \times 8"$) placed on the dashboard of the vehicle. Overlaid on the camera data are color-coded bounding boxes which highlight the objects that are detected in the scene. The detected objects are also mapped to physical space to illuminate a color-coded Light Emitting Diode (LED) strip around the windshield of the vehicle to show roughly where objects are in the real world. An experiment utilizing a software-in-the-loop simulation and survey is designed to test whether the proposed system improves human trust in the ADS and whether improving trust truly impacts the intent to adopt autonomous features. The simulation and survey are administered remotely using the Qualtrics platform to collect data from 517 people.

The results of the study in Chapter 3 confirm a variety of hypotheses in the field of autonomous driving. It is validated that the presented perception augmentation module does increase the driver's trust in the ADS. This increased trust, in turn, has a positive impact on the perceived benefits and a negative impact on the perceived risks of using the ADS. Improving trust is also shown to have a positive impact on the intention to adopt autonomous features, as do the constructs of perceived risks and benefits. As such, the results provide a strong argument for the inclusion of components like the perception augmentation module alongside consumer ADS to improve human trust in– and adoption of AVs.

1.2 Motivation for comparing loss functions in deep image restoration

Although user interfaces that provide insight into the underlying functionality of the system can improve the trust of the driver and encourage adoption of the system (see Chapter 3), these systems are limited by the quality of the sensor data and models on which they operate. Camera sensors on AVs are subject to *motion blur* effects, i.e., due to the fast movement of objects in the scene, that can cause severe degradation in the performance of down-stream models that may not have been trained specifically to handle the dynamics of the motion blurs (Kupyn et al. 2018). As such, prior research has investigated methods of restoring images that have

been degraded by external systems that have predictable patterns, either in the latent data, the degradation systems, or both.

Removing blur from natural images without underlying knowledge of the blur system is a challenging problem in digital image processing. Recently, deep learning methods have proven to be effective deblurring tools that demonstrate success when tested on public benchmarks. Although these networks continue to set new state-of-the-art metrics, there is a lack of principled understanding of what loss functions are truly effective in the learning process. Furthermore, leading state-of-the-art models produce poor results relative to some other methods when tested against real-world datasets, indicating high degrees of over-fitting to the training and validation data. Koh et al. (2021) have investigated the effect of single- versus multi-scale training, but their comparative study neglects elements such as loss functions and normalization settings. Lucic et al. (2018) have studied the effect of Generative Adversarial Network (GAN) loss functions on the standard image synthesis task, but the results of their study do not necessarily map directly to image-to-image translation tasks where the learned transformation is structurally different. Motivated by these limitations in the literature, this work presents a comparative study of recent deep learning innovations to provide an understanding of how different loss functions interact as it relates to image deblurring.

Two major types of loss functions are used in the context of deep image restoration, namely, content losses and adversarial losses. *Content losses* are computed using paired image data to enforce a generator model's output to match the expected images. A separate school of thought applies *adversarial losses* through the lens of an auxiliary discriminator network that is trained to detect sharp versus degraded images. Adversarial losses are notoriously difficult to stabilize during training and as such, authors that utilize GANs for image restoration tasks typically combine the adversarial loss with an auxiliary content loss or perceptual content loss to regularize the network (Nah et al. 2017, Kupyn et al. 2019). Although research continues to achieve state-of-the-art performance on the standard benchmarks for image deblurring (Chen et al. 2021), few works attempt to compare the innovations of new research on a granular

level. In particular, several works have suggested and demonstrated the use of adversarial loss functions and perceptual content loss functions, but no research provides a deep comparison between the results of these different losses.

The study in Chapter 4 provides a large comparative study of various losses that have been effective in prior research. In particular, the study addresses the following research questions.

1. Which content loss functions are the most effective for image deblurring?
2. Without using content losses, do adversarial losses stably converge?
3. How does the combination of content and adversarial losses affect deblurring performance relative to using adversarial loss or content loss in isolation?

The results of Chapter 4 provide empirical evidence that despite the popularity of Mean Squared Error (MSE) as a content loss function for image restoration tasks, Mean Absolute Error (MAE) frequently produces higher quality results. Furthermore, we show that generator models trained solely using a perceptual content loss produce outputs that are perceptibly better than the same model trained using a plain MAE or MSE loss despite validation metrics that would indicate otherwise. We show that adversarial losses do not produce generators capable of confidently deblurring images in the absence of auxiliary loss functions. Likewise, we show that the combination of adversarial and content losses in some cases produces higher quality results than either constituent loss when trained in isolation. Finally, we show examples where the best model in this study produces results that are in some cases perceptibly better than the current state-of-the-art models when tested against real-world blur data. To the best of our knowledge, this is the first work to comprehensively assess the impact of content and adversarial losses on deep learning image deblurring models.

1.3 Dissertation Organization

This dissertation is organized as follows. Chapter 2 provides a comprehensive review of the prior literature. In Chapter 3, a perception augmentation system is presented to measure

how system transparency affects user trust in autonomous vehicles. A psychological survey is described and the results of two trials and one primary study are presented and discussed. Chapter 4 describes a comparative study of losses in deep learning to summarize the effect of different innovations on model performance on deblurring tasks. A model is described that achieves worse metrics than state-of-the-art methods but generalizes better to real-world data. Chapter 5 concludes the dissertation with a summary of the contributions made and notes for future research.

Chapter 2

Literature Review

This chapter first presents a review of the literature surrounding AVs, AI, and trust in Section 2.1. In section 2.2, we go on to provide a comprehensive background of image processing, deep learning, adversarial networks, and state-of-the-art image restoration approaches.

2.1 Automated Driving Systems, Artificial Intelligence, and Trust

ADS are exciting technological advancements that could restructure the transportation engineering profession (Chan 2017, Lutin et al. 2013); however, ADS adoption is limited by various social and regulatory factors. Namely, legal liability, regulation, and public reception of the technology are known to be the primary barriers to adoption (Rosenzweig and Bartl 2015, Kyriakidis et al. 2015). Currently, liability in traffic incidents is based on *how* as opposed to *why* the human driver has failed. As the driving task shifts from the human pilot to the ADS, liability becomes harder to assign because the reason for the driving failure may place liability with the technology manufacturer instead of the driver of the vehicle (Goodall 2016). Furthermore, the opacity of current automatic systems (i.e., that are machine-learned) makes it difficult to understand if and/or why the system has failed. Regulation of ADS presents a challenge due to the difficulty of verifying the technology. Because it's infeasible to perform exhaustive edge case testing (Kalra and Paddock 2016, Schwarting et al. 2018), defining a robust set of laws surrounding ADS is nontrivial. Furthermore, the regulation requires an adept understanding of the moral and ethical implications of the driving task (Fleetwood 2017). Defining how the ADS should perform in mission-critical situations is not only challenging to accomplish in software, but also challenging to confer on as a society. A utilitarian belief is an accepted ethical model for an ADS; however, people tend to disavow this belief when the utilitarian decision results

in harm or misfortune (Bonnefon et al. 2016). From a regulatory adoption perspective, our perception augmentation module increases the transparency of the underlying technology by providing a clear means of describing how the ADS sees the world. For this same reason, our perception augmentation module partially resolves the issue of liability by explaining one-third of the ADS functionality – i.e., the perception module, but not the planning or control modules. The explanation of the planning and control modules is left for future work.

Public opinion is one of the most studied limiting factors to the adoption of ADS. Part of the problem lies in the current levels of autonomy on the market that achieve SAE levels 0–3, but not levels 4–5. It is known that younger audiences are more receptive to levels 4–5 that allow the driver to engage in other tasks without monitoring the vehicle or driving environment (Nees 2016); however, older audiences show resistance towards this level of autonomy and instead prefer features in the SAE level 3 classification (Abraham et al. 2017). Despite the increased interest in higher levels of autonomy, the public shows clear resistance due to unwillingness to pay, financial liability in the event of an incident, and general distrust of the technology. Although various ADS are available in various vehicles in the current market, these features are often part of premium packages or otherwise available only on high-end models. Furthermore, customers show concern about financial liability in cases where the ADS is involved in an automotive incident (Howard and Dai 2014). This financial barrier is a known social barrier to the advancement of ADS (Bansal and Kockelman 2018, Schoettle and Sivak 2014). With an increased number of data leaks and privacy violations in modern times, data collection has also become a major concern. Those worried about data collection and misuse by ADS manufacturers have shown resistance to adopting the technology (Kyriakidis et al. 2015). Additionally, trust in the technology itself has become a known problem. People have been shown to fear transferring control to an autonomous agent and express a clear desire to be able to override the system (König and Neumayr 2017). Furthermore, distrust of the technology tends to raise stress levels while using ADS (Morris et al. 2017). Such additional stress increases the likelihood that a driver will disable the autonomous systems because

the driver views the monitoring task as more work than the driving task it replaces (Koo et al. 2015). Finally, media coverage of ADS-related incidents can harm the public perception of the technology (Shariff et al. 2017). Although accidents related to ADS are rare relative to the number of human-induced accidents, the accidents receive massive media coverage, some of which may sensationalize or misrepresent the event. Our perception augmentation module (see Chapter 3) improves the trust and enjoyment in current and future levels of autonomy without dramatically increasing the cost of the ADS. In this way, our perception augmentation module improves the intent of all age groups to adopt ADS.

2.1.1 Trust in Artificial Intelligence

Because trust is a known barrier to the adoption of higher levels of autonomy from ADS (Choi and Ji 2015), improving trust in the technology is a highly relevant task. One approach to improving trust is through increasing the transparency of the technology. Because fully testing a level 4–5 ADS is infeasible (Kalra and Paddock 2016, Schwarting et al. 2018), various methods, namely Bayesian neural networks, have been proposed to quantify the uncertainty of the models powering the ADS (McAllister et al. 2017, Kendall and Gal 2017). This sort of technology can be applied to end-to-end systems – like the one proposed by Bojarski et al. (2016) – as a means of conveying model uncertainty to the driver, both to improve trust and to accurately signal a handover when the model is overly uncertain. Another way of improving trust is by notifying the driver preemptively when the ADS is going to behave in a way that deviates from its expected behavior (Haspiel et al. 2018). Further transparency that shows more general-purpose decision-making of the ADS has also been proposed by Gowda et al. (2014). It has been shown that ADS with AI that show high autonomy and anthropomorphic characteristics can increase the trust of the driver in the ADS (Lee et al. 2015). Although transparency can improve trust, it can also increase the cognitive load resulting in negative feelings from the driver. As such, determining the proper amount of transparency is a nontrivial task (Koo et al. 2015). By providing a reasonable level of transparency through a multi-modal

interface, our perception augmentation module (see Chapter 3) improves trust in the ADS without introducing a source of additional cognitive load, i.e., the driver enjoys using the system.

2.1.2 Artificial Intelligence and Human Interfaces

Designing proper computer-human interfaces for ADS that provide an enjoyable experience is a grand challenge in the literature (Dikmen and Burns 2016, Endsley 2017, Brown and Laurier 2017). The problem is two-sided in that (1) vehicles must be able to convey data to the driver using an information interface of any combination of visual and auditory alerts, and (2) the vehicles must embrace an appropriate control interface that allows the driver to intervene if necessary, or simply desired (Goodrich et al. 2008). Arguments have also been made for the application of AI to understand and respond to human actions (Khandelwal et al. 2017). The prevalence of using AI techniques in ADS coupled with the patterns of research surrounding AI and Human-Computer Interaction (HCI) has rendered AI systems that are opaque to the end-user and unable to explain the rationale behind the artificially intelligent logic (Grudin 2009). The notion of Explainable Artificial Intelligence (XAI) has emerged as a means of unifying the two fields of study – i.e., AI and HCI (Gunning 2017).

Because ADS may behave in unpredictable ways based on sensor data and complex nonlinear systems, informing the driver of the ADS state through an information interface is both a vital and challenging task (Surden and Williams 2016). To complicate matters, drivers of vehicles with ADS show interest in disengaging from the driving task to perform more enjoyable activities like listening to the radio, talking to passengers, etc. (Pfleging et al. 2016). This desire to disengage introduces a need to study multimedia systems capable of conveying the proper amount of information to the driver. Often, information interfaces are composed of a set of multi-modal approaches – i.e., using visual and auditory alerts. Alerts can be concrete like visually displayed text or a spoken phrase. Alerts can also be abstract like a "beep" played at high volume and/or frequency, or a static/flashing colored light. Because of the additional processing time required for concrete alerts to propagate and be understood by a

human, abstract alerts produce faster response times than concrete alerts in ADS handover simulations (Politis et al. 2017). Auditory alerts, like before Automatic Emergency Braking (AEB) activates or when Forward Collision Warning System (FCWS) detects an imminent collision, have been shown to reduce anxiety and increase enjoyment and perceived control of the driver (Koo et al. 2016, Bazilinskyy and de Winter 2015). In a similar vein, virtual assistants, like Amazon's "Alexa" or Apple's "Siri", can augment the experience in vehicles with ADS by responding to voice prompts (Lugano 2017). Likewise, visual alerts, including dashboards and infotainment systems, have also been proposed to accommodate the more autonomous nature of vehicles with advanced ADS (Udovicic et al. 2015, Gowda et al. 2014). Although information interfaces can improve trust and HCI, proper design and thorough study are necessary because the cognitive load introduced by the system could potentially evoke a negative emotional response from the driver to the point that they disengage or ignore the system (Koo et al. 2015, Casner et al. 2016). Our perception augmentation module, presented in Chapter 3, features a combination of abstract and concrete visual alerts. Although the system features concrete alerts, these are a secondary mechanism intended to improve the learn-ability of the abstract alerts.

2.2 Image Degradation and Restoration

2.2.1 Image Acquisition Model

In the context of this dissertation, the term *image* refers exclusively to *natural images* that are collected through Red-Green-Blue (RGB) image sensors, i.e., Complementary Metal-Oxide-Semiconductors (CMOSs) or Charge-Coupled Devices (CCDs). Although the details of image acquisition extend past the scope of this work, a brief discussion of the model of acquisition aids in the presentation of the concepts of image degradation and restoration. Image sensors integrate the reflection of light off of real-world objects over a period known as *exposure time*. To form two-dimensional images, image sensors are arranged into a grid array where each

sensor is responsible for the acquisition of a single discrete *picture element*, i.e., *pixel*, in the digital image. The cones of the Human Visual System (HVS) sense three primary bands of colored light: red, green, and blue. As such, color image sensors contain three separate planes of sensor grids that independently acquire the red, green, and blue bands of light into RGB pixel vectors. For CMOS sensors, image sensors are frequently arranged using the mosaiced *Bayer pattern* and reconstructed via de-mosaicing algorithms (Bayer 1976). Equation 2.1 introduces the model of image acquisition for a single image sensor where T represents the exposure time and $I(x, y, t)$ is the intensity measured by the sensor at row x and column y at the continuous point in time t. α represents the Camera Response Function (CRF), which is typically a positive sigmoid-shaped activation function that captures the nonlinear mapping of radiance to luminance in the human eye.

$$f(x, y) = \alpha\Big(\frac{1}{T}\int_0^T I(x, y, t)\mathrm{d}t\Big) \tag{2.1}$$

2.2.2 Image Degradation Models

The term *degradation* loosely includes a variety of destructive systems that may influence the acquisition of the image signal during the exposure window. A simple example of degradation is motion blur where an object in the scene quickly moves during the acquisition window, resulting in a low-pass smearing effect on the pixels containing the moving object. Other degradation systems include sensor noise and mosaicing artifacts. It is implied that *restoration* means the reconstruction of a clean image from one that has been affected by any of the aforementioned degradation systems. Image degradation is frequently modeled as an operator $\mathcal{H}$ that degrades an image and applies an additive noise term. In the special case where the degradation system $\mathcal{H}$ is linear and shift-invariant, it can be represented using Equation 2.2 as the convolution of the image $f(x, y)$ with the point-spread function $h(x, y)$ plus the noise image $\eta(x, y)$.

$$g(x, y) = (f * h)(x, y) + \eta(x, y) \tag{2.2}$$

The application of this operator to an input image $f(x, y)$ produces the degraded image $g(x, y)$. The noise term is optional and typically models the digital or analog noise that occurs as the result of over-exposure, compression, storage, transmission, or precision of the numerical system, to name a few causes. Applying the convolution theorem to Equation 2.2 allows the system to be rewritten in Equation 2.3 as the point-wise product of the Fourier coefficients of the image $F(u, v)$ and the point-spread function $H(u, v)$ plus the noise term $N(u, v)$.

$$G(u, v) = F(u, v)H(u, v) + N(u, v) \tag{2.3}$$

In many practical cases of image degradation, the operation $\mathcal{H}$ is nonlinear and shift-varying. For instance, motion blur that occurs as the result of moving objects in the scene, and out-of-focus blur that manifests due to objects in the scene existing at different depths are two examples of image degradation systems that are not strictly linear shift-invariant. The model of image degradation can be extended to more realistic blur systems by loosening the global constraint of the point-spread function in Equation 2.4. In this new model, each pixel has a blur kernel and the degraded image is the result of super-imposing each pixel's degradation. Because each pixel in the model can be affected by degradation individually, the model can better represent blur that results due to motion either of objects in the scene or the sensor itself. However, this model cannot account for the occlusion between objects of different depths in the scene.

$$g(x, y) = \left(\sum_{x', y'} (f(x^{'}, y^{'}) * h_{x', y'})(x, y) \right) + \eta(x, y) \tag{2.4}$$

2.2.3 Classical Image Restoration Approaches

A common problem in image processing is to invert the effect of the degradation system on a previously degraded image $g(x, y)$ to produce a restored image $\hat{f}(x, y)$ that is equal to- or approximately equal to the clean image $f(x, y)$. In the special case where $\mathcal{H}$ is known to be linear and shift-invariant, $h(x, y)$ is known, and $\eta(x, y)$ is known, the restoration of $\hat{f}(x, y)$

may be accomplished trivially using an inverse filter to perfectly recover $f(x, y)$, – i.e., by deconvolving in the frequency domain (see Equation 2.3). In most practical cases, $\eta(x, y)$ is not known and inverse filters will not recover $f(x, y)$. When $\eta(x, y)$ is unknown, a common technique is to apply regularized inverse filters, Wiener filters, or the Lucy-Richardson method that account for the noise in the image (Wiener 1949, Richardson 1972, Lucy 1974). Practical problems often embody a lack of knowledge of both the noise term $\eta(x, y)$ *and* the point-spread function $h(x, y)$. In the case where $\mathcal{H}$ is still guaranteed to be linear and shift-invariant, the data of $f(x, y)$ may provide a reasonable estimate of $h(x, y)$. In a large number of applications, $\mathcal{H}$ is nonlinear and shift-varying (see Equation 2.4). When a large amount of data exists describing the mechanics of the degradation systems at play, deep learning is a viable solution for learning to restore images that have been degraded by non-uniform blur systems (Nah et al. 2017).

2.2.4 Non-Uniform Image Degradation Dataset Synthesis

Because standard deep learning frameworks rely on paired samples, it is necessary to understand a method of obtaining a dataset of paired samples where each image pair embodies a clean version and a degraded version of the same real-world scene. A common approach is to approximate the acquisition process in Equation 2.1. Nah et al. (2017) use a camera that captures frames at $240Hz$ to record a video then approximate the continuous acquisition process using Equation 2.5.

$$\hat{f}[m, n] = \hat{\alpha}\Big(\frac{1}{K}\sum_{k=0}^{K}\hat{\alpha}^{-1}(f[m, n, k])\Big) \tag{2.5}$$

To generate an image pair, a linear sub-sequence of K frames of the source video f is averaged to approximate the integration of luminance over the time window of length T as $\hat{f}$. Nah et al. (2017) use K values between 7 and 13. Because each frame will have been influenced by the CRF of the camera, it is necessary to take this into account when averaging frames. Nah et al. (2017) use a gamma curve to approximate the CRF as $\hat{\alpha}(\cdot)$ and apply the inverse $\hat{\alpha}^{-1}(\cdot)$ before

averaging sharp images. Their final dataset, coined as the "GoPro" dataset, contains 2103 training samples and 1111 validation samples. Nah et al. (2019) extend the method using a camera with a source rate of $1920Hz$ to produce a blurred video with a cinema-standard rate of $24Hz$. The larger dataset of Nah et al. (2019), entitled "REDS", contains 24000 training samples and 3000 validation samples. It is worth noting, paired samples are not strictly necessary in all deep learning frameworks; Zhang et al. (2020) demonstrates the success of using unpaired samples to first build a generative model of the degradation system then use the model to generate paired training data to train a restoration model from.

2.2.5 Metrics

When paired image data are available, image processing models can be validated objectively using Peak Signal-to-Noise Ratio (PSNR) in Equation 2.6 which measures the log of the reciprocal MSE between source image x and restored image x_r times the squared maximum of the image data domain R^2. For 8-bit fixed-point RGB images, $R = 255$ and for 32-bit floating-point RGB images $R = 1$. The images must both be M pixels tall, N pixels wide, and have C channels of image data. For monochromatic images $C = 1$ and for color images, such as RGB, $C = 3$. PSNR reduces image tensor pairs to a scalar value measured in decibels (dB) that approaches infinity as the images become more similar. As such, a restored-sharp image pair will have a higher PSNR relative to the restored-degraded pair when the restoration of pixels is accurate.

$$PSNR = 10 \log_{10} \left(\frac{R^2}{\frac{1}{MNC} \sum_i^M \sum_j^N \sum_k^C (x[i,j,k] - x_r[i,j,k])^2} \right) \tag{2.6}$$

Because the HVS is non-linear and tuned to extract structured information from image data, PSNR is frequently a poor measure of image quality to humans. As an alternative, Wang et al. (2004) develop the Structural Similarity Index Measure (SSIM) which operates under the heuristic that structural information in the image is salient for human perception. SSIM is

computed from three independent comparisons of luminance values, contrast, and structure between two images to provide a quality measure that is better matched to the HVS than is PSNR. The mathematical definition of SSIM is verbose and extends past the scope of this work; we refer the reader to the work of Wang et al. (2004) for the full details of the computation of SSIM.

2.2.6 Generative Adversarial Networks

Specifically related to image synthesis, Goodfellow et al. (2014) formalized the method of GANs for learning a mapping between a Gaussian latent space and a structured image domain. They formulate the synthesis problem as an adversarial game between a generator and a discriminator network to train both networks using the same learning framework in tandem. This allows the discriminator to adapt to the improvements of the generator at run-time and also share its representational understanding with the generator through its upstream gradient. Several extensions to the GAN have been vetted in the literature. Mirza and Osindero (2014) propose the Conditional GAN (CGAN) to condition the generation/discrimination of images on known latent parameters. Opposed to the traditional usage of random latent spaces, this allows the authors to enforce conditional representation learning in the model which then allows for the synthesis of objects of particular classes (e.g., digits, object types). An alternative approach, first discussed by Odena et al. (2017), is known as Auxiliary Classifier GAN (AC-GAN). AC-GAN is similar to CGAN but refactors the discriminator as a dual network with one branch to discriminate real versus fake samples, and one branch to perform latent mapping, i.e., for classification. Empirically, this results in training a generator that produces more visually appealing images than either CGAN or the vanilla GAN. A final innovation in the general GAN framework worth mentioning is the Wasserstein GAN (WGAN) studied by Arjovsky et al. (2017). The WGAN replaces the standard loss function of cross-entropy for the Wasserstein loss (i.e., the Earth mover's distance). Put simply, this loss allows the loss network to better separate the distributions of the degraded and latent images by removing the limit on

the dynamic range that the usage of probability imposes. The WGAN model also uses a weight clipping strategy to impose a 1-Lipschitz continuity constraint on the model. Gulrajani et al. (2017) show that replacing this discontinuous weight clipping policy with a gradient penalty loss stabilizes the training.

Although vanilla GANs learn a mapping from a flattened latent space in a spatially unrelated domain to images, the idea can extend to *image-to-image transformations* where both the inputs and outputs are structured images. Isola et al. (2017) first studied this idea to map semantic segmentation maps to natural images. The proposed *pix2pix* model allows them to enforce semantic structure on the synthesized images. This runs contrary to the vanilla GAN that has no parameterization over where objects in the synthetic image will be placed. Zhu et al. (2020) extend the idea of image-to-image translation with the method of *cycle consistent translations* where a mapping back to the input image is also learned to find translations between domains without paired training samples. Image-to-image translation has also been applied to the problem of image super-resolution where a generator model up-samples an image and imputes missing data to produce a higher resolution image. Ledig et al. (2017) have demonstrated the efficacy of GANs on the image super-resolution problem.

The GAN learning framework embodies a min-max game between the discriminator and the generator. Nguyen et al. (2017) rework the discriminator into a duality between KullbackâĂŞLeibler (KL) and reverse KL divergences. By training one discriminator branch with a bias towards the generator and another branch with a bias toward the discrimination task, they can mitigate the impact of mode collapse. Li et al. (2017) propose a similar idea where the optimization of a GAN is restructured as a dual to resolve instability.

Adversarial Loss

Equation 2.7 describes the mini-max loss function for the vanilla GAN framework (Goodfellow et al. 2014). During training, the discriminator D attempts to maximize the loss while the generator G tries to minimize it. $D(x)$ represents the discriminator's calculation of the

probability that a sample of real data x is fake. Here p_{data} is the empirical distribution of the real training data and $\mathbb{E}_{x \sim p_{data}(x)}$ is the expectation over the real training data. z represents a sample vector from an unknown latent space that the generator G maps to a convincing realistic sample $x_f = G(z)$. $D(G(z))$ then expresses the discriminator's belief that x_f is fake. $\mathbb{E}_{z \sim p_z(z)}$ describes the expectation over the empirical distribution of the fake training data p_z based on the latent sample z. It is worth noting that Equation 2.7 is a dual formulation of the binary cross-entropy loss where real samples are labeled as 0 and fake samples are labeled as 1.

$$\mathcal{L}^D_{GAN} = \mathbb{E}_{x \sim p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.7}$$

Because the weights of the generator do not influence the calculation of the left term in the loss, Equation 2.7 only represents the loss of the discriminator. Equation 2.8 describes the loss of the generator, which follows the same derivation, but omits the left-hand term that the generator cannot back-propagate through.

$$\mathcal{L}^G_{GAN} = \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.8}$$

Goodfellow et al. (2014) also note that the form of the generator objective in Equation 2.8 that attempts to produce samples that are not like fake data causes saturation of the generator's gradient signal. As an amendment, they refactor the semantics of the generator loss to produce samples that are the most like real data. Equation 2.9 describes this *non-saturating* GAN loss that is typically used.

$$\mathcal{L}^G_{GAN-NS} = -\mathbb{E}_{z \sim p_z(z)}[\log(D(G(z)))] \tag{2.9}$$

Least Squares Adversarial Loss

Mao et al. (2017) note that sigmoidal losses like cross-entropy introduce vanishing gradients in the cases where the generator produces data that defeats the decision-boundary of the discriminator, but fails to adequately capture the distribution of real data p_{data}. This results in

generators that synthesize samples that lack adequate detail when evaluated by the HVS and discriminators that cannot produce a strong backward pass signal to improve the synthesis. To resolve this limitation, Mao et al. (2017) propose replacing the cross-entropy loss with a least-squares loss that targets these particular samples that fall far past the decision boundary but are far from the realistic samples. Equations 2.10 and 2.11 describe the revised minimization loss functions for the Least-Squares GAN (LSGAN) discriminator and generator, respectively.

$$\mathcal{L}^{D}_{LSGAN} = \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x)-1)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))+1)^2] \tag{2.10}$$

$$\mathcal{L}^{G}_{LSGAN} = \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-1)^2] \tag{2.11}$$

Wasserstein Adversarial Loss

The loss function in vanilla GANs tends to over-saturate resulting in poor separation of the distributions of real and fake data. Arjovsky et al. (2017) propose adopting the Wasserstein loss function (i.e., "earth movers distance") to mitigate this shortcoming in the vanilla GAN learning framework. The loss of the Wasserstein GAN (WGAN) restructures the discriminator D from estimating probits to estimating unconstrained logits as *the critic* C. Equation 2.12 displays the loss that the critic maximizes. This loss encourages the critic to produce large output values for data that derive from the real distribution and small output values for data that derive from the fake distribution.

$$\mathcal{L}^{C}_{WGAN} = \mathbb{E}_{x\sim p_{data}(x)}[C(x)] - \mathbb{E}_{z\sim p_z(z)}[C(G(z))] \tag{2.12}$$

Like the generator in the vanilla GAN framework, the generator in the WGAN has a modified loss that omits the pathways in the graph that are specific only to the discriminator. Equation 2.13 shows the loss that the generator maximizes in the WGAN framework. The generator loss encourages the critic to produce large output values based on the synthetic data of the generator

without influencing the parameters of the critic during back-propagation.

$$\mathcal{L}_{WGAN}^{G} = \mathbb{E}_{z\sim p_z(z)}[C(G(z))] \tag{2.13}$$

To enforce the 1-Lipschitz continuity during training of WGANs, Arjovsky et al. (2017) use a naive weight clipping strategy that results in unstable training. Gulrajani et al. (2017) demonstrate that an additional gradient penalty loss can be applied in place of this weight clipping strategy to enforce the 1-Lipschitz continuity while exhibiting a more stable training process. Equation 2.14 describes the gradient penalty loss used to enforce continuity of the updated WGAN Gradient Penalty (WGAN-GP). The loss is calculated based on random samples $\hat{x} \sim p_{\hat{x}}$ that are generated by randomly sampling a pair of real and fake data and then randomly sampling a linear interpolation between the two samples. The goal of the loss is to encourage the L_2 norm of the gradients to be at most 1.

$$\mathcal{L}_{GP} = \mathbb{E}_{\hat{x}\sim p_{\hat{x}}(\hat{x})}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{2.14}$$

Relativistic Adversarial Loss

Jolicoeur-Martineau (2018) proposes the concept of Relativistic average GANs (RaGANs) to address issues of instability with vanilla GAN and WGAN loss functions. In a vanilla GAN, the discriminator calculates a probability $D(\cdot)$ that represents whether data are real or fake. In a RaGAN, the output of the discriminator is modified to represent the probability that real/fake data are more realistic than randomly sampled fake/real data. Let C, *the critic*, be the logits of the discriminator D before the final sigmoid activation function. Equation 2.15 describes the calculation of the probability that real data x are more realistic than the expected critic of fake data $x_f = G(z)$ by the relativistic discriminator $\hat{D}$ where σ is the sigmoid activation function.

$$\hat{D}(x) = \sigma(C(x) - \mathbb{E}_{z\sim p_z(z)}[C(x_f)]) \tag{2.15}$$

Likewise, Equation 2.16 shows the calculation of the probability that fake data $x_f = G(z)$ are more realistic than the expected critic of real data x.

$$\hat{D}(x_f) = \sigma(C(x_f) - \mathbb{E}_{x \sim p_{data}(x)}[C(x)]) \tag{2.16}$$

The loss function for the relativistic average discriminator in Equation 2.17 can be formulated following the vanilla GAN discriminator loss function (see Equation 2.7). The discriminator is trained to minimize this value.

$$\mathcal{L}^D_{RaGAN} = \mathbb{E}_{x \sim p_{data}(x)}[\log(\hat{D}(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - \hat{D}(x_f))] \tag{2.17}$$

The loss function for the generator to minimize in the relativistic framework, shown in Equation 2.18, follows the same form as the discriminator, but with inverted labels. This updated formulation allows the generator's loss function to incorporate information from the distribution of real data that is neglected in the vanilla GAN generator loss function in Equation 2.8 and the WGAN generator loss function in Equation 2.13.

$$\mathcal{L}^G_{RaGAN} = \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - \hat{D}(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(\hat{D}(x_f))] \tag{2.18}$$

It is worth noting, RaGAN is a generalization of Relativistic GAN (RGAN) that addresses issues with the runtime complexity of the originally formulated RGAN loss during optimization. In RGAN, the expectation in the formulation of $\hat{D}(\cdot)$ in Equations 2.15 and 2.16 is simply replaced with a critic of a randomly sampled data of the opposing class referred to as $\tilde{D}$. The problem with the RGAN formulation is that it requires $O(n^2)$ combinations of paired data to visit the entire data space. By instead approximating the average critic using an expectation, this complexity is reduced to $O(n)$ in the RaGAN model.

Content Loss

Specific to image-to-image translation tasks with paired image data, content losses directly enforce that a generator model learns a particular transform between the image pairs. Unlike adversarial losses, there is no learned component and the loss is instead calculated based on image pairs. Common applications of content losses include image restoration tasks where degraded images are mapped back to sharp images and super-resolution tasks where down-sampled images are up-scaled. One such content loss is MSE shown in Equation 2.19 where x is a real image and $x_f = G(z)$ is a synthetic image.

$$\mathcal{L}_{MSE} = \mathbb{E}_{x \sim p_{data}(x), z \sim p_z(z)}[(x - x_f)^2] \tag{2.19}$$

MSE places more precedent on errors that diverge far from the mean. A choice of content loss that is piece-wise linear is MAE that replaces the square root operation in MSE with the absolute value operator in Equation 2.20.

$$\mathcal{L}_{MAE} = \mathbb{E}_{x \sim p_{data}(x), z \sim p_z(z)}[|x - x_f|] \tag{2.20}$$

For image synthesis tasks, MSE and MAE over luminance can introduce unwanted artifacts when applied as the primary loss function. To encourage stronger natural image priors in generator models, Johnson et al. (2016) apply a *perceptual* content loss based on pre-trained classification networks. Namely, the mean-squared error between shallow activation maps of a pre-trained model is calculated based on a real image x and corresponding fake image x_f. The particular activation map is the `block3_conv3` output of the Visual Geometry Group (VGG)-19 network after the ReLU activation. This particular layer is known to extract low-level representations and filter for object presence according to the ImageNet labels on which the network was trained (Simonyan and Zisserman 2014). This is a salient property, both for style

transfer and image restoration, as the perceptual content loss calculates image differences according to experimental natural image priors.

Some prior work has further developed content loss functions based on frequency-domain transformations, namely, the Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT). When applied to estimation-based image processing tasks, the DFT frequently introduces artifacts due to the periodic nature of the Discrete Fourier Series (DFS) that results in discontinuities around the window of the image. For image processing tasks, it is common to apply to DCT Type-2, which symmetrically extends the image in the spatial domain to prevent such discontinuities from occurring in the DFS of the extended image. It has further salient properties of being real-valued, as compared to the complex DFT, and good energy compaction capabilities. The DCT-2 is the standard transform applied by the Joint Photographic Experts Group (JPEG) compression algorithm (Wallace 1992). Equation 2.21 presents the Type-2 DCT in its orthogonal form (i.e., with scaling factor α shown in Equation 2.22). Because frequency-domain representations are sparse, the JPEG algorithm applies the transform over patches of size $M = N = 2^l$ to achieve a better compression ratio than using a DCT over the entire image would produce. Standard values of l are $l \in \{3, 4, 5\}$.

$$F[k,l] = \alpha(k,M)\alpha(l,N) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f[m,n] \cos\left(\frac{(2m+1)k\pi}{2M}\right) \cos\left(\frac{(2n+1)l\pi}{2N}\right) \tag{2.21}$$

$$\alpha(x,X) = \begin{cases} \sqrt{\frac{1}{X}} & \Longleftarrow \; x = 0 \\ \sqrt{\frac{2}{X}} & \Longleftarrow \; \text{otherwise} \end{cases} \tag{2.22}$$

Because the DCT is linear, it has a well defined analytical derivative (Reeves and Kubik 2006), which is necessary for the DCT to be used in deep learning pipelines. Khan et al. (2019) apply the DCT-II in their development of *Spectral Dropout*, which drops the activation of neurons that have DCT coefficients with low magnitude or excessive noise. Czolbe et al. (2020) study a GAN model using a loss function based on the p-norm of DCT coefficients of

images. Compared to using perceptual content losses, they find that their approach produces better results. Bhattacharya et al. (2018) argue that the sparse nature of frequency domain representations allows auto-encoder models to be more efficient by packing signal energy into smaller spatial forms This is because natural images have frequency spectra that are Gaussian centered around the DC coefficient. Sims (2020) specifically find that using DFT and DCT-based loss functions allows for the training of super-resolution models that produce higher PSNR values and better qualitative evaluation relative to the same models trained with spatial loss functions. The intuition behind their method derives from the fact that spatial loss functions place equal weighting over the frequency spectrum, but the average HVS is biased toward a specific band of the spectrum.

2.2.7 Training Deep Networks

Because the training of GANs relies on the parallel training of two networks and is known to be highly convex, stabilizing the training and convergence is a frequently discussed topic. During the training of GAN models, it is common for the discriminator to converge fully, resulting in a vanishing gradient for the generator. Arjovsky and Bottou (2017) provided analytical grounds for why adding noise to the inputs of the model during training would resolve the issue of vanishing gradients. The concept of adding noise to the inputs of neural networks to act as a regularization mechanism has been well studied empirically as well (Shorten and Khoshgoftaar 2019). Specifically related to training models using synthetic data, Carlson et al. (2018) have shown that applying a noise model inspired digital camera image acquisition to input images can improve the generalization of the model to real-world data. From the opposing angle, Xie et al. (2016) have studied the idea of disturbing target labels. They propose to randomly perturb training labels to prevent the classification model from converging on local optima.

Another common problem that is observed in training GANs is mode collapse where the generator converges on a small and sub-optimal subspace akin to emitting mean value estimates. Wang et al. (2016) expanded the idea of ensemble learning to the discriminator of

a GAN model to improve the learned distribution of the parallel generator model. This allows their generator model to produce synthetic images that better resemble test examples in their structured content. Metz et al. (2016) propose unrolling several steps of the discriminator optimization algorithm to prevent mode collapse. From a different perspective, Srivastava et al. (2017) suggest preventing mode collapse by altering the generator with an additional inverse network that maps synthesized outputs back to the latent noise space. One final way of preventing mode collapse, presented by Lin et al. (2020), is to modify the discriminator model to classify bins of samples, as opposed to individual samples.

Besides the issues specific to GANs, deep neural networks often benefit from normalization and regularization methods to counteract data bias and prevent over-fitting. Ioffe and Szegedy (2015) introduced the method of *batch normalization* where mini-batches of layer activation maps are normalized during training. The batch normalization technique has been widely adopted by countless works and remains a highly effective strategy for classification, object detection, and semantic segmentation, to name a few applications. Ulyanov et al. (2016) expanded upon the idea of batch normalization with the idea of an *instance normalization* layer that normalizes individual channels of the activation maps. Kupyn et al. (2018) have demonstrated the efficacy of instance normalization for image restoration tasks. *Layer normalization*, presented by Ba et al. (2016), takes a different approach by normalizing the entire activation map for each independent sample in a mini-batch. The benefit of layer normalization lies in the reduced impact of batch size on the optimizer and improved convergence of the training algorithm.

Traditional normalization strategies in deep learning act upon the activation maps that emit from individual layers. Salimans and Kingma (2016) introduce a novel method that involves a re-parameterizing the weights to produce a *weight normalization* effect similar to batch normalization, but without some of the challenges associated with batch normalization. They show that their method is effective for applications like generative synthesis that are sensitive to noise. Specifically related to GAN discriminator models, Miyato et al. (2018) suggest an

idea called *spectral normalization* to stabilize the model. Their idea is closely related to that of Gulrajani et al. (2017) in that they aim to constrain the Lipschitz continuity of the WGAN model. Spectral normalization is an important discovery that can replace the unstable weight clipping mechanism employed by the original WGAN.

2.2.8 State-of-the-art Image Restoration Approaches

Lai et al. (2016) provide a comprehensive survey of various algorithms used for single image blind de-blurring. Their analysis reveals that the algorithms perform well on synthetic data, but do not generalize well to real-world systems of degradation. Schuler et al. (2015) present some of the first works using Convolutional Neural Networks (CNNs) to deblur images. Their model estimates the kernel of uniform blur to deconvolve the degradation from the image using, e.g., regularized inverse filters. Nah et al. (2017) address the limitation of classical de-blurring algorithms for real-world non-uniform motion blur systems both by proposing a technique for approximating real-world blurs using digital systems (see Section 2.2.4) and by developing a deep auto-encoder method, "DeepDeblur", for learning to restore blurry images. Their auto-encoder method outperforms the existing algorithms in terms of PSNR, SSIM, and subjective evaluation. DeepDeblur is a *multi-scale* generator architecture that takes advantage of the encoder-decoder design to produce restored image outputs at full-size, half-size, quarter-size, etc. The multi-scale architecture aids in the perceptible quality of fine detail in the image as well as more global, large-scale features that require the larger receptive field that is achieved by down-sampling the image. DeepDeblur is the first work on deblurring to use an adversarial loss, namely GAN, in the training process.

Kupyn et al. (2018) further improve state-of-the-art image de-blurring using a GAN algorithm coined as "*DeblurGAN*". DeblurGAN applies a WGAN-GP loss function during training, which differs from the vanilla GAN loss used in DeepDeblur. A novelty in the loss function of DeblurGAN lies in the application of the method of Johnson et al. (2016). Kupyn et al. (2018) note that MSE does not correlate with quality to the HVS and propose replacing the MSE

content loss function with a *perceptual loss function*. Using the `block3_conv3` layer outputs of the VGG-19 network as an auxiliary content loss, DeblurGAN can produce perceptibly better results than prior works. To solve issues with training stability, Kupyn et al. (2018) also introduce the notion of a *global residual skip connection* that forwards the input image to the output of the network for a residual update. Kupyn et al. (2019) improve upon DeblurGAN with an update called "*DeblurGAN-v2*". The updated DeblurGAN-v2 can utilize various backbone generator architectures to fit the needs of the network designer in terms of the ratio of SSIM to Floating Point Operations (FLOPs). They also replace the WGAN-GP loss with a more stable relativistic loss, namely, RaGAN.

Tao et al. (2018) extend upon the multi-scale model of DeepDeblur with the concept of recurrent restoration. Their "Scale-Recurrent Network (SRN)" iteratively restores and up-scales down-scaled versions of the input in a recurrent graph to produce the final full-resolution output. This is driven by the intuition that the problem solved at each scale is the same and will thus benefit from parameter sharing. Zhang et al. (2020) note that existing synthetic blur datasets poorly reflect the dynamics of realistic blur systems due to invalid assumptions made when generating paired sets of sharp and blurry images using methods of, e.g., Nah et al. (2017). To address this limitation, they design an auto-encoder that learns to realistically blur images then deblur them in cascade. The strength of their method lies in the training procedure that no longer relies on paired images and can thus be trained with independent sets of sharp and degraded images. Zhang et al. (2020) also point out that the outputs of `block3_conv3` are sparse due to the REctified Linear Unit (ReLU) activation function. They propose developing the perceptual loss from the activation maps before this loss function to provide a denser gradient signal. Gao et al. (2019) replace the local and global residual skip connection structures used in Kupyn et al. (2018) and Kupyn et al. (2019) with high-order skip connections between scales in the auto-encoder network. They argue that the global skip and local skip connections have gradient pathways that do not cross in training due to their first-order nature. Replacing these simpler

structures with high-order residuals allows their model to learn more complex representations without sacrificing the benefits of residuals in terms of reducing the vanishing gradient problem.

Zhang et al. (2019) note that end-to-end image restoration models suffer from a high capacity to performance ratio. To improve the statistical performance of deep learning image restoration, they propose using *hierarchical multi-patch* models, where images are restored in non-overlapping chunks that grow in size until reaching the full-size image output at the final stage. This allows them to efficiently compute progressive updates in a fine-to-coarse manner, whereas prior methods have all used coarse-to-fine approaches. Their "Deep Multi-Patch Hierarchical Network (DMPHN)" achieves state-of-the-art performance on the GoPro dataset at the time of their publication. Suin et al. (2020) extend upon the hierarchical multi-patch concept by introducing a self-attention mechanism in their "Spatially-Attentive Patch-Hierarchical Network (SAPHNet)" to improve the PSNR metric while simultaneously reducing the inference time. The self-attention module allows their model to learn data-dependent masks for applying non-uniform restoration in the case of, e.g., motion blur of a single object in the scene. Purohit and Rajagopalan (2020) develop a similar technology that further focuses on the non-uniform nature of real-world blurs using "Region-Adaptive Deblurring Networks (RADNets)". A novelty of their approach lies in the Dense Deformable Module (DDM) that introduces awareness of motion trajectories to the model. Such mechanics are normally modeled using affine transformations of pixel coordinates, which cannot be implicitly learned using traditional convolutional layers. Tsai et al. (2021) extend the self-attention concept with their Blur-Aware Modules (BAMs) that address the slow inference time of existing works. Their "Blur-Aware Network (BANet)" can replace the recurrent self-attention structures of Suin et al. (2020) and Purohit and Rajagopalan (2020) and achieve a faster inference time and higher PSNR as a result. The BAM applies a combination of strip-pooling and attention-refinement to better identify blurred regions of the still image. Zamir et al. (2021) approach the image restoration problem with a multi-stage architecture, "Multi-Stage Progressive Image Restoration Network (MPRNet)", that is capable of adapting to several image restoration tasks, including deblurring, deraining, and denoising.

They also propose a combined loss function based on the error between image gradients, i.e., they calculate the MSE between images after being convolved with a Laplacian kernel. Chen et al. (2021) extend the idea of MPRNet with their Half-Instance Normalization (HIN) block in a model entitled “Half-Instance Normalization Network (HINet)”. The key innovation of HINet lies in the HIN block that applies instance normalization to only half of the data. HINet is the current state-of-the-art for the image restoration datasets studied in this work.

Table 2.2 presents a comparison of the choice of optimization parameters of state-of-the-art image deblurring models. All prior methods use the Adam optimizer with a patch size of 256×256 when training other than Zhang et al. (2020) who use a patch size of 128×128. The choice of learning rate varies little between the various works and is frequently selected as the default value of $1\mathrm{e}{-4}$ (Kingma and Ba 2014). Most authors do not provide the β_1 and β_2 parameters, but those who do report using the default values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ (Tao et al. 2018, Gao et al. 2019, Tsai et al. 2021). It is worth noting, Zhang et al. (2020) do not explicitly state the optimizer of choice, but provide parameters $\alpha = 0.005$ and $\beta = 0.01$ that do not resemble usual selections for Adam. There is a high degree of variance in the batch size with some authors choosing as few as $b = 1$ sample per batch and others going as high as $b = 64$ samples per batch. Likewise, there is a wild amount of variation in the training duration, measured in *epochs*, and the learning rate schedules. This is to be expected as each work presents a different model, but provides little insight into the appropriate choice of optimization parameters.

There is little consensus on the choice of learning rate schedule, other than the confirmed usage of either linear, exponential, or Cosine (Loshchilov and Hutter 2016) decay strategies. Nah et al. (2017) apply an exponential decay strategy using a rate of 0.1 after $3\mathrm{e}5$ epochs have transpired at the initial learning rate. Kupyn et al. (2018) train at the initial rate for 150 epochs then apply a linear decay to 0 for another 150 epochs. Tao et al. (2018) use an exponential decay throughout the entire training that reduces the rate to $1\mathrm{e}{-6}$. Kupyn et al. (2019) follow their original method and decay to $1\mathrm{e}{-7}$ after the first 150 epochs. Gao et al. (2019) use an

Model	Loss			Architecture		
	Adversarial	**Perceptual**	**Content**	**Attention**	**Multi-Scale**	**Recurrent**
DeepDeblur (Nah et al. 2017)	GAN	✗	MSE	✗	✓	✗
DeblurGAN (Kupyn et al. 2018)	WGAN-GP	✓	✗	✗	✗	✗
SRN (Tao et al. 2018)	✗	✗	MSE	✗	✓	Scale
DeblurGAN-v2 (Kupyn et al. 2019)	RaGAN	✓	MSE	✗	✗	✗
Zhang et al. (2020)	RGAN	✓	MSE	✗	✗	✗
Gao et al. (2019)	✗	✗	MSE	✗	✗	✗
DMPHN (Zhang et al. 2019)	✗	✗	MSE	✗	✓	Patch
SAPHNet (Suin et al. 2020)	–	✗	–	✓	✓	Patch
RADNet (Purohit and Rajagopalan 2020)	–	✗	–	✓	✗	✗
BANet (Tsai et al. 2021)	✗	✗	–	✓	✗	✗
MPRNet (Zamir et al. 2021)	✗	✗	Charbonnier, MSE	✓	✓	Patch
HINet (Chen et al. 2021)	✗	✗	PSNR	✓	✓	Image
This Study	Multiple	Multiple	Multiple	✗	✗	✗

Table 2.1: A comparison of state-of-the-art deep learning-based image restoration models. A mark of "–" indicates that a value was not specified by the authors.

exponential decay schedule with a power of 0.3. (Zhang et al. 2019) also use exponential decay, but with a rate of 0.1. Suin et al. (2020) describe a quantized exponential decay strategy that halves the learning rate every $1e5$ epoch. (Tsai et al. 2021) train at the initial learning rate for 50 epochs before linearly decaying to $1e{-}7$ for the next 150 epochs. Zamir et al. (2021) and Chen et al. (2021) both utilize the Cosine annealing strategy of Loshchilov and Hutter (2016) to decay the learning rate to $1e{-}6$ and $1e{-}7$, respectively.

Data augmentation is scantly mentioned in the literature surrounding image restoration; however, some researchers report success from using data augmentation approaches. Nah et al. (2017), (Tsai et al. 2021), (Zamir et al. 2021), and (Chen et al. 2021) all perform random horizontal flips, vertical flips, and rotations during training to regularize their models against the spatial properties of natural images. Nah et al. (2017) additionally apply a noise augmentation model (see Equation 2.2) during training to regularize their model against subtle Gaussian noises that arise in digital images.

When validating against the GoPro testing data, most authors use models trained using only the GoPro training set; however, there are exceptions to this practice. Kupyn et al. (2018) use supplementary training data from the Microsoft Common Objects in Context (COCO) Lin et al. (2014) dataset that they synthetically blur using a uniform blur synthesis algorithm that resembles camera shakes. Kupyn et al. (2019) re-calculate the estimated blurs of the GoPro dataset by interpolating the data up to a higher sampling rate before generating the blurry training samples. They also train their models using a mixed training dataset containing samples from the GoPro dataset, the Deep Video Deblurring (DVD) dataset (Su et al. 2017), and the Need For Speed (NFS) dataset (Kiani Galoogahi et al. 2017). Gao et al. (2019) use the GoPro data and also utilize a bespoke dataset that is not publicly available. Finally, Zhang et al. (2020) use real-world blurred and sharp images in addition to GoPro data in their end-to-end model that learns to blur and to deblur. It is standard for researchers in this area to provide PSNR and SSIM metrics on the GoPro testing data, but some authors also elect to provide metrics on the

Human-aware Image DEblurring (HIDE) dataset or the REalistic and Dynamic Scenes (REDS) dataset in addition to the GoPro validation (Shen et al. 2019, Nah et al. 2019).

Model	Batch Size	Epochs	Learning Rate		Validation	
			Initial	Decay	REDS	HIDE
DeepDeblur (Nah et al. 2017)	2	9e5	1e−5	Exponential	✗	✗
DeblurGAN (Kupyn et al. 2018)	1	300	1e−4	Linear	✗	✗
SRN (Tao et al. 2018)	16	2000	1e−4	Exponential	✗	✗
DeblurGAN-v2 (Kupyn et al. 2019)	–	300	1e−4	Linear	✗	✗
Zhang et al. (2020)	4	–	1e−4	Linear	✗	✗
Gao et al. (2019)	16	4000	1e−4	Exponential	✗	✗
DMPHN (Zhang et al. 2019)	6	3000	1e−4	Exponential	✗	✗
SAPHNet (Suin et al. 2020)	6	–	1e−4	Exponential	✗	✓
RADNet (Purohit and Rajagopalan 2020)	16	1e6	1e−4	–	✗	✗
BANet (Tsai et al. 2021)	–	3000	1e−4	Linear	✗	✓
MPRNet (Zamir et al. 2021)	16	4e5	2e−4	Cosine	✗	✓
HINet (Chen et al. 2021)	64	4e5	2e−4	Cosine	✓	✗
This Study	4	150	1e−4	Exponential	✓	✗

Table 2.2: A comparison of training and validation parameters used by state-of-the-art deep learning-based image restoration models. A mark of "–" indicates that a value was not specified by the authors.

Chapter 3

Does Trust Influence Autonomous Vehicle Adoption?

In this chapter, we present the design of a perception augmentation module that demonstrably improves the driver's trust in the underlying AI technology. The perception augmentation module describes the visual surroundings of the AV to the driver using a video camera, an object detection model, and a screen. In this case, the object detection model is a deep learning algorithm that detects five generic classes of objects in the stream of images generated by the camera sensor. The detected objects are shown to the driver using a screen that could be placed on the dashboard of the vehicle. The screen contains an overlay with color-coded bounding boxes describing the objects that are detected in the scene by the AI. A LED strip provides an additional medium for providing alerts to the driver about objects in the scene but in more localized ways. Using a software-in-the-loop simulation and a psychological survey, we test whether the proposed system improves human trust in the ADS and whether improving trust truly impacts the intent to adopt autonomous features. The simulation and survey are administered remotely using the Qualtrics platform to collect data from 517 people.

This chapter validates that the perception augmentation module that we design improves the driver's trust in the ADS. We also show that trust in AI has a positive impact on the perceived benefits and a negative impact on the perceived risks of using the ADS. We also show that trust has a positive impact on the intention the adopt or use an ADS. The results demonstrate that auto-makers should consider the development of user-interface components like our perception augmentation module in their consumer ADS to improve the adoption and usage of their products.

The remainder of the chapter is organized as follows. In Section 3.1, we describe a framework for the design of human interfaces for ADS and the implementation of a perception

augmentation module for an ADS. We go on to stage the validation of the system through an experimental procedure in Section 3.2. A discussion of the results of the study can be found in Section 3.3.

3.1 Perception Augmentation Module Design

3.1.1 Framework

Following the extant conceptual model of an ADS, – as a collection of sensors, a perception module, a route planning module, and a vehicle control module (Badue et al. 2020) – we present a framework for understanding HCI in ADS. Figure 3.1 illustrates the framework for a User Interface (UI) layer between the human driver and each of the modules in the ADS. For each of the preexisting modules, we present an independent interface component responsible for relaying information to the driver, as well as responding to prompts from the passenger (e.g., by voice).

The first UI module, the *perception augmentation module*, receives as input the data from the perception module of the ADS, as well as the sensor data. The primary duty of this perception augmentation module is to improve the human driver's understanding of their surroundings through the sensors and computational models. One implementation of a perception augmentation module may be an alert system that flashes a light and emits a sound if the vehicle is approaching an object, like a pedestrian, at a dangerous rate. Another implementation of a perception augmentation system may wait for questions from the driver related to the physical surroundings of the car, like whether any pedestrians are nearby.

The second module in the ADS HCI framework is the *route guidance module*. As a downstream module, the route guidance component receives as input the data from the previous steps in the pipeline that are related to perception. It also receives the output data from the planning component of the ADS. This module's function is to provide the human driver with information related to the planned route for the vehicle. This could include functions like

alerting the driver when the route is changing due to traffic and/or weather conditions. The module could also contain reactive features, such as answering questions from the driver about the route. Many of the features that fall into this component can be found in contemporary Global Navigation Satellite System (GNSS) receivers and map applications.

The final component, the *control notification module*, is responsible for relaying information to the driver about the physical control of the vehicle. Much like the route guidance module, the control notification module is downstream and takes as input all prior data, as well as data from the control module in the ADS. An example function of the module would be explaining to the driver why the ADS may have performed unexpectedly or alarmingly. Because the definition of alarming is subjective, the component may also await prompts from the driver to explain behaviors that the driver personally found alarming or unconventional.

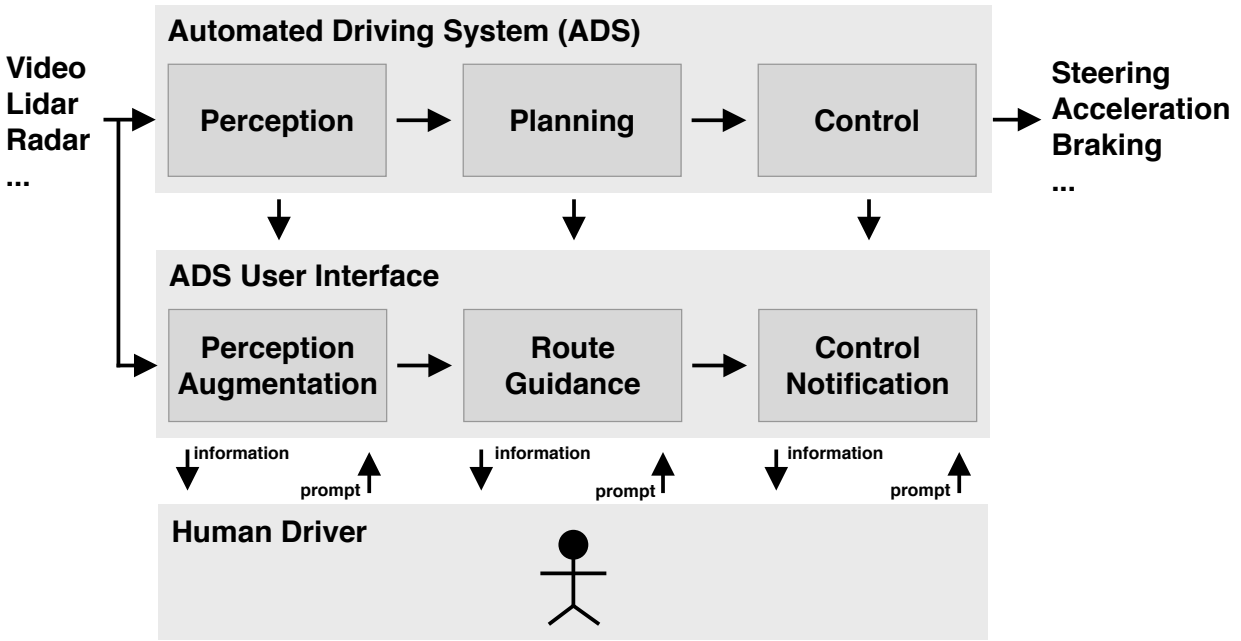


Figure 3.1: A framework for understanding Human-Computer Interaction (HCI) in Automated Driving Systems (ADS).

3.1.2 Perception Augmentation Module

This study presents an implementation of one component of the conceptual framework for HCI in an ADS (see Figure 3.1), namely, the perception augmentation module. We base the system on a previous prototype presented by Kauten et al. (2018). The proposed module observes the environment of an AV through the vehicle's sensors, aiming to show alerts to the driver based on the semantics of the scene. Our implementation uses a single camera mounted on the roof of the AV to capture frames of the scene in front of the vehicle at a resolution of 1280×720p and a capture rate of $30Hz$. Semantic meaning is parsed from singular frames using an object detection model that maps images to collections of rectangular regions with labels describing the object in the rectangle (e.g., "person", "banana", "bicycle"). Alerts are generated through two mechanisms, namely, (1) a video stream of the rooftop camera bearing color-coded rectangles, and (2) a color-coded LED strip around the windshield. It is worth noting that the LED strip is reserved for high-priority alerts about vehicles and pedestrians; the LED controller ignores data about generic objects, traffic signals, animals, and the like.

Figure 3.2 describes a high-level architecture of the proposed perception augmentation module. The architecture illustrates the collaborations among the components in terms of data flow. More specifically, the components include a roof-mounted camera, vision model, dashboard display, object space to LED space transformation function, and a windshield LED strip. The roof-mounted camera is an input device, whereas the dashboard display and windshield LED strip are both output devices. The intermediary modules are conceptual compute components of the underlying system.

Real-time pixel data acquired by the roof-mounted camera flows through the *vision model* to extract a set of labeled rectangles, the possible labels of which are: "generic object" (e.g., backpack, skateboard), "vehicle" (e.g., car, bus, truck), "pedestrian", "traffic signal" (e.g., stoplight, stop sign), and "animal". The labeled rectangles pass to a *dashboard display* to be overlaid onto the camera frame using a color code. The color-coded annotated frame is then shown to the driver using a display on the dashboard of the vehicle. The labeled rectangles

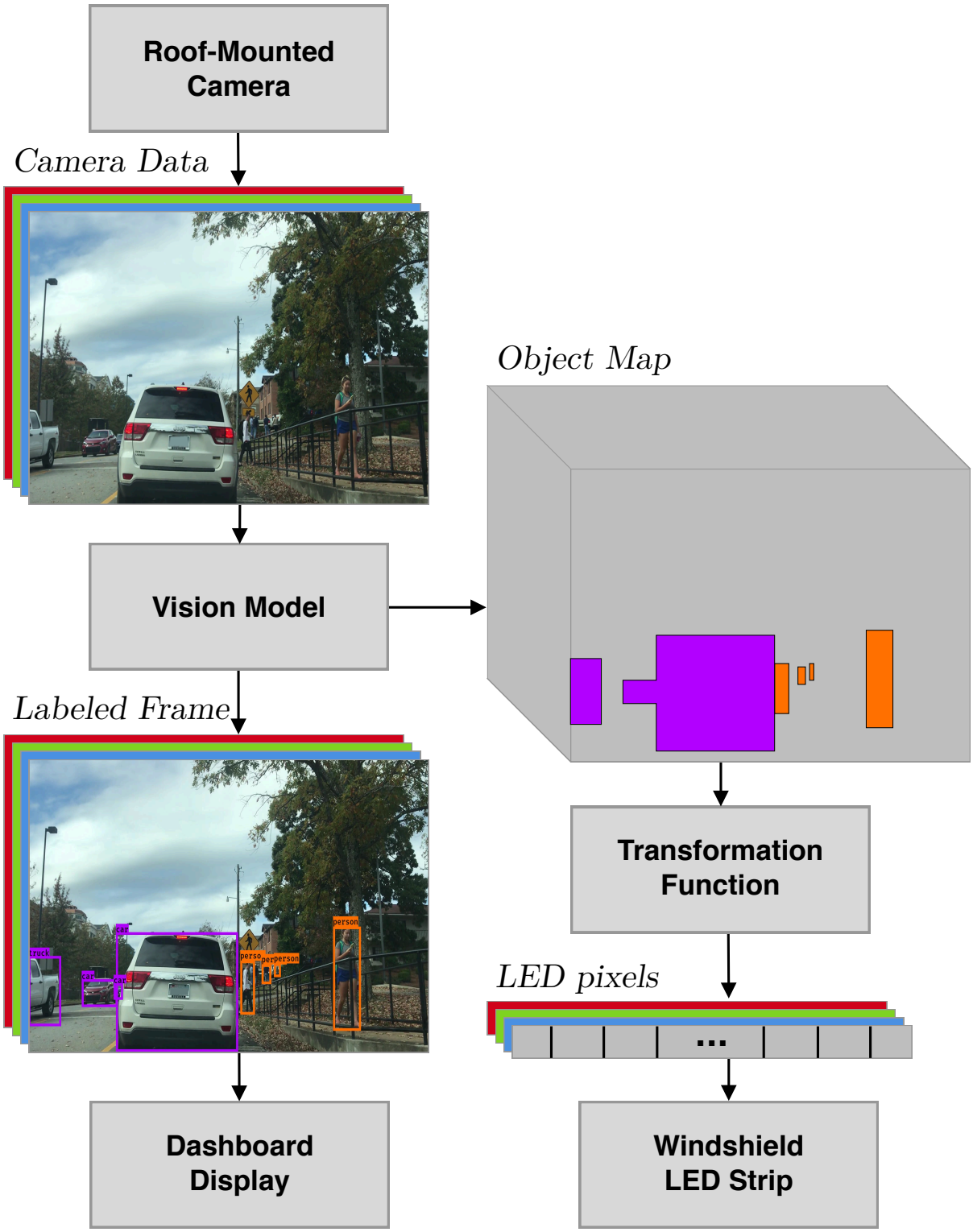


Figure 3.2: The high-level architecture and data flow of the proposed perception augmentation module.

also flow through an *LED controller* to be interpolated to coordinates around the windshield and shown to the driver using an LED strip.

3.1.3 Roof-Mounted Camera

We equip a vehicle with a rooftop camera to produce a view of the scene in front of the car. Although modern cameras are capable of producing high-resolution image data (i.e., 4K) at a fast rate of capture (i.e., $120Hz$), processing big data in a vehicle bears a significant cost for marginal gains. As such, the rooftop camera in our setup runs at 1280×720p resolution with a capture rate of $30Hz$. To simulate driving experiences from the perspective of the driver, we place a second camera (a.k.a., cockpit camera) inside the cabin to capture frames of the windshield from the driver's perspective. Compared with the rooftop camera, this cockpit camera captures data at a higher resolution of 1920×1080p and a faster refresh rate of $60Hz$ to produce a smooth and high-definition video. This second camera is for experimental purposes only (see Section 3.2.4), only the roof-mounted camera is necessary for the functionality of the perception augmentation module.

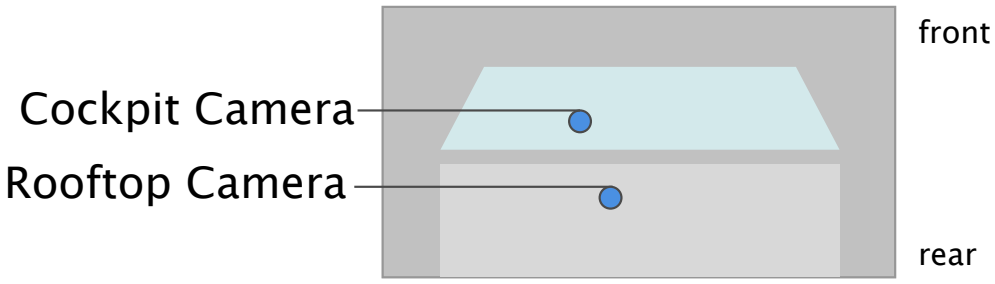


Figure 3.3: The layout of the roof-mounted and cockpit cameras installed on the vehicle.

3.1.4 Vision Model

The perception augmentation module leverages the vision model of the ADS to detect objects in the driving environment in the form of annotated rectangular regions. This particular strategy is commonly used in contemporary ADS (Schwarting et al. 2018). The vision model acts as a "black box" transformation function from pixel space to object space. For simplicity, we apply a temporally agnostic vision model that considers singular frames (i.e., images) as opposed to sequences of frames (i.e., videos).

In what follows, we briefly introduce the object-detection model implemented in the perception augmentation module. Let Eqn. 3.1 be the output domain $\mathbb{I}$ of the 1280×720p RGB rooftop camera.

$$\mathbb{I} = \mathbb{R}^{1280 \times 720 \times 3} \tag{3.1}$$

We refer to an object domain as $\mathbb{O}$ (Eqn. 3.2), where class c is one element in a predefined set of classes $\mathbb{C}$. We denote tuples (x, y) and (w, h) as (1) the starting point and (2) the dimensions of the rectangular region containing the object, respectively.

$$\mathbb{O} = \{c \in \mathbb{C}, (x, y), (w, h)\} \tag{3.2}$$

Now we are positioned to define the vision model as function f_{vision} in Eqn. 3.3, where n represents the total number of objects in a scene conforming to the semantics of the labels in set $\mathbb{C}$. In practice, it is impractical to exhaustively detect all objects in a scene. Therefore, we argue that n is bounded.

$$f_{vision} : \mathbb{I} \rightarrow \mathbb{O}^n \tag{3.3}$$

In practice, f_{vision} is approximated using deep learning algorithms Girshick et al. (2013), Redmon et al. (2015) trained on large large-object-detection datasets Lin et al. (2014). We commonly refer to the input to f_{vision} as the tensor $\mathbf{I} \in \mathbb{I}$ and the output as the matrix $\boldsymbol{O} \in \mathbb{O}^n$ – i.e., $\boldsymbol{O} = f_{vision}(\mathbf{I})$.

For our study, we implement f_{vision} using the "You Only Look Once" (YOLO) object detection model. Specifically, we apply the third version of the model, which achieves near state-of-the-art detection metrics while maintaining a higher throughput than competing models. YOLO relies on a CNN to map images from $\mathbb{I}$ to objects in $\mathbb{O}$. We obtain trained weights for the model directly from a prior study reported in Redmon and Farhadi (2018).

YOLO Object Detector

The "You Only Look Once" (YOLO) vision model is a neural network that automatically detects objects in an image. The architecture has been iterated over three times resulting in a highly accurate and computationally efficient detector; this section briefly discusses only the third version. For a detailed history, we refer readers to v1 Redmon et al. (2015), v2 Redmon and Farhadi (2016), and v3 Redmon and Farhadi (2018) in order.

Training the YOLO parameters relies on large datasets that are iterated over through back-propagation. The three distinct datasets studied in computer vision research include ImageNet (an object *classification* dataset), Visual Object Classes (VOC), and COCO (object *detection* datasets) Deng et al. (2009), Everingham et al. (2010), Lin et al. (2014). Although the abstract vision model presented in Section 3.1.4 represents objects using their direct dimensions, YOLO outputs dimensions as *offsets* from the dimensions of bounding box priors. These priors are extracted from the bounding boxes in the VOC and COCO datasets using a clustering algorithm, such as k-means.

YOLO is a fully convolutional neural network, meaning that the input and output shapes can dynamically change to accommodate different image sizes. YOLO typically uses images with size 416×416 to ensure the output dimension has an odd dimension – 13×13 in this case – so a single cell falls in the middle of the image. This is performed under the heuristic that larger objects typically occupy the center of an image.

The architecture contains two distinct parts, namely, a feature extractor and an object detector. The *feature extractor* applies a series of convolutional, pooling, and residual layers to reduce the image size by a factor of 32 and extract meaningful semantics. The *object detector*, a single convolutional layer, then maps the extracted features to object space in the form of (x, y) coordinates, (h, w) offsets from bounding box priors, an "objectness" score (i.e., the likelihood that the region contains *any* object), and a set of conditional class probabilities.

Figure 3.4 describes the output grid of YOLO with a hypothetical bounding box. It is worth noting, there are five bounding box priors, so there are five bounding boxes predicted for each

cell in the output grid. p_h and p_w refer to the height and width of the bounding box prior. YOLO predicts (t_h, t_w) to scale the prior dimensions along an exponential curve. Eqn. 3.4 and Eqn. 3.5 formulate how the bounding box priors p_h and p_w are scaled using the network outputs t_h and t_w to produce the bounding box dimensions b_h and b_w, respectively.

$$b_h = p_h e^{t_h} \tag{3.4}$$

$$b_w = p_w e^{t_w} \tag{3.5}$$

The YOLO algorithm directly predicts the center of the box (t_x, t_y) – as opposed to predicting an offset from a prior – relative to the location of the grid cell whose upper left corner is at pixel (c_x, c_y). Eqns. 3.6 and 3.7 describe the calculation of the center of the bounding box (b_x, b_y) from (c_x, c_y) and (t_x, t_y).

$$b_x = \sigma(t_x) + c_x \tag{3.6}$$

$$b_y = \sigma(t_y) + c_y \tag{3.7}$$

In addition to the bounding box, the network produces an "objectness" score as the probability of the bounding box containing *any* object (i.e., $P(object)$) multiplied by the Intersection over Union (IoU) between the predicted and ground truth bounding box. During inference, this score reduces to just $P(object)$ as the IoU is undefined for unknown targets. Lastly, the model estimates a series of class conditional probabilities (i.e., $P(c|object) \forall c \in \mathbb{C}$) that determine the likelihoods of the bounding box containing specific object classes.

YOLO is trained using a combination of ImageNet, COCO, and VOC datasets. This strategy allows the model to effectively detect over 1000 object classes. The datasets are combined using a concept graph to link the disparate topics between datasets – i.e., "golden retriever" in ImageNet is also "dog", which is also "animal" in COCO, and so on. The implementation of the training procedure escapes the scope of this work. We refer readers to Redmon et al.

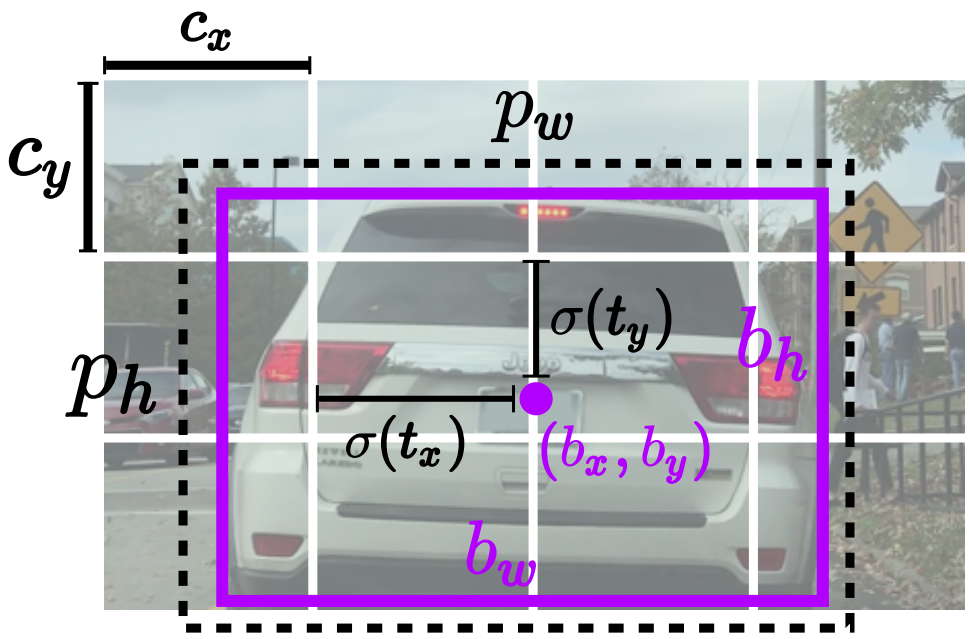


Figure 3.4: A hypothetical YOLO output grid with a bounding box output for a given prior.

(2015), Redmon and Farhadi (2016), and Redmon and Farhadi (2018) for the full training algorithm.

3.1.5 Object Color Map

Before describing the dashboard display and the LED controller, we articulate an object color map based on a generalization set, $\mathbb{C}'$, of the class set $\mathbb{C}$. This simplification focuses on five core object groups, like "vehicle", as opposed to specific objects, like "car", "truck", and the like, thereby making the results of the model easier to interpret in fast-paced environments. Figure 3.5 outlines the color-coding in terms of the five generic class groups in $\mathbb{C}'$: "generic object", "vehicle", "pedestrian", "traffic signal", and "animal". "generic object" describes skateboards, bananas, and the like; "vehicle" encompasses cars, trucks, buses, and motorcycles; "pedestrian" stands for people and bicyclists; "traffic signal" represents traffic lights and stop signs; and "animal" denotes non-human beings like birds, dogs, and cats, just to name a few.

It is prudent to carefully select colors in the color map and the reason is two-fold. First, we ensure that colors are aligned with preconceived symbolic meanings for better integration with existing driving knowledge. For example, the color orange symbolically represents the concept of "caution". Second, we advocate that colors intended for externally visible features must not accidentally imitate emergency vehicles or other traffic signals. For instance, the windshield LED strip in our module may be seen by other drivers and should not be misconstrued as traffic signals or emergency vehicle signals. We assign "generic objects" a subdued gray color due to their low significance to the driving task. Importantly, "pedestrians" are assigned bright safety orange, similar to U.S. road construction signs, to convey a similar message of caution for surrounding human beings. "Traffic signals" are designated a red color in a similar vein of thought; stop signs are red and stoplights are easily associated with red. There is no singular color that objectively provides a symbolic representation of motor vehicles or animals. As such, we assign "vehicle" and "animal" a purple color and green color, respectively.

Figure 3.5: The color-coding for the generalized object classes.

3.1.6 Dashboard Display

To provide a real-time feed of the performance of the ADS perception component, a simple dashboard display shows the outputs of the vision model combined with the image data from the rooftop camera (see also Figure 3.2). The controller of the dashboard display uses the color-coding described by Figure 3.5 to color each of the annotated rectangular regions output by the vision model. These colored regions are then overlaid with the image to produce an image of the scene with colored and labeled rectangles surrounding the detected objects.

Figure 3.6 illustrates an example output from the dashboard display. Similar to how previous works have visualized object space $\mathbb{O}$, the display renders detected objects in the form of bounding boxes around the pixels that illustrate the object. The color of each box is defined by the class label c of the object in the box (i.e. "vehicle") and the symbolic color map (see Section 3.1.5). A string label describing the object's specific class is attached to the top left corner of the box. This label serves both to reinforce the color-map, and to provide more detailed information about the object than the color alone – e.g., "car" or "truck" as opposed to the general class, "vehicle".

Our dashboard display logic is simple yet effective, arguably lacking the aesthetic polish of modern consumer software. For instance, scenes cluttered with an excessive number of objects can be hard to understand due to the many overlaying boxes (see the clump of pedestrians in Figure 3.6). The simple design is furnished intentionally to reduce the number of independent variables of our study. Future work may explore how to improve the aesthetic of the display to be easily and intuitively parsed by the human driver in object-dense scenes.

Figure 3.6: An example object detection output.

3.1.7 Windshield LED Strip

An LED strip provides a mechanism to direct the attention of the driver to areas of interest on the windshield. Figure 3.7 illustrates that an LED strip with seven separate zones (i.e., five on the top and one on each side) lines the windshield of a vehicle. The LED strip conveys alert data only for "pedestrian" and "vehicle" classes (i.e., $\mathbb{C}'$), which are treated as mission-critical classes. To generate the signal for the LED strip, the annotated regions from the vision model are aggregated into a tensor of cumulative class confidence scores $\mathbf{L} \in \mathbb{R}^{1280 \times 1920 \times |\mathbb{C}'|}$ described by the iterative process in Algorithm 1. The transformation between the pixel coordinates (i.e., object coordinates) from the camera perspective and LED strip coordinates from the perspective of the driver is nontrivial; however, we assume the transformation between these coordinate planes is reasonably approximated by a straight line. As such, we apply a simple linear interpolation technique to resize the tensor $\mathbf{L}$ to a matrix $\boldsymbol{L} \in \mathbb{R}^{l \times |\mathbb{C}'|}$ where l is the number of LED zones around the windshield – i.e., $l = 7$ in Figure 3.7. We define an ad-hoc threshold for determining whether an LED zone should be illuminated. It is worth noting, "pedestrians"

and “vehicles” are given a weight of 0.29, and “void” is given a weight of 0.3 where void is any class that is neither a vehicle nor a pedestrian.

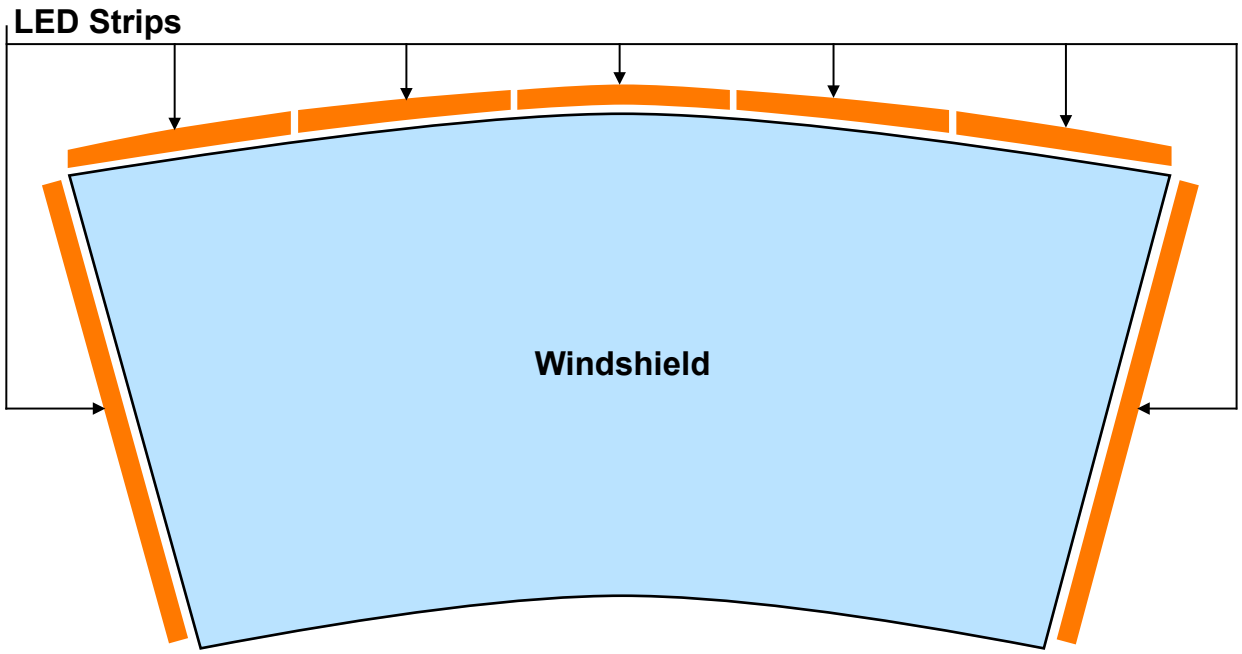


Figure 3.7: A windshield LED strip divided into seven zones.

Algorithm 1: Object Space to LED Space Transformation Function

$\boldsymbol{O}$: The set of objects from the object detection model
$\boldsymbol{w}$: The actuation weights for the various classes

$\mathbf{L} \sim 0 * \mathbb{R}^{1280\times1920\times|\mathbb{C}'|}$ // initialize L with zeros
$\mathbf{L}_{:,:,\text{void}} += \boldsymbol{w}_{void}$
foreach $\boldsymbol{o} \in \boldsymbol{O}$ **do** // accumulate scores for all objects
  $\mathbf{L}_{x_o:x_o+w_o,y_o:y_o+h_o,c_o} += P(c_o)$
end
$\boldsymbol{L} \leftarrow interpolate(\mathbf{L})$ // transform from camera space to LED space
foreach $\boldsymbol{c} \in \mathbb{C}'$ **do** // apply a weight to each class channel
  $\boldsymbol{L}_{:,c} += \boldsymbol{w}_c$
end
$\boldsymbol{l} \leftarrow \arg\max_{\boldsymbol{c}\in\mathbb{C}'} \boldsymbol{L}_{:,c}$ // get class with highest P for each LED
return $\boldsymbol{l}$ mapped to RGB space using the color-coding

The design of the LED strip is still in its infancy. Our windshield LED strip is the first of its kind aiming at capturing the attention of the driver. We don’t imply by any means that our design is an optimal one in terms of layout, format, and number of zones. It is arguably true

that the windshield LED strip might be (1) substituted by an augmented reality technology (e.g., Google Glass (Muensterer et al. 2014)) or (2) integrated into a media center (e.g., Android Auto (Udovicic et al. 2015)). Our LED strip design opens a door to explore advanced mechanisms for directing the attention of the driver in autonomous vehicles. Regardless of the various potential implementations, data acquired and managed by our proposed LED strip controller remains unchanged.

3.2 Psychological Study

To validate the designed system, a research model is designed and a simulation study is conducted. In this section, we first describe the theoretical foundation for the study including a brief review of Social Contract Theory (SCT) and an extension of SCT to the adoption of AVs. On this theoretical basis, we go on to introduce a research model and series of hypotheses about how the designed system (see Section 3.1) will affect human drivers. To measure the effectiveness of the perception system on human participants through the lens of our research model, we go on to design a simulation environment and questionnaire. We finish the section with a presentation of data collected from two rounds of trial studies and a final study.

3.2.1 Theoretical Development

This study uses SCT as the theoretical lens to understand the initial consumer acceptance of ADS. In the below subsections we first review the philosophical roots of SCT and build the connections with our research context. Then, we further extend the social contract model of health IT (Li et al. 2010) to the context of ADS adoption by considering the effects of trust in AI, joy, personal innovativeness, perceived risks, perceived benefits, and the perception augmentation system.

Social Contract Theory (SCT) posits that there exists an implicit social contract governing the relationship between two parties in situations involving uncertainty (Dunfee et al. 1999).

The core assumption of SCT is that individuals are subject to bounded moral rationality, i.e. "individual moral agents lack the information, time, and emotional strength to make perfect judgments" (Donaldson and Dunfee 1994, p. 18). AI systems supporting an ADS are very complex and difficult to understand. As such, AI is criticized for being non-transparent and is comparable to a "black box". AI not only leads to less transparency in information but also challenges the basic human need for control. In an autonomous mode, users essentially relinquish some or all their control to the ADS in exchange for the benefits of autonomy. Thus, the adoption of ADS comes with inherent uncertainty due to the lack of transparency and reduced control, demanding the existence of an implicit social contract to govern the relationship between users and ADS.

A social contract entails implicit norms defining the rights and responsibilities of two parties. The specific norms embedded in a social contract vary from context to context. SCT has been applied in many different contexts such as market exchange (Dunfee et al. 1999), personal information disclosure (Li et al. 2010), adoption of health IT (Li et al. 2014), and technology governance by block-chain (Reijers et al. 2016). Li et al. (2014) contend that individual intention to adopt health IT is driven by a trust-enabled social contract that governs the relationship between patients and health IT. They suggest that 1) the core of the social contract involves a cost-benefit trade-off analysis, and 2) trust and sufficient benefits are two key factors for individuals to enter a social contract with embryonic health IT to overcome the potential privacy risks of patients.

3.2.2 Extending the Social Contract Model of Health IT to AV Adoption

In this study, we further extend the social contract model of health IT proposed by Li et al. (2014) to the context of the ADS adoption to develop the social contract model involving new IT artifacts. In particular, we argue that the adoption of ADS is driven by a trust-enabled hedonic social contact between drivers and AI systems of ADS. The ADS users rely more on the underlying AI systems than the traditional car. Trust in AI could play a key role in mitigating

drivers' perceived risks and explaining people's intention to use an ADS. Additionally, the social contract between drivers and AI systems of ADS have a hedonic nature since ADS could bring enjoyment to users (Raue et al. 2019). ADS can be considered as a mixed system since it serves both utilitarian (e.g. improved safety and fuel efficiency) and hedonic (i.e. enjoyment) purposes. For systems with hedonic value, perceived enjoyment has been suggested to be a strong predictor of individuals' adoption intention (van der Heijden 2004). In situations with uncertainty, emotions such as enjoyment can also act as important information cues for individuals to evaluate the risk and benefits of using the systems (Li et al. 2011). Therefore, we incorporate perceived enjoyment as another important lever influencing an individual's decision to adopt ADS.

Besides developing the social contract model of IT, our study also attempts to explore mechanisms for cultivating trust in AI and joy. The lack of trust in the ADS underlying AI system may be largely attributed to the complexity of such AI, which makes them opaque to drivers. According to trust literature, first-hand knowledge is an important mechanism for building trust (Gefen et al. 2003). Therefore, we designed and implemented an alert system, also called the perception augmentation module, to increase the transparency of the AI of ADS. The alert system is a module that augments drivers' perception about how the underlying AI works, which may serve as a knowledge-based antecedent for promoting trust in AI. Additionally, personal innovativeness, as a personal propensity to try out new technologies (Agarwal and Prasad 1998), is pertinent to our research context for explaining drivers' trust in ADS and perceived enjoyment. Those with high personal innovativeness have been suggested to increase trust in IT (Schweitzer and van den Hende 2016) and more easily enter the state of cognitive absorption that exhibits partly through heightened enjoyment (Agarwal and Karahanna 2000).

From the above, we posit that the decision to adopt ADS is influenced by a trust-enabled hedonic social contract with cost-benefit trade-off analysis, trust in AI, and perceived enjoyment being three vital components. The perception augmentation module and personal innovativeness are important external factors that help cultivate trust in AI and perceived enjoyment.

3.2.3 Research Model and Hypotheses

The research model (Figure 3.8) shows how individuals' willingness to use an ADS is driven by a cognitive assessment of risks and benefits embedded in a trust-enabled hedonic social contract. We postulate that individuals' willingness to use an ADS could be increased by 1) increasing people's trust in AI, 2) enhancing people's perceived enjoyment from using an ADS, and 3) providing adequate benefits such as fewer traffic accidents and lower stress related to driving. Our model also suggests that trust in AI can be elevated by the add-on perception augmentation module and personal innovativeness.

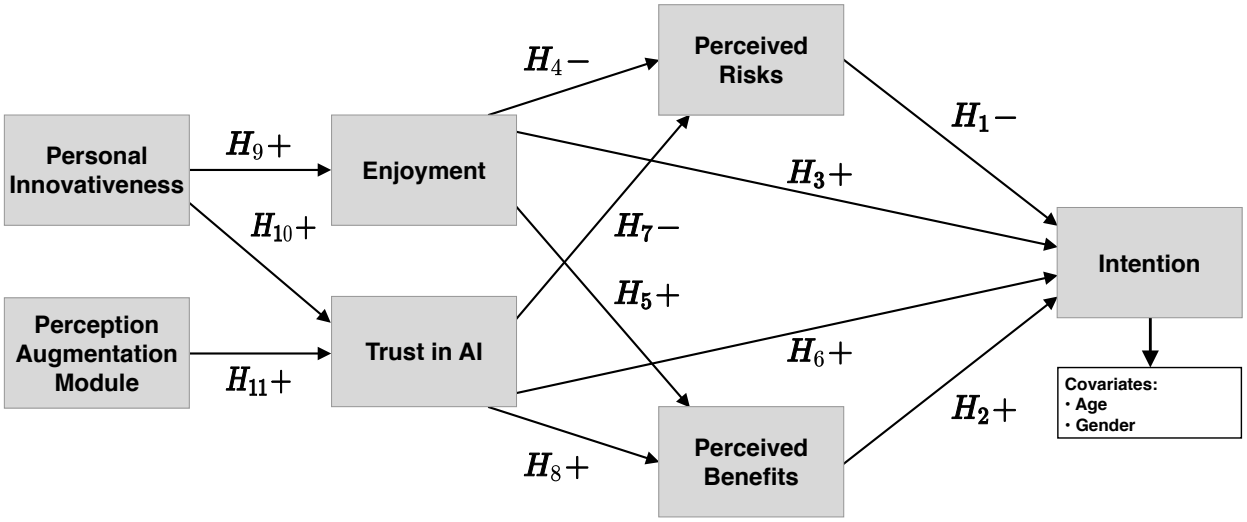


Figure 3.8: The research model describing the constructs and hypothesized paths.

Perceived Risk, Perceived Benefit, and Behavioral Intention ADS may bring various risks to drivers, including performance risk, safety risk, psychological risk, and social risk. The performance of an ADS is influenced by hard-to-predict road conditions and bad weather such as heavy rain or snows, thus exposing drivers to performance and safety risks (Guo et al. 2019). People may also perceive psychological risk in terms of reduced self-image and social risk of embarrassment before one's social group (Luo et al. 2010). An individual may think that an ADS does not fit his or her self-image and could even negatively influence how (s)he is

viewed by others. Drivers who perceive high risks from ADS will be more cautious in their ADS adoption decisions. Therefore,

H1: Perceived risk has a negative impact on drivers' intention to use ADS.

On the other hand, theories behind ADS suggest that the technology can reduce the number of deaths on the road, increase the amount of free time the driver of the vehicle has, and improve the flow of traffic on public roads (Chan 2017). To a certain extent, these perceived benefits resolve challenges that all drivers face on the road (in varying degrees based on their geographic location, etc.). Because these benefits make ADS appear more useful, we posit that

H2: Perceived benefit has a positive impact on drivers' intention to use ADS.

Perceived Enjoyment and Behavioral Intention Perceived enjoyment is the extent to which a system is perceived to be enjoyable in its own right, apart from any anticipated performance consequences (Lowry et al. 2013). van der Heijden (2004) emphasized the important role of perceived enjoyment in influencing individuals' decisions to use technology with hedonic value. A considerable number of studies have found that perceived enjoyment increases people's intention to use technology with hedonic value (Agarwal and Karahanna 2000, Teo and Noyes 2011, van der Heijden 2004). Considering the enjoyment value associated with the use of ADS (Raue et al. 2019), we posit that

H3: Perceived enjoyment has a positive impact on drivers' intention to use ADS.

Perceived Enjoyment and Cost-benefit Analysis In uncertain situations, emotions tend to influence people's judgment in a congruent manner such that positive emotions (e.g. joy) are often associated with more positive judgment and lower negative judgment (Forgas 1995). Similarly, perceived enjoyment, reflecting the extent of joyfulness one expects to glean from using technology, may also exert a congruent effect on their judgment. Such a congruent effect of perceived enjoyment has been supported in prior studies. For example, perceived enjoyment was found to increase perceived ease of use and perceived usefulness (Venkatesh

2000, Venkatesh et al. 2002), perceived benefits of online payment systems (Rouibah et al. 2016), and reduce perceived risks of mobile banking (Koenig-Lewis et al. 2015). Due to the embryonic and "black box" nature of ADS, consumers often do not have complete information to evaluate the costs and benefits of ADS. Thus, those who perceive more enjoyment from driving ADS are expected to form a more favorable judgment about ADS. Therefore, we posit that

H4: Perceived enjoyment has a negative impact on perceived risks from using ADS.

H5: Perceived enjoyment has a positive impact on perceived benefits from using ADS.

Trust in AI and Behavioral Intention In this study, trust in AI refers to one's belief that the AI in ADS is robust and provides the necessary safeguards to protect drivers. It corresponds to the structural assurance component in institution-based trust (McKnight et al. 2002), reflecting one's general trust belief toward AI enabling ADS. Structural assurance has been suggested to increase one's trusting intention. In the context of e-commerce, trust in the general Internet environment was proposed to positively influence consumers' intention to follow advice, disclose persona information or make purchases (McKnight et al. 2002). Similarly, drivers who trust in the AI of ADS would be more willing to depend on ADS. Thus, we have

H6: Trust in AI underlying ADS has a positive impact on intention to use ADS.

Trust in AI and Cost-benefit Analysis In line with the social contract model of health IT by Li et al. (2014), trust in AI may also indirectly influence behavioral intention through modifying the cost-benefit trade-off analysis embedded in the social contract. Trust plays a key role in mitigating perceived risks in situations involving uncertainty (McKnight et al. 2002). Trust in the wireless Internet platform has been found to alleviate perceived risk in mobile banking services. Likewise, drivers who trust in AI supporting ADS are expected to be more likely to overcome their perceived risks of using ADS. In addition to risk mitigation, trust could also serve as one form of subjective guarantee for increasing the chance for people to attain the

potential benefits in a situation (Luo et al. 2010). As a result, those who trust in AI are likely to form more favorable perceptions about the benefits of ADS. Therefore, we posit that

H7: Trust in AI has a negative impact on the perceived risk of ADS.

H8: Trust in AI has a positive impact on the perceived benefit of ADS.

Personal Innovativeness and Perceived Enjoyment Personal innovativeness reflects an individual's disposition to try out new technologies (Agarwal and Prasad 1998). Empirical evidence suggests that those with high personal innovativeness tend to have more favorable perceptions about new technologies (Rouibah et al. 2016). It was found to increase perceived enjoyment in using online payment systems (Rouibah et al. 2016), and mobile video calling in a leisure context (Zhou and Feng 2017). ADS are an embryonic innovation enabled by AI technology. Highly innovative people are likely to perceive the use of ADS to be more enjoyable.

H9: Personal innovativeness has a positive impact on perceived enjoyment from using ADS.

Personal Innovativeness and Trust in AI Highly innovative people tend to be more curious and willing to take the risk from using new technologies than those with low personal innovativeness (Schweitzer and van den Hende 2016). Due to its association with curiosity and risk-taking propensity, personal innovativeness has been suggested to increase people's trust in innovative technologies such as health-monitoring devices and smartphone apps (Schweitzer and van den Hende 2016). Likewise, we submit that personal innovativeness helps build people's trust in AI underlying ADS.

H10: Personal innovativeness has a positive impact on trust in AI.

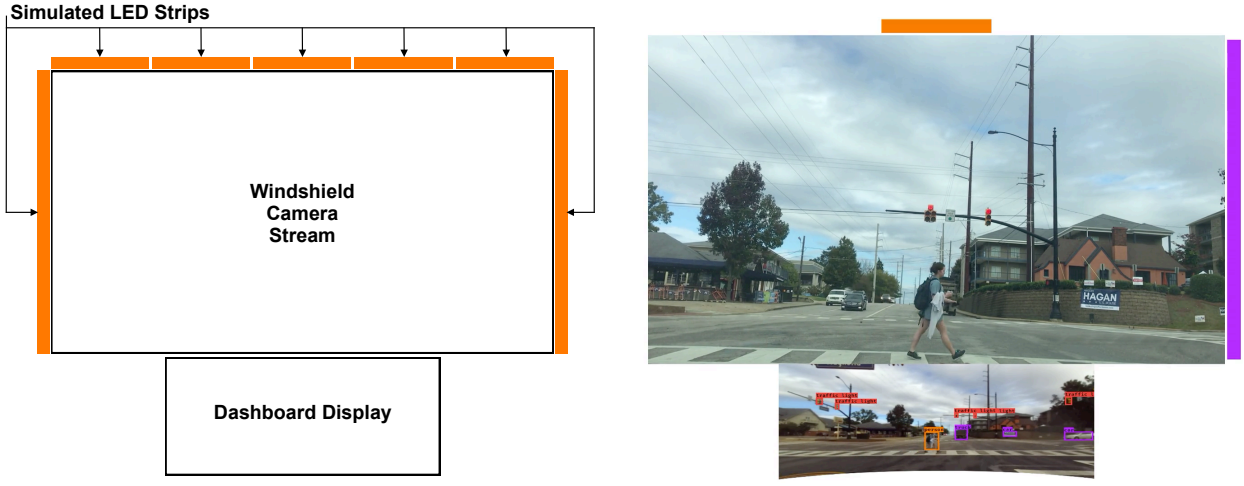
Perception Augmentation Module and Trust Extant studies have shown that the technological opacity of the AI that powers ADS results in reduced trust from the drivers that use these systems (Kalra and Paddock 2016, Kyriakidis et al. 2015). The add-on perception augmentation module developed in this study attempts to open the "black box" AI, which helps users better understand how AI that enables ADS works. Personal experience or knowledge has been

suggested as one of the most fundamental mechanisms for building trust (McKnight et al. 2002). Therefore, we argue that the perception augmentation module increases users' trust in AI through imparting knowledge or building user experience.

H11: Perception augmentation module has a positive impact on trust in AI.

3.2.4 Research Methodology

We conduct extensive experiments to validate the proposed hypotheses. Because it is costly, potentially dangerous, and prohibitively difficult to engage human subjects to validate the system in the field in an ADS, we design a simulated environment to deploy online to remote participants. The simulation takes the form of a video where a viewer (i.e., driver) assumes the position of the driver of the ADS performing a monitoring task. Figure 3.9 depicts our simulated ADS cockpit, which is comprised of three primary components, namely, (1) the windshield camera stream, (2) the dashboard display, and (3) the simulated LED strips. The windshield camera stream shows the frames captured by a windshield camera (see Figure 3.2). The dashboard display shows the output of the dashboard view controller as described in section 3.1.6. Finally, the simulated LED strip shows the data output by the LED controller (see Section 3.1.7) using groups of pixels as a model of an LED strip.



(a) The abstract layout of the simulator.

(b) An example frame from the simulator.

Figure 3.9: The design of the perception augmentation module simulator.

Although the LED strips and dashboard display are simulated, the videos displayed in the simulated windshield and dashboard are real-world driving scenarios captured by a cockpit camera and a rooftop camera mounted on the autonomous vehicle (i.e., Lincoln MKz). More specifically, the windshield videos were recorded by the cockpit camera, whereas the dashboard videos were acquired by the rooftop camera (see Section 3.1.3). Driving data are recorded directly from the vehicle while driving in the real world.

The driving scene data is organized by the subjective "density" of vehicles and pedestrians in the scene – i.e., the number of pixels in the scene that belong to either object class. Specifically, we create four distinctive DDT datasets based on object density, namely, (1) low vehicle and pedestrian density, (2) high vehicle and pedestrian density, (3) low vehicle and high pedestrian density, and (4) high vehicle and low pedestrian density. Separating the data into distinct situations allows us to study the effect of the perception module under a wide variety of driving conditions. It is worth noting, the densities are so-called subjective because the four scenarios are selected by hand from a larger collection of driving scene data generated by the data collection vehicle. This is a limitation imposed by the large size of the data collected by the car coupled with the rural nature of the tested driving environment.

Procedure

We relied on experimental design and an online survey to test our research model. The availability of perception-based alert systems was manipulated at two levels, i.e. with and without alert systems. The alert system consists of a dashboard display and a light strip, providing alerts of vehicles and pedestrians detected around the front and sides of the vehicle.

The first page of our online survey provides the informed consent form. The second page consists of two filter questions about whether they agree to fill out the survey honestly and whether they have any prior experience of driving a car or other types of automobiles. Only subjects who answered "Yes" to both filter questions are allowed to proceed to the next page showing an introduction video about the alert systems in an AV. The next survey page, exhibited

> A **Self Driving Car (SDC)** relies on an **Artificial Intelligence (AI)** system to function. The AI system collects data from sensors and cameras to make decisions in various driving situations. Typically, such decisions are made by the human driver in a traditional car. The control of SDC can be interrupted and taken over by the human driver at any time.
> One of the key components of the AI system is an alert system. The below video provides a tutorial of the alert system. Please click on the play button in the middle of the video to watch the video. Once you finish watching, click the "Next" button to move to the next survey page.

Figure 3.10: The survey scenario that participants read before embarking on the study. Here we use the less technical term "Self Driving Car (SDC)" to refer to an ADS or AV for ease of communication to a non-technical audience.

in Figure 3.10, provides a short description of the AI system of an ADS and prompts subjects to watch a short-minute video clip recorded in an AV. Here each subject was randomly assigned to one of the two treatment groups, i.e. with and without alert systems. Subjects assigned to the group with alert systems saw outside street view through the front window together with the alert systems. Those assigned to the group without alert systems only saw the street view. All subjects were required to imagine themselves as the driver of an AV while watching the video clip. After watching the manipulation video clip, subjects answered a manipulation check question on whether the video provided the alert systems. This manipulation question also serves as a filter question in the final study such that only those who answered correctly are allowed to work on the rest of the online survey.

Before the final study, we conducted two rounds of pilot studies. The first round of pilot study yielded 50 complete responses from students in STEM programs in a university in the southeastern U.S. and a university in the southwestern U.S. The results from the first round were used to check the effect of video manipulation and the wording of survey questions. We received many useful comments and suggestions about the color precision of dashboard display in the alert systems and grammar errors and wording ambiguity in some of the questions. The original version of the alert systems was based on a semantic segmentation algorithm, which gives no explicit boundaries of objects such as cars or pedestrians. Following student

comments, we recreated the alert systems using an object detection algorithm, which clearly outlined different types of objects. We also provided an introduction video about the alert systems and improved the wording clarity of questions in the survey.

In the second round of the pilot study, we collected 65 complete responses from the same two universities as those in the first pilot study and 45 subjects from the Qualtrics research panel. The panelists were randomly contacted by Qualtrics and stayed anonymous to the researchers. Qualtrics applies sophisticated digital fingerprinting technology to ensure all survey responses are from different subjects. The demographic profile of subjects in the two pilot studies is provided in Table 3.1.

Gender		Age		Driving Experience (yr.)		Furthest Education	
Male	58.8%	$18-24$	41.3%	≤ 1	6.9%	High school	15.0%
Female	41.2%	$25-34$	35.0%	$2-4$	15.0%	Professional school	1.9%
		$35-44$	15.0%	$5-10$	40.0%	Undergraduate	37.5%
		$45-54$	5.6%	$11-15$	11.3%	Graduate	39.4%
		$55-64$	3.1%	>15	26.9%	Doctoral	6.3%
		>64	0.0%				

Table 3.1: The demographic distribution of survey respondents from two pilot studies.

In the second round of the pilot study, we did not find any issues with the experiment and the survey questions. In the final data collection, we only changed the manipulation check question to be a filter question so only valid responses to this question were retained. In the end, we received a total of 517 usable responses (143 males and 374 females). The demographic profile of respondents in the final study is summarized in Table 3.2. The survey respondents in the final study are older than student subjects and are mostly female.

Covariates

Besides the core constructs in the research model, we also controlled for age, gender, and driving experience, which may influence people's willingness to use an AV in the future. For example, younger people may be more willing to use an AV than elderly people.

Gender		Age		Driving Experience (yr.)		Furthest Education	
Male	27.7%	$18-24$	4.8%	≤ 1	0.8%	High school	30.4%
Female	72.3%	$25-34$	18.6%	$2-4$	4.8%	Professional school	10.6%
		$35-44$	19.7%	$5-10$	8.3%	Undergraduate	31.9%
		$45-54$	18.8%	$11-15$	11.0%	Graduate	25.0%
		$55-64$	21.3%	>15	75.0%	Doctoral	2.1%
		>64	16.8%				

Table 3.2: The demographic distribution of survey respondents in the final study.

Variable measurement

We primarily used existing validated scales in prior research with slight modifications to fit the context. The scale measuring perceived benefits was developed based on the results of a recent survey on consumer opinions on automated vehicles conducted by Bosch LLC. (2019). Perceived benefit was operationalized as a formative first-order instrument while other first-order scales were operationalized as reflective instruments. All these core constructs were measured on a seven-point scale. The detailed items in each core construct are provided in Table 3.3.

Joy (Adapted from Lowry et al, 2013)		
JOY1	I would feel enjoyment while driving the SDC.	1 2 3 4 5 6 7
JOY2	I would have fun using the SDC.	1 2 3 4 5 6 7
JOY3	Using the SDC would be pleasant to me.	1 2 3 4 5 6 7
Trust in AI Systems (after (McKnight et al., 2002) (Strongly Agree/Strongly Disagree)		
TAI1	The AI systems has enough safeguards to make me comfortable using the SDC.	1 2 3 4 5 6 7
TAI2	I feel assured that the AI systems would adequately protect me from accidents on the road.	1 2 3 4 5 6 7
TAI3	I feel confident that the AI systems makes it safe for me to use the SDC.	1 2 3 4 5 6 7
TAI4	In general, the AI systems provides a robust and safe environment for me to use the SDC.	1 2 3 4 5 6 7
Performance Risk (Modified after Luo et al. DSS paper).		
PFR1	The SDC might not perform well and create problems while driving. (1-strongly disagree, 4-not sure either way, 7-strongly agree)	1 2 3 4 5 6 7
PFR2	The safety features built into the SDC might not strong enough to protect me. (1-strongly disagree, 4-not sure either way, 7-strongly agree)	1 2 3 4 5 6 7
PFR3	What is likelihood that there will be something wrong with the performance of the SDC or that it will not work properly? (1-low, 7-high)	1 2 3 4 5 6 7
PFR4	Considering the expected level of performance of the SDC, for you to purchase and use it would be ___. (1-Not risky at all, 7- risky)	1 2 3 4 5 6 7
PFR5	SDC may not perform well and process road information incorrectly. (1-strongly disagree, 4-not sure either way, 7-strongly agree)	1 2 3 4 5 6 7
Safety Risk		
SFR1	Considering the place and time you use a SDC, what are the chances that you stand to safety risk? (1- low, 7-High)	1 2 3 4 5 6 7
SFR2	My using a SDC would expose me to increased safety risks related to traffic accidents. (1-strongly disagree, 4-not sure either way, 7-strongly agree)	1 2 3 4 5 6 7
SFR3	There would be high potential for safety risks associated with my using a SDC.	1 2 3 4 5 6 7
Psychological Risk (Modified after Luo et al. DSS paper).		
PSR1	Driving a SDC will not fit in well with my self-image or self-concept. (1-strongly disagree, 4-not sure either way, 7-strongly agree)	1 2 3 4 5 6 7
PSR2	The usage of a SDC will lead to a psychological loss for me because it would not fit in well with my self-image or self-concept. (1-Improbable, 7-probable)	1 2 3 4 5 6 7
Social Risk (Modified after Luo et al. DSS paper).		
SCR1	What are the chances that using SDC will negatively affect the way others think of you? (1- Low, 7-high social risk)	1 2 3 4 5 6 7
SCR2	My usage of SDC would lead to a social loss for me because my friends and relatives would think less highly of me. (1-Improbable, 7-probable)	1 2 3 4 5 6 7
Perceived Benefits (Developed for this study) (Strongly Agree/Strongly Disagree)		
PB1	To me, using SDC would result in fewer traffic accidents.	1 2 3 4 5 6 7
PB2	To me, using SDC would result in more free time on the road.	1 2 3 4 5 6 7
PB3	To me, using SDC would help reduce driving-related stress.	1 2 3 4 5 6 7
PB4	To me, using SDC would help improve fuel economy.	1 2 3 4 5 6 7
PB5	To me, using SDC would improve my productivity due to the free time while driving.	1 2 3 4 5 6 7
Personal Innovativeness in IT (Agarwal and Prasad, 1998a, 1998b). (Strongly Agree/Strongly Disagree)		
PI1	If I heard about a new information technology, I would look for ways to experiment with it.	1 2 3 4 5 6 7
PI2	Among my peers, I am usually the first to try out new information technologies.	1 2 3 4 5 6 7
PI3	In general, I am hesitant to try out new information technologies.	1 2 3 4 5 6 7
PI4	I like to experiment with new information technologies.	1 2 3 4 5 6 7
Intention (Venkatesh et al, MIS Quarterly 2003) (Strongly Agree/Strongly Disagree)		
INT1	I intend to use SDC in the near future.	1 2 3 4 5 6 7
INT2	My general intention to use SDC is very high.	1 2 3 4 5 6 7
INT3	I will think about using SDC in the near future.	1 2 3 4 5 6 7

Table 3.3: The survey instrument for collecting data from the participants of the study.

3.2.5 Data Analysis and Findings

SmartPLS (Ringle et al. 2015), a type of component-based structural equation modeling (SEM) technique, was applied to check the quality of the measurement model and test the research hypotheses. SmartPLS technique is particularly suitable for exploratory theory building and testing (Lowry and Gaskin 2014), which fits one of the fundamental purposes of our study, i.e. exploring the theoretical causes for people to use the embryonic AV. SmartPLS also has an unparalleled advantage for testing complex research models consisting of both reflective and formative constructs. Our research model is fairly complex, including reflective and formative constructs together with second-order constructs. We operationalize perceived risk as a second-order formative construct consisting of four first-order risk dimensions, i.e. performance risk, safety risk, psychological risk, and social risk. Therefore, the SmartPLS technique is appropriate for this study. In the following subsections, we first confirm the reliability and validity of our measurement and then perform path modeling to test our research hypotheses.

Measurement Model Results

In line with the typical practice in the literature, we followed different criteria and procedures to test the measurement quality of reflective and formative scales. The measurement quality of formative scales, i.e. perceived benefit and perceived risk were evaluated based on the significance level of path weights and the extent of multicollinearity of formative indicators as suggested by MacKenzie et al. (2005). All formative indicators except social risk have significant path weights. The variance inflation factor (VIF) was then computed for each of the formative indicators to check the extent of multicollinearity. Excessive multicollinearity is suggested when VIF values are above 10. The VIF values range from 1.8 to 4.2 for indicators of perceived benefits and from 1.4 to 2.5 for those of perceived risks. As all formative indicators have acceptable VIF values, multicollinearity is not a concern for the two formative scales.

In the prior literature, keeping non-significant indicators is a recommended practice, especially in the situation involving low multicollinearity (Mathieson et al. 2001, Bollen and Lennox

1991, MacKenzie et al. 2005). Retaining non-significant indicators helps ensure the content validity of the construct. The VIF value of social risk is only 1.4, suggesting that social risk has little overlap with the other three risk dimensions. Dropping the social risk dimension would omit a unique part of risk perceptions. Therefore, despite the insignificance of social risk, we retained this risk dimension in the following data analysis.

After examining the measurement quality of formative scales, we assessed the measurement quality of eight first-order reflective constructs based on their reliability, convergent validity, and discriminant validity. A measurement scale is considered reliable if its composite reliability (CR) is 0.7 or higher and its average variance extracted (AVE) reaches 0.5 as suggested by Bagozzi and Yi (1988). All eight reflective scales were found to satisfy these two criteria for reliability. Convergent validity evaluates whether the items measuring the same construct load closely together. The convergent validity of a latent construct is supported if all its measurement items are 0.6 or higher (Bagozzi and Yi 1988) and are statistically significant (D. and D. 2005). We found that all items except the third item measuring personal innovativeness (i.e. *PI3*) have significant loadings above 0.6. Therefore, after dropping *PI3*, we re-run the data analysis using SmartPLS. All data analysis reported in this section reflects the results without *PI3*.

Following the suggestion by Fornell and Larcker (1981), we then checked the discriminant validity of our measurement model based on loading and cross-loading values (Table 3.4) and the correlation matrix (Table 3.5). Discriminant validity is suggested when items tapping a latent construct load more strongly on that construct than on any other constructs. Each item should have its loading value higher than its cross-loading values. Also, the square root of the AVE of each construct should exceed the correlations between that construct and any other constructs. From Tables 3.4 and 3.5, both criteria for discriminant validity were satisfied by the eight first-order reflective constructs. Therefore, our measurement model has sufficient reliability and validity.

Similar to other cross-sectional studies measuring all variables at one point in time, our findings may be subject to the bias of Common Method Variance (CMV). To test the potential

Constructs / Items		Loading / Cross-Loading							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1. Personal Innovativeness (PI)	PI1	**0.911**	0.608	0.614	–0.353	–0.306	–0.295	–0.052	0.621
CR $= 0.93$	PI2	**0.888**	0.475	0.486	–0.278	–0.212	–0.158	0.075	0.533
AVE $= 0.82$	PI4	**0.920**	0.517	0.502	–0.283	–0.257	–0.239	–0.024	0.524
2. Perceived Joy (JOY)	JOY1	0.577	**0.968**	0.813	–0.547	–0.472	–0.369	–0.011	0.789
CR $= 0.98$	JOY2	0.577	**0.975**	0.812	–0.518	–0.451	–0.382	–0.022	0.778
AVE $= 0.95$	JOY3	0.581	**0.976**	0.815	–0.540	–0.481	–0.399	–0.036	0.798
3. Trust in AI (TAI)	TAI1	0.577	0.796	**0.940**	–0.620	–0.538	–0.356	–0.018	0.739
CR $= 0.98$	TAI2	0.550	0.766	**0.950**	–0.580	–0.517	–0.326	–0.020	0.714
AVE $= 0.91$	TAI3	0.583	0.814	**0.969**	–0.601	–0.535	–0.347	–0.043	0.745
	TAI4	0.562	0.814	**0.958**	–0.597	–0.548	–0.366	–0.055	0.750
4. Performance Risk (PFR)	PFR1	–0.264	–0.445	–0.486	**0.863**	0.559	0.324	0.132	–0.414
CR $= 0.92$	PFR2	–0.267	–0.409	–0.478	**0.860**	0.565	0.345	0.183	–0.360
AVE $= 0.69$	PFR3	–0.311	–0.455	–0.516	**0.815**	0.716	0.430	0.274	–0.431
	PFR4	–0.332	–0.521	–0.604	**0.781**	0.692	0.404	0.209	–0.522
	PFR5	–0.214	–0.422	–0.491	**0.844**	0.544	0.315	0.122	–0.381
5. Safety Risk (SFR)	SFR1	–0.249	–0.397	–0.468	0.652	**0.849**	0.351	0.260	–0.401
CR $= 0.91$	SFR2	–0.202	–0.380	–0.445	0.589	**0.871**	0.449	0.261	–0.327
AVE $= 0.77$	SFR3	–0.301	–0.483	–0.554	0.726	**0.918**	0.544	0.313	–0.419
6. Psychological Risk (PSR)	PSR1	–0.296	–0.422	–0.384	0.426	0.505	**0.933**	0.463	–0.390
CR $= 0.93$	PSR2	–0.189	–0.313	–0.297	0.401	0.446	**0.932**	0.536	–0.270
AVE $= 0.87$									
7. Social Risk (SCR)	SCR1	–0.042	–0.084	–0.080	0.254	0.320	0.522	**0.935**	–0.064
CR = 0.92	SCR2	0.036	0.048	0.020	0.158	0.266	0.467	**0.917**	0.045
AVE $= 0.86$									
8. Intention (INT)	INT1	0.583	0.752	0.718	–0.472	–0.389	–0.299	0.014	**0.943**
CR $= 0.96$	INT2	0.582	0.773	0.720	–0.484	–0.397	–0.328	0.013	**0.948**
AVE $= 0.89$	INT3	0.589	0.762	0.742	–0.504	–0.451	–0.371	–0.066	**0.930**

Table 3.4: Composite Reliability (CR), Average Variance Extracted (AVE), and loadings and cross-loadings of reflective scales.

Construct	1	2	3	4	5	6	7	8
1. Personal Innovativeness (PI)	**0.906**							
2. Perceived Joy (JOY)	0.594	**0.973**						
3. Trust in AI (TAI)	0.595	0.836	**0.954**					
4. Performance Risk (PFR)	−0.340	−0.550	−0.628	**0.833**				
5. Safety Risk (SFR)	−0.289	−0.481	−0.560	0.751	**0.880**			
6. Psychological Risk (PSR)	−0.260	−0.394	−0.365	0.443	0.510	**0.933**		
7. Social Risk (SCR)	−0.005	−0.024	−0.036	0.226	0.318	0.536	**0.926**	
8. Intention (INT)	0.622	0.811	0.773	−0.517	−0.439	−0.354	−0.014	**0.941**

Diagonal elements (bold) are the square root of the Average Variance Extracted (AVE) values of all reflective constructs. Off-diagonal elements are the correlations among latent constructs.

Table 3.5: Discriminant validity of reflective measurement scales.

effect of CMV, we applied the marker-variable technique proposed by Lindell and Whitney (2001). This technique requires the identification of the second smallest positive correlation among the manifest variables as a more conservative estimate of CMV (i.e., r_m). The second smallest positive correlation was found to be 0.004 for our data. To assess the potential impact of CMV, we further computed CMV-adjusted correlations among latent constructs by partialing out r_m from the bivariate correlations in Table 3.5. The CMV-adjusted correlations are only slightly different from the original correlations with differences smaller than 0.0006. The significance levels of all correlations stay the same. Therefore, CMV was not a significant source of bias influencing the results of our study.

Hypothesis Testing

Figure 3.11 summarizes the results of path modeling, showing completely standardized path coefficients and significance levels along each path. Our research model explains 35.3% variance in perceived enjoyment, 36.8% variance in Trust in AI, 39.6% variance in perceived risk, 73.6% variance in perceived benefit, and 69.9% variance in intention. In SmartPLS, we performed bootstrapping of 5000 samples to test the statistical significance of hypothesized paths in our research model. All hypothesized paths except *H1*, i.e. the path between perceived risk and intention, are statistically significant. Overall, our research model is well supported. Among the three covariates, driving experience is significant with those more experienced drivers having lower use intention. Age and sex have no significant impact on intention.

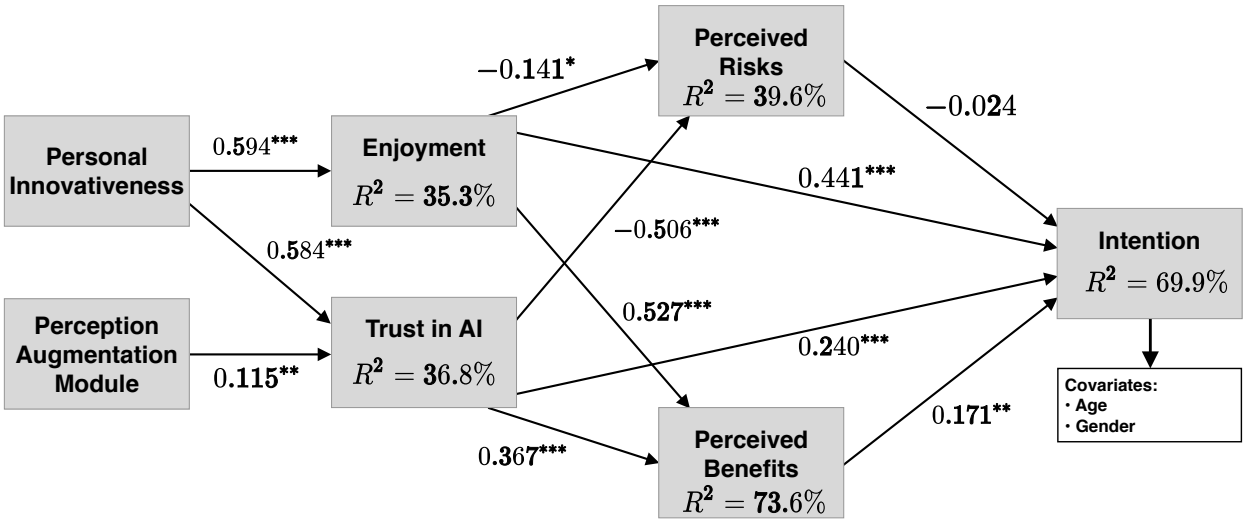


Figure 3.11: Results of testing hypotheses using Partial Least Squares (PLS) analysis.

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

3.3 Discussion

This work aimed to contribute to the extant literature surrounding ADS by providing an understanding of how a perception augmentation module could potentially improve the driver's trust in the underlying AI technology of the ADS. The study first introduced a conceptual framework of the interface between a human driver and an ADS (see Figure 3.1). Based on the framework, a perception augmentation module was developed that interacted with a hypothetical perception layer of an ADS to relay information about the driving environment – observed through the sensors of the vehicle – to the driver (see Figure 3.2). A research model was developed to validate the perception augmentation module through the lens of SCT.

Of the eleven hypotheses posed by this study (see Section 3.2.3) ten were supported by the collected data from the survey participants. The hypothesis that was not found to be significant, *H1*, posited that perceived risk (i.e., physical, financial, psychological) of using an ADS would have a negative impact on the intention to use an ADS (Guo et al. 2019, Luo et al. 2010). There are three possible reasons for this insignificant result. First, the drivers were informed that the ADS would allow them to intervene in the automation if the ADS began to

fail (see research scenario in Figure 3.10,) which may reduce the effect of perceived risks on adoption. Second, the majority of the participants in the study reported having more than ten years of driving experience, which may help increase their confidence in intervening in case of ADS failure. Lastly, the effect of perceived risks may be overridden by that of other competing factors that directly influence intention. To check this issue, we performed a robustness test by building three alternative models. In the first model (*M1*), we removed the link between trust and intention from the original model. In the second model (*M2*), we removed the path from perceived enjoyment to intention from the original model. In the third model (*M3*), we removed the two paths from trust and perceived enjoyment to intention. *H1* becomes significant in *M1* and *M3* but not in *M2*. This suggests that trust is the major factor overriding the effect of perceived risks. Without trust, perceived risks become a significant factor reducing adoption intention. The detailed results of the robustness test are available in Table 3.6.

Path	**Model**			
	Original	**M1**	**M2**	**M3**
H1: Perceived Risks → Intention	–0.024	**–0.070**	–0.041	**–0.154**
H2: Perceived Benefits → Intention	**0.171**	**0.244**	**0.371**	**0.652**
H3: Enjoyment → Intention	**0.441**	**0.553**		
H4: Enjoyment → Perceived Risks	**–0.141**	**–0.142**	**–0.141**	**–0.142**
H5: Enjoyment → Perceived Benefits	**0.527**	**0.527**	**0.527**	**0.527**
H6: Trust → Intention	**0.240**		**0.434**	
H7: Trust → Perceived Risks	**–0.506**	**–0.506**	**–0.506**	**–0.506**
H8: Trust → Perceived Benefits	**0.367**	**0.367**	**0.367**	**0.367**
H9: Personal Innovativeness → Enjoyment	**0.594**	**0.594**	**0.594**	**0.594**
H10: Personal Innovativeness → Trust	**0.584**	**0.584**	**0.584**	**0.584**
H11: Perception Augmentation Module → Trust	**0.115**	**0.115**	**0.115**	**0.115**
Age → Intention	0.013	0.007	0.003	–0.016
Gender → Intention	0.008	0.009	–0.005	–0.010
Driving Exp. → Intention	**–0.073**	**–0.074**	**–0.090**	**–0.104**
R-square	9.9%	8.6%	5.6%	9.9%

Significant path coefficients are shown in bold text.

Table 3.6: Path coefficients of alternative testing models.

We propose that, under the lens of SCT, the adoption of ADS would be governed by a trust-enabled hedonic social contract between drivers and the underlying AI systems. This

led to four core hypotheses surrounding the intent to adopt an ADS, namely, *H1*, *H2*, *H3*, and *H6*. Although *H1* was found to be insignificant, the remaining three hypotheses of this subset were validated by the data. This provides support for the application of SCT to the adoption of ADS. *H2* presented the idea that perceived benefit would have a positive impact on the intention to adopt an ADS. This was based on the previous work of Chan (2017), who theorize as to the potential benefits from ADS, both to society and individuals. Notably, we found no empirical works connecting perceived benefit to the adoption of ADS. As such, this work is the first to confirm this relationship in the context of ADS. From a different angle, we postulated *H3*: perceived enjoyment would have a positive impact on the intent to adopt an ADS. The idea that perceived enjoyment will positively impact an individual's intent to use technology was first studied by Agarwal and Karahanna (2000), Teo and Noyes (2011), and van der Heijden (2004), and is further confirmed by this study. Although Raue et al. (2019) noted that individuals find ADS enjoyable, this study is the first to empirically show the positive effect of perceived enjoyment on ADS adoption. The final hypothesis relating external constructs to the adoption of ADS, *H6*, poses that an individual's trust in the AI powering the ADS would positively influence that individual's intention to adopt or use an ADS. This is an idea that has been studied in the context of e-commerce by McKnight et al. (2002) and is further confirmed in the context of ADS by this work.

H4 (i.e., perceived enjoyment will have a negative effect on perceived risks) also proved significant in the structural model. This result agrees with prior literature which suggests a strong link between trust and perceived enjoyment in other discipline areas, namely, IT (Venkatesh 2000, Venkatesh et al. 2002) and mobile banking (Koenig-Lewis et al. 2015). Similarly, *H7*, which theorizes that the driver's trust in AI will have a negative impact on the perceived risks from using the ADS, was found to be significant. No other empirical studies have reported this result to our knowledge.

H5 and *H8* were theorized in regards to the effect of perceived enjoyment and trust in AI, respectively, on the perceived risks. Likewise, based on previous work in online payment

systems by Rouibah et al. (2016), *H5* states that perceived enjoyment would have a positive effect on the perceived benefits of the system. *H8* similarly proposes that the trust in AI would positively impact the perceived benefits of the system, which is an idea originally coined by Luo et al. (2010). Both of these hypotheses were supported by the data.

H9, which posits that those who perceive themselves as personally innovative would perceive enjoyment from using an ADS, was found to be significant. This aligns with prior empirical researches by Rouibah et al. (2016) and Zhou and Feng (2017). Similarly, *H10* proposed that those same individuals who perceive themselves as personally innovative would be more likely to trust AI. This result was also strongly confirmed, further cementing the research surrounding personal innovativeness in regards to trust, which was originally suggested by Schweitzer and van den Hende (2016).

A major contribution of this work is the perception augmentation system that is intended to improve the driver's trust in the AI technology that powers the ADS. Although the data shows a less significant link between the perception augmentation module and trust in AI than other links in the model, the link is still strong enough to be considered significant. One reason that the result does not appear more significant could be the simulated nature of the study. Participants were exposed only to a loose approximation of the system (through software-in-the-loop simulation) for a short time. As a result, participants may not have been able to garner appropriate experience with– or knowledge of the system to trust it (McKnight et al. 2002). Irrespective, the result shows that the perception augmentation module does improve the driver's trust in the AI technology. This confirms that reducing the technological opacity of an AI system can improve human trust in the system.

3.3.1 Contributions to Theory and Research

The findings of this study provide vital contributions to the research on trust and the decision to use emerging AI-enabled technologies. First, we extend the social contract model of health IT of Li et al. (2014) in three notable ways. 1) We contextualize the model to examine

the implicit social contract between drivers and AI systems of ADS. Our study represents an initial effort of applying the social contract lens in examining the relationship between human beings and AI-enabled new technologies and verifies the usefulness of the social contract lens in such a research context. The results of data analysis suggest that the lens of a social contract is useful for us to understand the relationship between drivers and AI as a black box, a context characterized by embryonic IT artifacts with a hedonic angle. 2) We extend the model by incorporating a hedonic angle to reflect the nature of the product for both utilitarian and hedonic purposes. The trust-enabled hedonic social contract model proposed in our study lays a theoretical foundation for advancing the research of complex technologies meant for both utilitarian and hedonic purposes. Our study opens a new avenue for further applying the trust-enabled hedonic social contract model to other similar IT artifacts such as augmented reality drones. 3) We expand perceived risks in the original model from privacy risk to multi-dimensional risks involving performance risk, safety risk, psychological risk, and social risk to fit the driving-related research context. Among the four risk dimensions, all but social risk emerge as significant. The other risk dimensions are all important risks factoring into the social contract between drivers and AI. Future studies on ADS or in a similar research context such as augmented reality could leverage and further validate the risk dimensions identified in this study.

Second, our study is the first that designs, and implements a perception augmentation module to open the black box of AI and empirically tests the mechanism for developing human beings' trust in AI supporting ADS. This innovative research methodology closely integrates the design of IT artifacts and the empirical test of human and IT interactions, which not only enhances the relevancy of our study but also spawns a new channel for advancing theory. Particularly, our results shed important light on the mechanisms for building trust in AI. Besides one's inherent personal innovativeness, imparting knowledge about the underlying AI through the perception augmentation module is found to be effective for building trust in AI. Future research could benefit from implementing different multi-media elements in the perception

augmentation module to gain a fine-grained understanding of why and how certain elements are effective or not effective for nurturing trust.

Third, our study confirms the crucial role of trust in AI for drivers to enter the social contract with a black-box AI. It not only indirectly influences the behavioral intention through the risk-benefit calculus but also exerts a direct effect. From the robustness test results (see Table 3.6) discussed above, trust is instrumental for drivers to overcome their perceived risks on their adoption intention. Thus, future research on technology use decisions should not neglect the pivotal role of trust in IT.

Lastly, our study also provides insights into the effect of perceived risks in the literature. The study by Li et al. (2014) supports the direct effects of both trust and perceived risks on intention simultaneously. However, our study finds perceived risks to be no longer significant in the existence of a direct link between trust and intention. The divergence from the study by Li et al. (2014) may be partly attributed to the differences in the characteristics of embryonic technologies and the types of trust. Our study examines trust in AI and the ADS that is equipped with a perception augmentation module for building the driver's trust in AI while Li et al. (2014) investigate patients' preexisting trust in firms providing Personal Health Record Systems. The conflicting findings regarding the effect of perceived risks suggest the importance of future studies to explicitly differ trust types and consider the characteristics of technologies when studying the effect of perceived risks.

3.3.2 Implications for Practice

Besides the theoretical contributions outlined above, two practical implications may be drawn from this work. First, the results suggest that designers of consumer AVs should prioritize the development of features that drivers find enjoyable, as opposed to features that may strictly improve the trust in the underlying AI. Such features may have a stronger downstream effect both on the intention to adopt an AV, and on the perceived benefits from using the ADS. Because people who perceive themselves to be personally innovative are shown to already

derive some enjoyment from using ADS, AV designers may also consider developing interfaces that make the system more enjoyable for those with lower degrees of interest in the technology. In the context of modern autonomy, this may include features like fully hands-free operation, stop-and-go traffic management, and road-sign detection, to name a few. By crafting ADS that are highly enjoyable, automakers can catalyze the adoption and encourage the continued usage of such systems.

Although this study primarily concerned autonomy for commercial vehicles, the results may be extrapolated to other markets with similar properties. In particular, military and aviation markets have a long history of applied autonomy for the control of vehicles. These markets embody significantly higher degrees of uncertainty and perceived risk than is typical of the commercial vehicle market. It is noteworthy that trust in AI has a far stronger negative effect on the perceived risks from using the system than the construct of enjoyment. This suggests that designing systems for these markets that prioritize the improvement of the operator's trust in the AI can be an effective strategy for encouraging the adoption and usage of such autonomous systems.

3.3.3 Limitations and Implications for Future Work

Two noteworthy limitations of this current work give way to implications for future work. First, higher degrees of realism could be achieved in the experiment to reduce any distortion of the driving experience. The most intuitive approach to combating this limitation is to deploy the technology in an actual ADS and have human participants ride in the vehicles. However, this solution imposes technical and safety challenges that are difficult to resolve. Bearing this in mind, previous works have developed state-of-the-art driving simulators that aim to reproduce as much of the driving experience as possible, without causing potential harm to the human participants (Rosique et al. 2019, Shahrdar et al. 2019). The extension of the perception augmentation system studied in this paper to more advanced simulators or experimental procedures can help in further confirming the results shown by this work. More sophisticated

driving simulators can also allow for the study of more complex perception augmentation systems. This work focused primarily on the design of an interface that relayed information to the driver using a screen and LED strip. Future work may introduce more nuanced HCI to improve the enjoyment from using the ADS. For instance, augmented reality interfaces used in modern video games are enjoyable. The integration of such technology into the ADS through the framework presented in this work marks an interesting avenue for future research and innovation.

This work focused on a single type of vision model for a particular mode of data, namely still images. In the future, we will explore the application of additional computer vision techniques, such as semantic segmentation (i.e., pixel classification), instance segmentation, 3-dimensional data processing, and video processing to determine how the different technologies can be leveraged to provide monitoring tools for the human drivers. Additionally, questions of whether virtual reality or augmented reality provides a meaningful interface between the ADS and the human will be investigated. A final future research direction concerns designing a more transparent vision model. In this work, we developed a system that provided transparency down to the object detection level, but a deeper understanding of why the object detection model chooses the classifications that it does is not currently understood. This is an active area of research in computer vision that can significantly improve the capabilities of the perception augmentation system.

Chapter 4

Choosing a Loss Function for Deep Image Deblurring

Two different kinds of loss functions are used in deep learning-based image restoration, namely, content losses and adversarial losses. *Content losses* are computed using paired image data to enforce a generator model to produce outputs that match the expected images. An alternative is to use *adversarial losses* via an auxiliary discriminator network that is trained to detect sharp versus degraded images. Adversarial losses can be difficult to stabilize during training and as such, authors that use GANs for image restoration tasks typically combine the adversarial loss with an auxiliary content loss or perceptual content loss to stabilize the training procedure (Nah et al. 2017, Kupyn et al. 2019). Although research continues to achieve state-of-the-art performance on the standard benchmarks for image deblurring (Chen et al. 2021), few works attempt to compare the innovations of new research on a granular level. In particular, several works have suggested and demonstrated the use of adversarial loss functions and perceptual content loss functions, but no research provides a deep comparison between the results of these different losses.

The results of this chapter show that despite the popularity of MSE as a content loss function for image restoration tasks, MAE frequently produces higher quality results. Furthermore, we show that generator models trained solely using a perceptual content loss produce outputs that are perceptibly better than the same model trained using a plain MAE or MSE loss despite validation metrics that would indicate otherwise. We show that adversarial losses do not produce generators capable of confidently deblurring images in the absence of auxiliary loss functions. Likewise, we show that the combination of adversarial and content losses in some cases produces higher quality results than either constituent loss when trained in isolation. Finally, we show examples where the best model in this study produces results that are in some

cases perceptibly better than the current state-of-the-art models when tested against real-world blur data. To the best of our knowledge, this is the first work to comprehensively assess the impact of content and adversarial losses on deep learning image deblurring models.

The remainder of this chapter is organized as follows. In Section 4.1, we describe the design of the experimental setting and the independent variables in the study. We go on to present the results of the quantitative and qualitative validation in Section 4.2. A discussion of the results of the study can be found in Section 4.3.

4.1 Methodology

4.1.1 Architecture

Following Lucic et al. (2018), we hold the architecture of our networks constant for comparing loss functions. We define our generator model based on extant literature on image deblurring and design the discriminator model from classifier model research. The models are kept simple to prevent the external influences of techniques like dropout, activation normalization, weight normalization, and the like.

Generator

We base the architecture of the generator model on the idea of residual learning. *Residual learning* is a well-studied method for reducing the vanishing gradient problem in deep networks (He et al. 2016). Deep residual models replace the direct feed-forward architecture of traditional neural pipelines with a residual structure. In residual blocks, activation maps propagate through one or more layers before being added back to the output of the parallel network. Figure 4.1 illustrates the architecture of the residual block used in the generator model where f is the Leaky REctified Linear Unit (Leaky ReLU) activation function. Leaky ReLU is used instead of ReLU because ReLU produces sparse gradients (Xu et al. 2015). The residual structure is typically repeated several times in a cascade and can be adapted with

more complex internal filtering mechanisms, such as in the Inception-v2 network of Ioffe and Szegedy (2015). Besides classification tasks, residual learning is common in image-to-image transformation models not only for the properties of the strengthened gradient in deeper models but also because it reduces the impact of error propagation on the model outputs, resulting in more stable and quicker optimization (Kupyn et al. 2018).

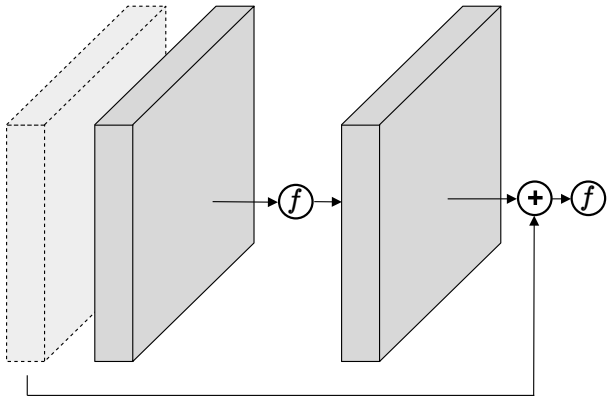


Figure 4.1: A depiction of a simple residual block. Two convolutional layers, shown in gray, are applied to the inputs, shown in light gray, and added back to the inputs before the final activation function f.

Figure 4.2 illustrates the architecture of the generator in terms of the activation map outputs of individual layers. The generator is composed of a series of cascaded convolutional layers in an encoder-decoder structure. The first layer in the generator, a convolutional filter followed by Leaky ReLU activation, projects the input image to 128 channels. Following this, the generator contains three down-sampling blocks composed of a convolutional layer with a stride of two, a Leaky ReLU activation, and a residual block. At the finest resolution, the generator cascades an additional three residual blocks before up-sampling. In the decoder network, the generator maps the encoded activation maps back to full size using a series of transposed convolutional layers, each followed by a Leaky ReLU activation and a residual block. The final layer projects the activation maps back to three channels and adds them to the input image acting as a global residual skip connection, which Kupyn et al. (2018) find stabilizes the training processes and reduces over-fitting. All convolutional kernels have bias terms and are size 3×3 other than the final output layer which is size 7×7. Different from Kupyn et al. (2018), we apply the

residual addition *before* the final hyperbolic tangent activation function, which we find produces better perceptive results by preventing over-saturation of the pixel illumination space. Based on the work of Gao et al. (2019), we apply local residual skip connections at each scale in the encoder-decoder. Namely, the activation maps before each down-sampling stage are skipped and added to the output activation map of each up-sampling stage. There are no feature pyramids, such as deployed by Kupyn et al. (2019), used in the baseline generator model. Also, no attention mechanisms, like those developed by Zamir et al. (2021) and Chen et al. (2021), are employed in the generator. Images are normalized to the domain of $[-1, 1]$ before being processed by the network. In total, the generator network contains $\mathrm{card}(\boldsymbol{\theta}_G) = 3,564,419$ parameters, which is comparable to the MobileNet architecture of Howard et al. (2017) that is designed for embedded systems, edge nodes, mobile phones, and such.

Discriminator

In cases where an adversarial loss is utilized, we exploit the simple discriminator architecture shown in Figure 4.3. The model is composed of two blocks of two convolutional layers. The second layer in each block down-scales the activation maps by a factor of two using convolution with a stride of two. All convolutional layers contain 128 filters, each of size 3×3 with a bias term, and a Leaky ReLU activation function. The final feature maps are down-scaled globally using the spatial pyramid feature pooling strategy of He et al. (2015) before being flattened into a one-dimensional feature vector. This feature vector is mapped to logits using a single dense layer with one output unit. For loss functions based on probabilities, a sigmoid activation function is applied to the output. Otherwise, the raw logits act as the output, and the discriminator is instead a *critic*. When the outputs of the model are probits, input images are normalized to the domain $[0, 1]$ to match the range of $[0, 1]$. For critic models that are logit-based, input images are normalized to the domain of $[-1, 1]$. In total, the discriminator network contains $\mathrm{card}(\boldsymbol{\theta}_D) = 450,817$ parameters.

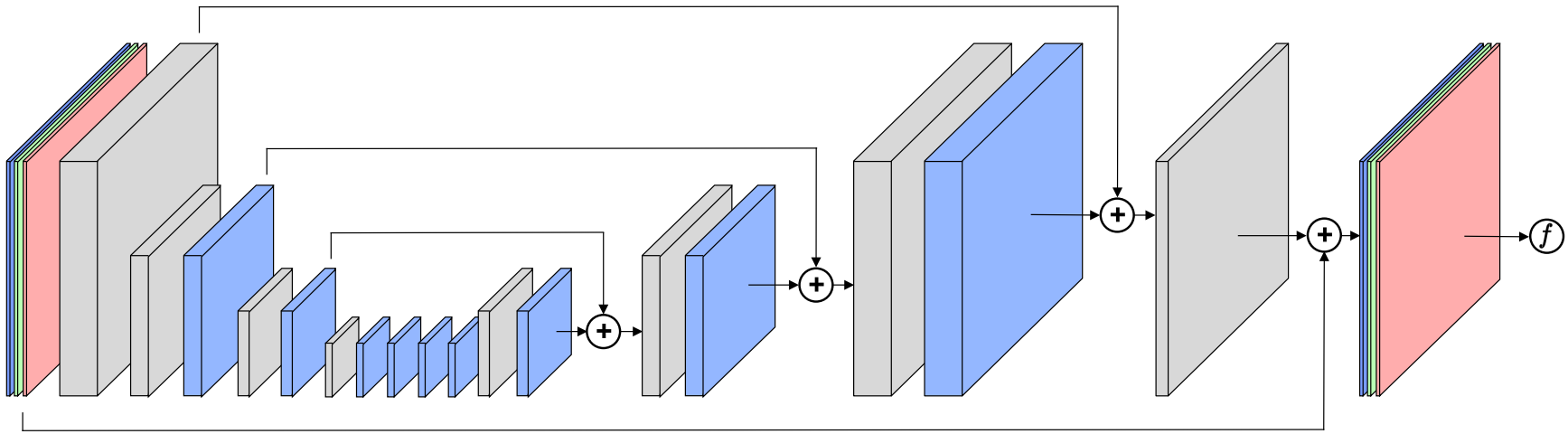


Figure 4.2: The convolutional generator network. The inputs and output of the model are RGB images of size $(M, N, 3)$. Gray blocks denote convolutional layer activation maps after a Leaky ReLU activation function and blue blocks indicate residual sub-network outputs (see Figure 4.1).

To validate the ability of the discriminator in our adversarial pipeline, we conduct a small experiment to ensure that the discriminator can learn to detect blurry images in isolation. Using the GoPro dataset, we train an independent discriminator model on the classification task of detecting blurry images among sharp ones in the training data. We train this discriminator model for 25 epochs using a batch size of 64 and the Adam optimizer with $\eta = 1\mathrm{e}{-4}$, $\beta_1 = 0.5$, and $\beta_2 = 0.9$. Images are cropped randomly to squares of size 256×256 during training. Because the pyramid pooling module produces the same number of activation maps irrespective of the size of the image, the discriminator can be validated against full-sized images even when trained on only small patches. The trained discriminator model can detect sharp and blurry images with 93.1% and 84.4% accuracy, respectively, on the GoPro testing set. It is worth noting, we do not use a pre-trained discriminator in our adversarial losses; the training of this one model is simply to test the capacity of the architecture.

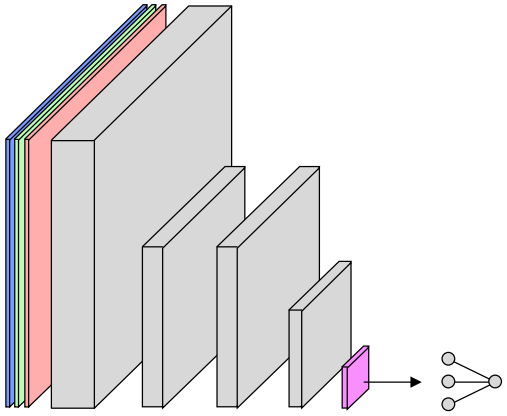


Figure 4.3: The convolutional discriminator network. The inputs to the model are RGB images of size $(M, N, 3)$. Gray blocks denote convolutional layer activation maps after a Leaky ReLU activation function. The pink block describes the outputs of a pyramid pooling layer. The flattened outputs of the pyramid pooling layer pass to a dense network with a single layer and single output unit.

4.1.2 Loss Functions

We train generator models using different combinations of loss functions to measure the effect of the loss on deblurring performance. We first use a single adversarial loss, which has

not been investigated in the literature to the best of our knowledge. Next, we investigate the effect of using only content loss. Although MSE has been well studied in the literature, MAE and frequency domain representations have not. Finally, we combine the adversarial losses with the content losses in a cross-study to determine the effect of the combination on the learning process.

Adversarial Losses

GANs are an active research area that undergoes constant innovation. In the context of image restoration and deblurring, adversarial losses have shown some success in a variety of forms. Specifically, GAN (Nah et al. 2017), WGAN-GP (Kupyn et al. 2018), and RaGAN (Kupyn et al. 2019, Zhang et al. 2020) have set state-of-the-art records in the past. A flavor of GAN that has not been studied in the context of image restoration is the LSGAN. The LSGAN has salient properties in that it can encourage the generator to produce samples that are both realistic to the discriminator and also close to the distribution of the real data. This prevents cases where the generator produces samples that fool the discriminator but are not drawn from a distribution that resembles real data and thus look fake to human observers. As such, Kupyn et al. (2019) propose that LSGANs are well suited for image restoration applications where adherence to the sharp data distribution is paramount.

In total, we study both the standard non-saturating GAN used by Nah et al. (2017) and the saturating GAN to set baseline metrics. Following Kupyn et al. (2018), we next utilize the WGAN-GP for training and measure its effect on the metrics. We also apply the standard WGAN loss with its weight clipping policy. Kupyn et al. (2019) and Zhang et al. (2020) both report good results using the RaGAN loss. As such, we study both the RGAN and RaGAN loss functions. Lastly, based on our aforementioned motivations we include the LSGAN in our study. Table 4.1 illustrates the losses of the discriminator and generator applied in this study. A comprehensive review of these losses and their foundation can be found in Section 2.2.6.

Framework	Loss	
	Discriminator	**Generator**
GAN	$\mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$	$-\mathbb{E}_{z\sim p_z(z)}[\log(D(G(z)))]$
GAN (Saturating)	$\mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$	$\mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$
LSGAN	$\frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(C(x) - 1)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(C(G(z)) + 1)^2]$	$\frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(C(G(z)) - 1)^2]$
WGAN	$\mathbb{E}_{x\sim p_{data}(x)}[C(x)] - \mathbb{E}_{z\sim p_z(z)}[C(G(z))]$	$\mathbb{E}_{z\sim p_z(z)}[C(G(z))]$
WGAN-GP	$\mathbb{E}_{x\sim p_{data}(x)}[C(x)] - \mathbb{E}_{z\sim p_z(z)}[C(G(z))] + \mathcal{L}_{GP}$	$\mathbb{E}_{z\sim p_z(z)}[C(G(z))]$
RGAN	$\mathbb{E}_{x\sim p_{data}(x)}[\log(\tilde{D}(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - \tilde{D}(x_f))]$	$\mathbb{E}_{x\sim p_{data}(x)}[\log(1 - \tilde{D}(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(\tilde{D}(x_f))]$
RaGAN	$\mathbb{E}_{x\sim p_{data}(x)}[\log(\hat{D}(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - \hat{D}(x_f))]$	$\mathbb{E}_{x\sim p_{data}(x)}[\log(1 - \hat{D}(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(\hat{D}(x_f))]$

Table 4.1: Adversarial loss functions used to train models in this study.

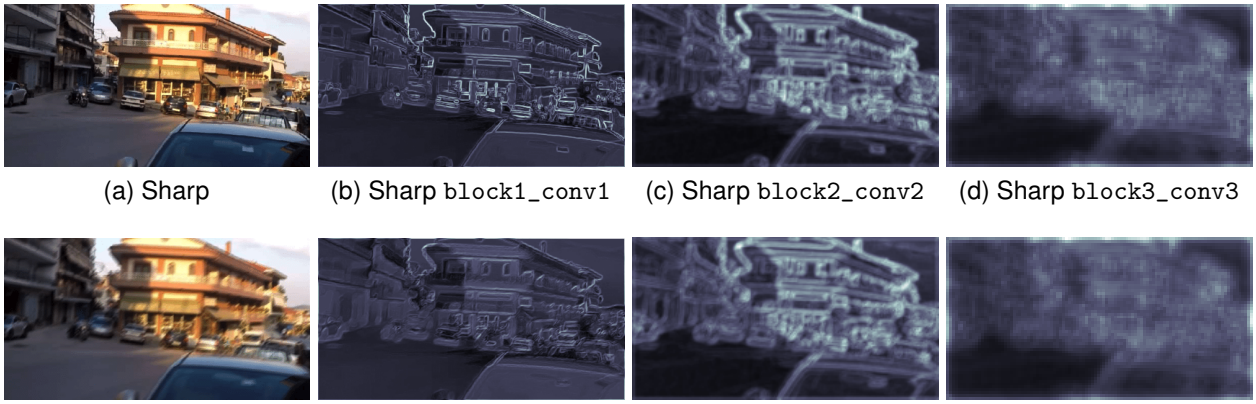
Content Losses

MSE over RGB luminance values is a frequent choice of loss function for image restoration tasks for its simplicity. MAE does not appear in the literature around deep image restoration but has advantages over MSE in that it places equal weight on all errors (Steffens et al. 2020). Because MSE computes squared errors, it is sensitive to large pixel differences (i.e., outliers) that may cause sub-optimal convergence. Sims (2020) further note that errors computed in spatial domains place equal weight to all frequency bands, which does not accurately represent how the HVS perceives images. As such, they derive inspiration from Wallace (1992) and calculate errors based on DCT coefficients instead of direct pixel illuminances. In this study, we apply both MSE and MAE loss functions, as well as their frequency domain equivalents MSE-DCT and MAE-DCT.

Perceptual Content Losses

For perceptual content losses based on pre-trained priors, `block3_conv3` activation maps of the VGG-19 network are often used as the auxiliary loss function due to this particular layer's ability to filter for object presence (Simonyan and Zisserman 2014, Johnson et al. 2016). This is a salient property for style transfer approaches because it orients the loss function towards pixel content with recognizable objects by masking low-entropy regions of the image. Intuitively, such a loss could extend to image restoration algorithms, which Kupyn et al. (2018), Kupyn et al. (2019), and Zhang et al. (2020) have shown. However, no current work related to image restoration addresses the selection of the layer from the VGG-19 network. Although the `block3_conv3` activation maps filter for object presence, lower level layers act as simpler edge detectors that are conditioned to natural image priors. These lower layers could provide valuable gradient information as an auxiliary loss function for image restoration approaches where sharpness does not necessarily correlate with object presence. Figure 4.4 illustrates the mean activation output for the `block1_conv1`, `block2_conv2`, and `block3_conv3` layers of the VGG-19 network after the ReLU activation function for a sharp-blurry image pair taken

from the GoPro training data. For visualization purposes, the activation maps are normalized by the L_∞ norm before mapping to pixel space using the "bone" color-map. To account for the down-sampling between blocks in the model, images are up-scaled to the same size, which introduces some blurring in the visualizations of the deeper layers. Due to the propagation effect of the ReLU non-linearity in the model, the activation maps of deeper layers become sparser as signal content is progressively filtered in a cascade. This is noted by Zhang et al. (2020), who elect to use the final layer's activation maps *before* the ReLU activation in an attempt to make the gradient signal less sparse. For this example, it is trivial to confirm that `block3_conv3` filters for the presence of the cars and buildings in the scene. The signal of `block1_conv1` is denser, containing strong activation only for basic edges in the image. `block2_conv2` exhibits strong edge detection, and also a degree of object presence filtering indicated by the sparseness of the low-entropy pixels describing the road and sky.



(a) Sharp (b) Sharp `block1_conv1` (c) Sharp `block2_conv2` (d) Sharp `block3_conv3`

(e) Blurry (f) Blurry `block1_conv1` (g) Blurry `block2_conv2` (h) Blurry `block3_conv3`

Figure 4.4: Example average activation outputs from the `block1_conv1`, `block2_conv2`, and `block3_conv3` layers of the VGG-19 network for a sharp-blurry image pair.

We investigate the effect of using each one of the `block1_conv1`, `block2_conv2`, and `block3_conv3` layers as an auxiliary loss function in terms of qualitative and quantitative metrics. We also measure the impact of combining the losses of the `block1_conv1`, `block2_conv2`, and `block3_conv3` layers into an ensemble loss.

4.1.3 Training

The training dataset used is the GoPro dataset of Nah et al. (2017), which was generated by the authors using the method described in Section 2.2.4. Specifically, we use the *gamma* subset of the data that applies a nonlinear model of the CRF to the synthetically blurry images to better represent the image acquisition process (see Section 2.2.1). Images are cropped randomly to windows of size $(256, 256)$ without flipping or rotating. Random noise is added to samples from the real-world data following the method of Nah et al. (2017). For each sample (in the floating-point domain $[0, 1]$), a standard deviation is randomly sampled from the Gaussian distribution $\sigma = \mathcal{N}(0, \frac{2}{255})^2$. Each pixel is randomly perturbed by a random noise generated by sampling from $\mathcal{N}(0, \sigma^2)$. Before training, weights of convolutional filters are initialized uniformly (Glorot and Bengio 2010) and bias terms are initialized to zeros. During training, mini-batches of size b are drawn to calculate gradients and update model weights of the generator, and batches of size $b/2$ are drawn for the discriminator. Both the discriminator and generator are trained using an independent Adam optimizer with $\eta = 1\mathrm{e}{-4}$, $\beta_1 = 0.5$, and $\beta_2 = 0.9$ for a total of 150 epochs (Kingma and Ba 2014, Gulrajani et al. 2017). In cases where there is no discriminator model, the generator is trained following the same policy and parameters for all content losses. The learning rate is decayed exponentially over the full duration of training to half of its original value, i.e., the final learning rate is $\eta_f = 5\mathrm{e}{-5}$. After the optimization algorithm completes, the best model weights are kept in terms of training PSNR. Table 4.2 specifies the exact training parameters used for each learning framework. It is worth noting, the WGAN framework requires the use of the RMSprop optimizer because momentum-based approaches like Adam will fail to converge (Arjovsky et al. 2017). WGANs discriminators are also over-trained to optimality by taking a ratio of five training steps for every one generator step.

Framework	Batch Size		Optimizer	$D : G$ Ratio
	Discriminator	Generator		
GAN	$b/2$	b	Adam	1
GAN (Saturating)	$b/2$	b	Adam	1
LSGAN	$b/2$	b	Adam	1
WGAN	b	b	RMSprop	5
WGAN-GP	b	b	RMSprop	5
RGAN	b	b	Adam	1
RaGAN	b	b	Adam	1

Table 4.2: Training parameters that are held constant in this study based on associated loss configurations.

4.1.4 Validation

Quantitative validation is performed using the GoPro test dataset of Nah et al. (2017), which contains 1111 paired validation samples. We additionally perform quantitative validation using the REDS dataset of Nah et al. (2019), which contains 3000 images that were generated using a similar, but refined method as compared to Nah et al. (2017). The generator can be validated objectively using PSNR and SSIM (see Section 2.2.5). For qualitative analysis, we use samples from the dataset of Lai et al. (2016) that contains images degraded by real-world blur samples. Because these are real-world blurs, there are no corresponding sharp images to compute hard metrics from. We also use samples from Köhler et al. (2012) for qualitative analysis. Code to train the models and generate the validation results in this study can be found at `https://github.com/Kautenja/choosing-a-loss-function-for-deep-image-deblurring`.

4.2 Results

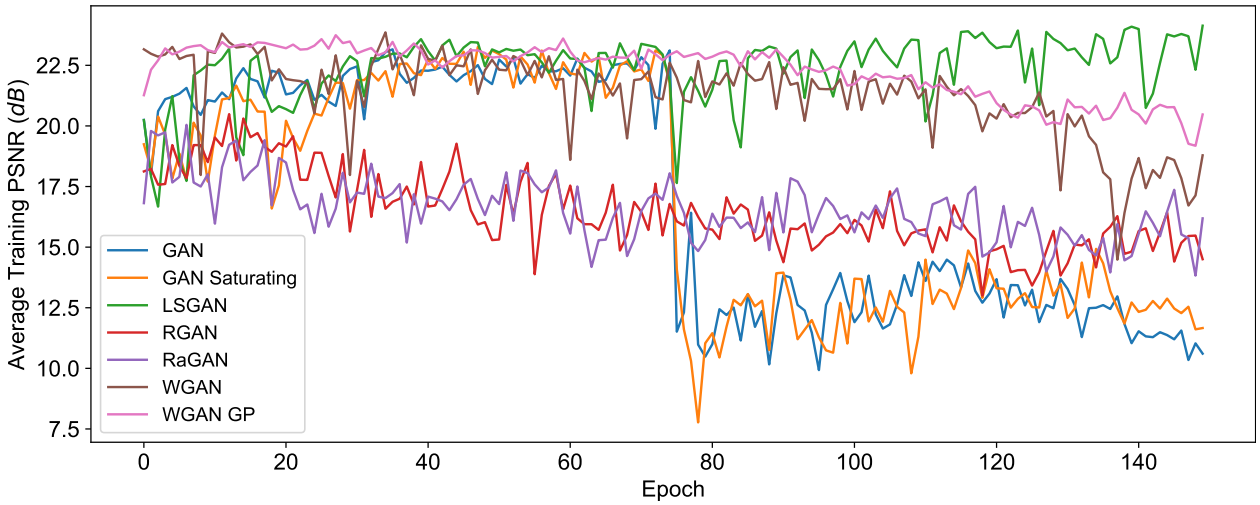
4.2.1 Adversarial Losses

To provide an understanding of training stability for adversarial loss functions, Figure 4.5 plots the average training metrics for each model over the batches of each training epoch. Because each epoch represents a random sample of the training data, some noise can be expected in the data. In terms of both PSNR and SSIM, LSGAN achieves the highest overall

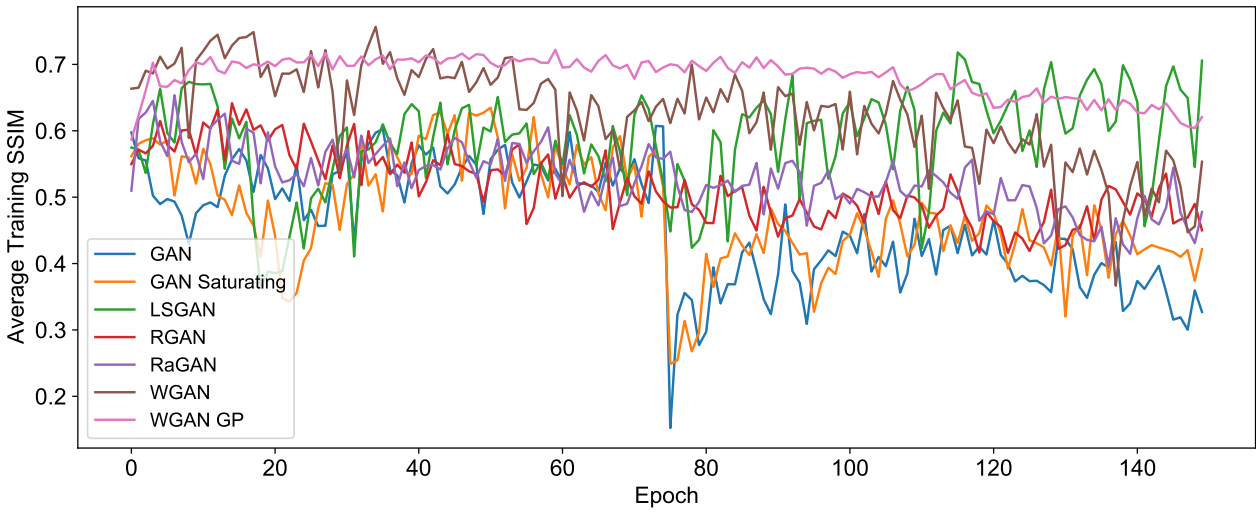
values and is the only adversarial loss to achieve a nearly increasing trend on average. WGAN and WGAN-GP perform similarly, but WGAN-GP reaches higher values and is overall smoother than WGAN. Despite converging on high training values, both WGAN and WGAN-GP losses begin to diverge after epoch 100. Both RGAN and RaGAN losses fail to improve the metrics past the values produces by the initial weights. Although GAN and saturating GAN models both converge on competitive PSNR readings around epoch 75, the SSIM improves very little suggesting that the images are not improving in perceptive quality. Both vanilla GAN models encounter a collapse around epoch 80 where the generator model begins to diverge.

Training metrics only provide insight into the interaction of the model and the optimization algorithm. To understand the final generator model from each learning framework, the models are tested against GoPro and REDS testing sets to measure the PSNR and SSIM on data that was not observed during training. Table 4.3 displays the validation metrics for each adversarial loss tested in this study. Overall, the LSGAN loss function produces the highest validation metric across all datasets. No one loss function produces the second-best result, but it is worth noting that the WGAN does produce the second-highest validation metrics on the GoPro set. No adversarial loss produces a generator model capable of improving the baseline metrics of the degraded images relative to the sharp images. This does not necessarily imply that adversarial loss alone cannot produce a viable generator in all architecture and hyperparameter configurations but is an important consideration when comparing the results of the adversarial losses against those of the content losses and combined losses.

Because PSNR does not always correlate to perceptible quality in the HVS, it can sometimes be a misleading metric when applied to image synthesis problems. SSIM can provide a more robust estimate of quality to the HVS, but it is frequently still useful to evaluate generator outputs manually using the HVS. Furthermore, in the context of deblurring, all datasets available for computing hard metrics are the result of blur approximation models because it is not currently possible to produce sharp-blurry pairs of real-world blur systems. As such, evaluation using the HVS is the only way to examine how deblurring models generalize to real-world blur



(a) PSNR



(b) SSIM

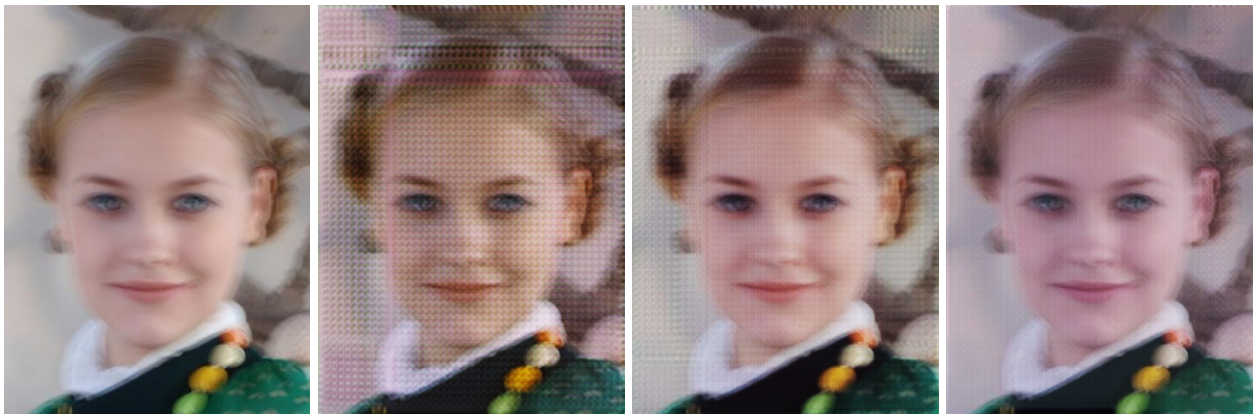
Figure 4.5: Average PSNR and SSIM per metric during the training procedure for models trained with adversarial losses.

Adversarial Loss	GoPro		REDS	
	PSNR	SSIM	PSNR	SSIM
Degraded Images	25.6	0.792	26.2	0.770
GAN	21.0	0.555	21.9	0.568
GAN (Saturating)	22.3	0.688	<u>22.8</u>	0.673
LSGAN	**23.5**	**0.744**	**24.1**	**0.725**
WGAN	<u>22.7</u>	<u>0.741</u>	21.9	0.701
WGAN-GP	20.9	0.629	21.0	0.629
RGAN	19.0	0.647	19.8	0.636
RaGAN	21.1	0.722	21.0	<u>0.705</u>

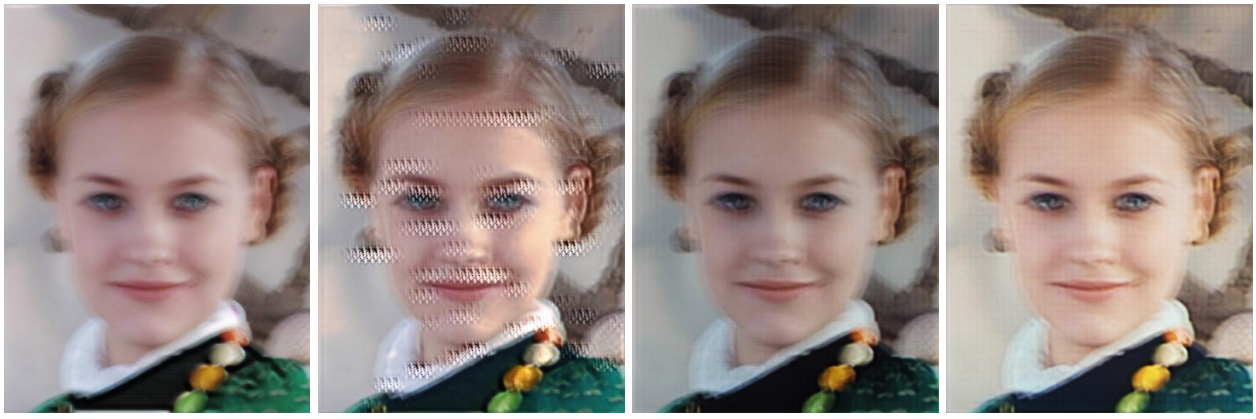
Table 4.3: PSNR and SSIM metrics on the GoPro and REDS test benchmarks based on generators trained with different adversarial loss functions. The best values are shown in bold and the second-best values are underlined.

systems. Figure 4.6 provides a comparative illustration of the generator outputs for an image taken from the dataset of Lai et al. (2016), namely, `face2`. This image is selected due to the HVSs natural ability to detect human faces.

The GAN model produces an output with a large number of checkerboard artifacts caused by poor learning of transposed convolutional layer weights. The same artifacts are observable in the saturating GAN and LSGAN networks, but with far less intense of an effect. Interestingly, the saturating GAN produces a perceptibly higher quality image than the GAN in this case but introduces a high degree of image saturation relative to the blurred image and all other model outputs. The LSGAN produces an output that appears no less blurry than the original image and has undergone some non-uniform DC coefficient distortion that causes shifts in hue relative to the other images. The output of the WGAN closely resembles that of the blurry image. Despite achieving a high validation metric, the model generalizes poorly to validation data outside the stylized testing set. The WGAN-GP produces outputs with non-uniform checkerboard artifacts indicative of GAN collapse. Both RGAN and RaGAN exhibit mild degrees of checkerboard artifacts and uniform DC bias. In the case of the RGAN, the image appears darker, but without color-shift, indicating uniform removal of signal content between channels. The RaGAN model behaves inversely, producing outputs with uniformly more signal content resulting in a brighter image without hue adjustment. It is a common misperception of the HVS that brighter, i.e.,



(a) *Blurred* (b) GAN (c) GAN (Saturating) (d) LSGAN



(e) WGAN (f) WGAN-GP (g) RGAN (h) RaGAN

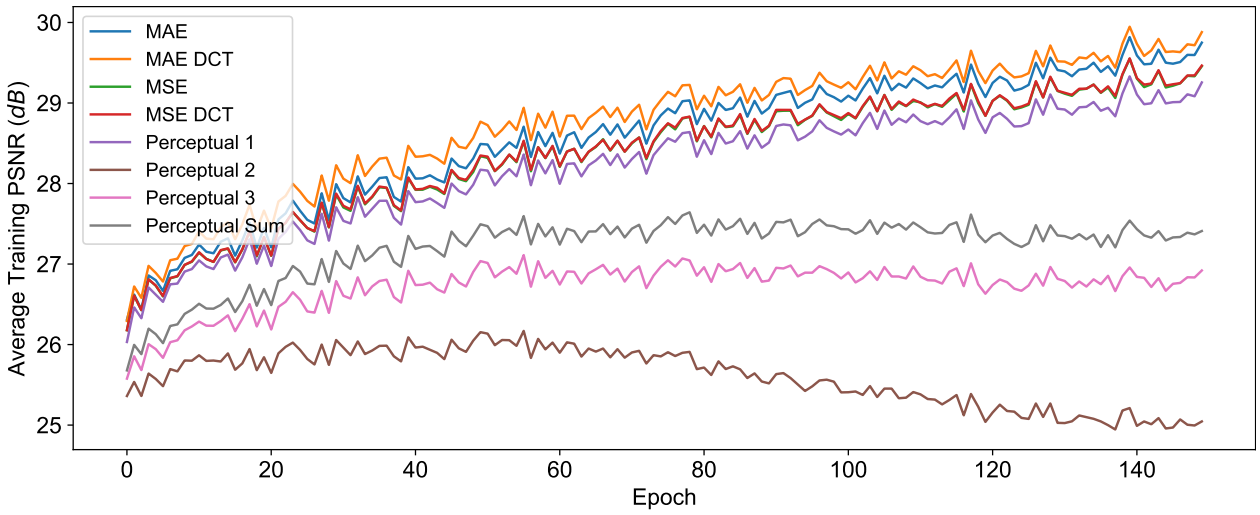
Figure 4.6: Examples restorations of "`face2`" from the dataset of Lai et al. (2016) based on generators trained with different adversarial loss functions.

higher energy, image signals are perceptually better than darker counterparts. For this reason, the output of the RaGAN model may appear better than other models. Upon close inspection, the RaGAN does produce a sharpening effect but fails to remove ghosted edges, such as around the face and collar.

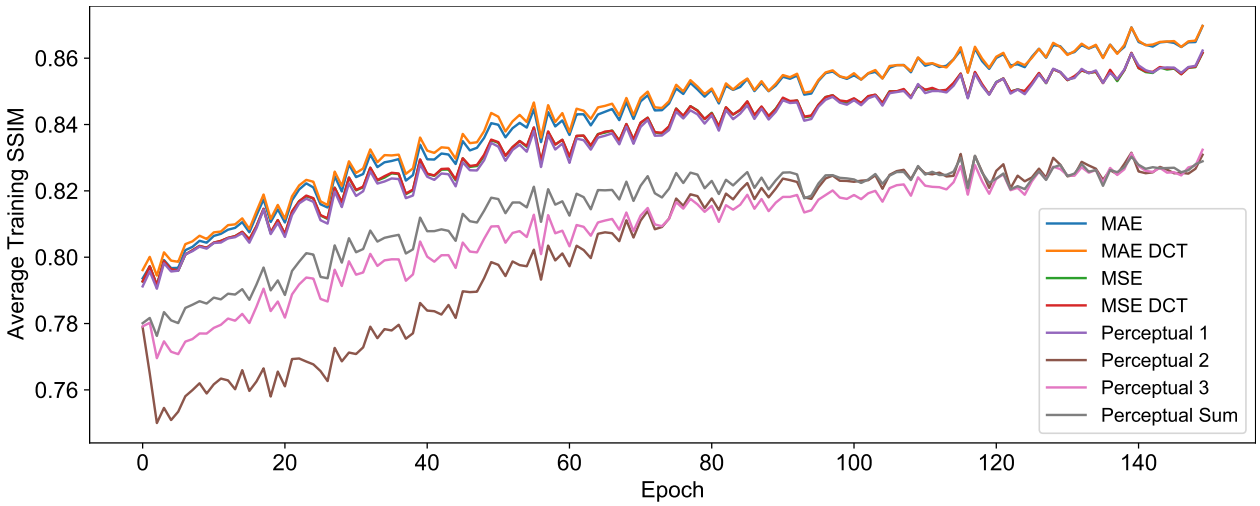
4.2.2 Content Losses

To compare the training performance of the content losses, Figure 4.7 displays the average PSNR and SSIM per training epoch over the full training duration. Unlike the trends of the adversarial losses, the trends of the content losses are smooth and relatively similar. Because the dataset loader was conditioned on the same random seed for all optimization runs, the influence of challenging/trivial subsets of data causes similarity in the shapes of both PSNR and SSIM curves. Notably, perceptual content losses based on the pre-trained layers of VGG-19 produce consistently lower training metrics than do the standard pixel-based and frequency-domain loss functions. In terms of SSIM, `block2_conv2`, `block3_conv3`, and the VGG-19 ensemble all converge to the same value. This is despite each having a different trend in terms of PSNR. Interestingly, the `block2_conv2` loss function results in diverging PSNR values, but stably converging SSIM values. Because `block1_conv1` is a relatively shallow layer that has not been as heavily influenced by the sparseness of the cascaded ReLU functions, it has a performance trend that is more comparable, but strictly worse than, all of the pixel-based and frequency-domain losses. The DCT-based MAE loss function achieves the highest training metrics among all other methods; however, the improvement over the spatially computed MAE is marginal in terms of PSNR and immeasurable in terms of SSIM. The DCT-based MSE and the spatial MSE behave identically, both falling slightly below the MAE loss functions.

During the validation stage, perceptually based losses continue to underperform relative to the simpler MSE and MAE solutions, as indicated by the results in Table 4.4. On both the GoPro and the REDS dataset, the DCT-based MAE produces the highest validation metrics in terms of both PSNR and SSIM. The spatial MAE produces the second-highest results across



(a) PSNR



(b) SSIM

Figure 4.7: Average PSNR and SSIM per metric during the training procedure for models trained with content losses.

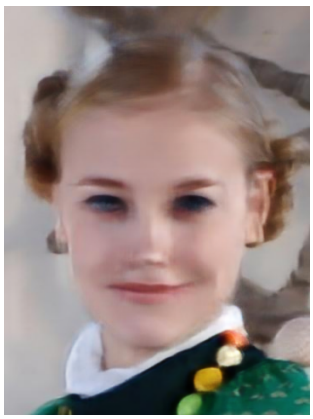
the board. The improvement of MAE over both MSE loss functions on the GoPro testing set is marginal, but on the REDS data, it is more significant. Similar to the behavior observed in the training metrics, `block1_conv1` produces validation metrics that are comparable to the standard MAE and MSE losses. `block2_conv2` and `block3_conv3`, and the VGG-19 ensemble all fail to achieve comparable metrics to the other loss functions. It is worth noting, the perceptual losses are intended to improve the quality to the HVS based on pre-trained natural image priors. Although the performance on the quantitative metrics is poor, prior literature often reports that using perceptual content losses improves the qualitative results.

Content Loss	**GoPro**		**REDS**	
	PSNR	SSIM	PSNR	SSIM
Degraded Images	25.6	0.792	26.2	0.770
MAE	<u>27.6</u>	<u>0.848</u>	<u>26.1</u>	<u>0.783</u>
MAE-DCT	**28.0**	**0.855**	**26.4**	**0.794**
MSE	27.5	0.842	25.4	0.765
MSE-DCT	27.5	0.843	25.5	0.770
`block1_conv1`	27.4	0.845	25.3	0.778
`block2_conv2`	25.1	0.785	24.5	0.739
`block3_conv3`	25.4	0.790	24.4	0.744
Perceptual Ensemble	26.0	0.806	24.4	0.737

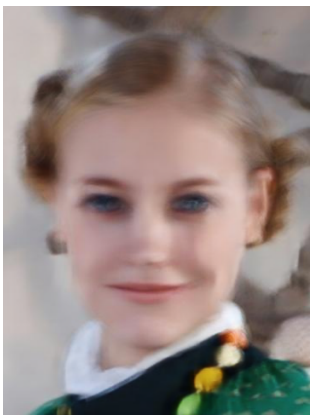
Table 4.4: PSNR and SSIM metrics on the GoPro and REDS test benchmarks based on generators trained with different content loss functions. The best values are shown in bold and the second-best values are underlined.

Figure 4.8 provides a qualitative assessment of the generator models learned using a single content or perceptual content loss function. The image used is the same `face2` image of Lai et al. (2016) that was used during the qualitative validation of the content losses. Consistent with the results of prior literature, the usage of a perceptual content loss produces models that are perceptually more crisp, less blurry, and generally containing fewer artifacts than the simpler MSE and MAE approaches. This contradicts the expectation based on training and validation metrics where the four MSE and MAE approaches all outperformed the perceptual content losses. As was seen in the quantitative assessment of adversarial losses, the `block1_conv1` loss produces outputs that closely resemble those of the models trained using MAE and MSE

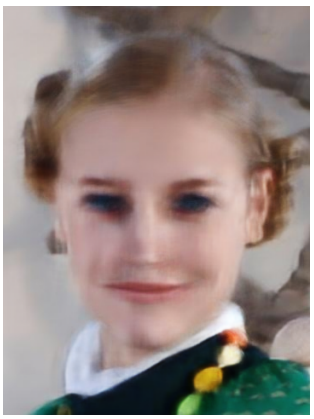
losses. Compared to the adversarial loss functions, all these examples demonstrate a higher degree of deblurring capability indicated by the collar of the woman's blouse where the ghosting has been mostly removed. However, the spatial and frequency domain variants of both MSE and MAE loss functions all fail the improve the quality of the woman's face, and in many cases, introduce new artifacts. `block2_conv2`, `block3_conv3`, and the VGG-19 ensemble loss all produce relatively similar results that are qualitatively better than those of the four MSE and MAE variants. In this case, `block2_conv2` produces a smoother result that some may perceive to be higher quality than the result of `block3_conv3` that is crisper, but with added artifacts. The VGG-19 ensemble loss appears to have a very literal effect of averaging both the benefits and the artifacts of constituent layers. For instance, the same artifacts in the hair can be observed from the `block3_conv3` example, but the smoothness of the `block1_conv1` and `block2_conv2` examples can be seen in the face and eyes.



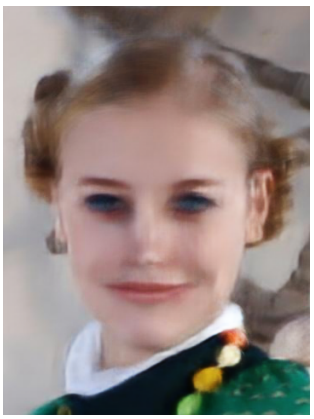
(a) MAE



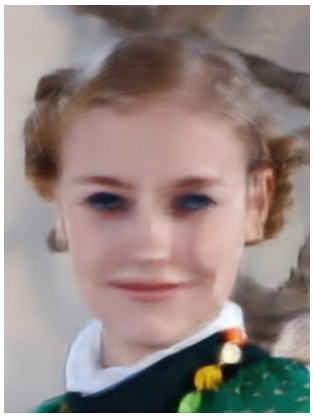
(b) MAE-DCT



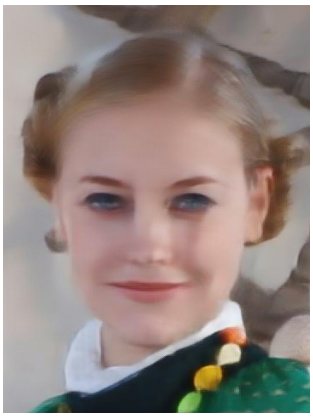
(c) MSE



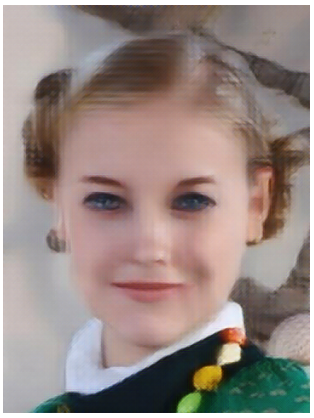
(d) MSE-DCT



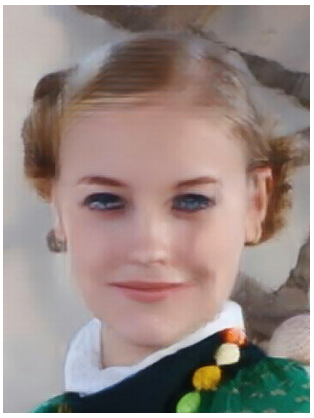
(e) `block1_conv1`



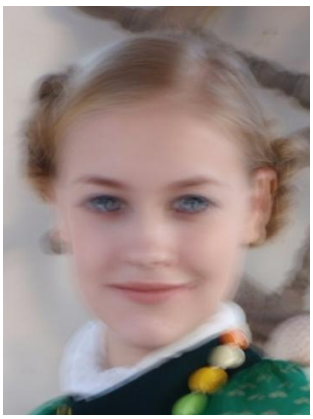
(f) `block2_conv2`



(g) `block3_conv3`



(h) Perceptual Ensemble



(i) *Blurred*

Figure 4.8: Examples restorations of "`face2`" from the dataset of Lai et al. (2016) based on generators trained with different content loss functions.

4.2.3 Combined Losses

Because WGAN and WGAN-GP models exhibited high degrees of model collapse, and require five times as many computational resources to train, we omit them from inclusion in this experiment. Performance on training metrics for the combined loss models closely resembles those of the content losses. This coupled with a large number of adversarial and content loss function pairs drives us to omit illustration of the combined loss metrics during training.

Table 4.5 provides a cross-reference table between the validation metrics of each adversarial and content loss function combination based on GoPro and REDS testing sets. The combinations of any adversarial loss with any of the four MSE and MAE-based losses relatively consistently produce higher metrics on both datasets than do any of the perceptual losses used in combination with an adversarial loss. Although in isolation the LSGAN produces the highest metrics across the board, in combinations the basic non-saturating GAN loss produces better metrics. However, it is worth noting that the improvement of the vanilla GAN over the LSGAN model is relatively marginal. the Saturating GAN produces metrics that are between the non-saturating GAN and the LSGAN. Interestingly, both RGAN and RaGAN models produce consistently lower validation metrics across the board.

Although the combinations featuring simple GAN adversarial loss combined with MSE and MAE content losses produced the highest validation metrics, we find the best perceptible results with the `block3_conv3` variant. Of the adversarial losses studies, we find that GAN produces the best qualitative results overall. Figure 4.9 provides an image comparison matrix of the four best content losses with the three best adversarial losses. Overall, there is a high degree of variability between the different results despite having marginally similar metrics.

Adversarial Loss	Content Loss	GoPro		REDS	
		PSNR	SSIM	PSNR	SSIM
Degraded Images		25.6	0.792	26.2	0.770
GAN	`block1_conv1`	26.8	0.834	25.2	0.770
	`block2_conv2`	26.6	0.826	24.5	0.743
	`block3_conv3`	26.8	0.829	24.8	0.746
	Perceptual Ensemble	26.6	0.832	24.3	0.743
	MAE	27.4	**0.845**	25.9	0.785
	MAE-DCT	27.1	0.833	**26.0**	**0.786**
	MSE	**27.6**	0.843	25.9	0.781
	MSE-DCT	**27.6**	0.843	25.7	0.777
GAN-S	`block1_conv1`	26.6	0.828	24.9	0.758
	`block2_conv2`	26.8	0.828	24.4	0.743
	`block3_conv3`	26.8	0.829	24.8	0.749
	Perceptual Ensemble	27.0	0.838	24.7	0.751
	MAE	27.2	0.843	25.2	0.772
	MAE-DCT	24.4	0.774	24.9	0.759
	MSE	27.5	0.840	25.7	0.778
	MSE-DCT	27.5	0.841	25.7	0.777
LSGAN	`block1_conv1`	26.4	0.822	24.7	0.753
	`block2_conv2`	26.6	0.829	24.6	0.756
	`block3_conv3`	26.8	0.829	25.1	0.752
	Perceptual Ensemble	27.0	0.836	24.7	0.751
	MAE	27.3	0.844	25.4	0.775
	MAE-DCT	26.4	0.808	24.8	0.742
	MSE	27.4	0.840	25.8	0.776
	MSE-DCT	27.3	0.838	25.9	0.777
RGAN	`block1_conv1`	25.7	0.804	24.3	0.733
	`block2_conv2`	26.5	0.832	24.8	0.761
	`block3_conv3`	26.0	0.808	24.1	0.728
	Perceptual Ensemble	26.5	0.837	24.4	0.757
	MAE	25.5	0.807	24.0	0.742
	MAE-DCT	26.1	0.814	24.6	0.758
	MSE	27.2	0.833	25.8	0.776
	MSE-DCT	26.9	0.828	25.6	0.772
RaGAN	`block1_conv1`	26.4	0.821	24.5	0.755
	`block2_conv2`	25.6	0.814	24.1	0.747
	`block3_conv3`	24.8	0.783	23.2	0.701
	Perceptual Ensemble	25.3	0.796	23.6	0.712
	MAE	26.5	0.830	25.2	0.774
	MAE-DCT	25.5	0.792	23.9	0.731
	MSE	26.8	0.826	25.3	0.766
	MSE-DCT	26.7	0.826	25.2	0.766

Table 4.5: PSNR and SSIM metrics on the GoPro and REDS test benchmarks based on generators trained with different combinations of adversarial and content loss functions. The best values are shown in bold and the second-best values are underlined.

(a) GAN-block2_conv2 (b) GAN-block3_conv3 (c) GAN-MAE-DCT (d) GAN-MSE

(e) GAN (Saturating)-block2_conv2 (f) GAN (Saturating)-block3_conv3 (g) GAN (Saturating)-MAE-DCT (h) GAN (Saturating)-MSE

(i) LSGAN-block2_conv2 (j) LSGAN-block3_conv3 (k) LSGAN-MAE-DCT (l) LSGAN-MSE

Figure 4.9: Examples restorations of "face2" from the dataset of Lai et al. (2016) based on generators trained with different combinations of adversarial and content losses.

4.2.4 Comparisons to State-of-the-art

Table 4.6 outlines the mean validation metrics (i.e., PSNR and SSIM) for current state-of-the-art deblurring methods on the 1111 samples in the GoPro benchmark testing dataset. The works of Xu et al. (2013) and Sun et al. (2015) predate the existence of the GoPro dataset, whereas the remaining models are trained directly using the GoPro training data. Up to the current state-of-the-art, there is a nearly monotonically increasing trend of PSNR and SSIM suggesting that the learned models are generalizing better to the distribution of the testing data. Although our model does not attempt to breach state-of-the-art metrics, the performance of our best model is shown for reference. The performance in terms of PSNR is comparable to that of the MobileNet model of Kupyn et al. (2019), although the SSIM metric is more comparable to that of Xu et al. (2013).

Model	**PSNR**	**SSIM**
Xu et al. (2013)	25.1	0.842
Sun et al. (2015)	24.6	0.890
DeepDeblur (Nah et al. 2017)	29.2	0.916
DeblurGAN (Kupyn et al. 2018)	28.7	0.958
SRN (Tao et al. 2018)	30.2	0.934
DeblurGAN-v2 (Kupyn et al. 2019)	29.5	0.934
DeblurGAN-v2 (MobileNet) (Kupyn et al. 2019)	28.1	0.925
Gao et al. (2019)	31.5	0.947
DMPHN (Zhang et al. 2019)	31.5	0.948
Zhang et al. (2020)	31.1	0.942
SAPHNet (Suin et al. 2020)	32.0	0.953
RADNet (Purohit and Rajagopalan 2020)	32.1	0.956
BANet (Tsai et al. 2021)	32.4	0.957
MPRNet (Zamir et al. 2021)	32.6	0.959
HINet (Chen et al. 2021)	32.7	0.959
Ours (GAN-`block3_conv3`)	26.8	0.829

Table 4.6: PSNR and SSIM metrics on the GoPro test benchmark. Metrics for previous works were derived from the papers.

Although the combinations featuring simple non-saturating GAN adversarial loss combined with MSE and MAE content losses produced the highest validation metrics, we found the best perceptible results with the `block3_conv3` variant. Figure 4.10 provides a qualitative

comparison of the GAN-block3_conv3 model from this study against the current state-of-the-art approaches for the 385/11_01_003028 image from the GoPro testing set. At low resolution, the results of this work are comparable to those of the state-of-the-art despite achieving far lower metrics. The primary improvements of the DMPHN (Zhang et al. 2019), MPRNet (Zamir et al. 2021), and HINet (Chen et al. 2021) relative to our model exist in the fine details such as around the structural beams in the image. It is worth noting, the DeblurGAN-v2 model of Kupyn et al. (2019) produces significantly blurrier regions along these support beams than our simpler and under-trained model. Although not obvious in this example, the HINet model introduces structured grid artifacts in the image that result from its patch architecture.

Figure 4.11 illustrates the same face2 image from the dataset of Lai et al. (2016) that has been used in prior qualitative evaluations to this point. Notably, the current state-of-the-art models with regards to GoPro testing metrics, Zamir et al. (2021) and Chen et al. (2021), produce significant artifacts for this particular example. The patch architecture of HINet results in grid-variant image restoration that is visible along the edges of the patches of the image. DMPHN and MPRNet both introduce artifacts and excess smoothing to the image that corrupts significant existing detail. Furthermore, an inspection of the woman's collar reveals that all three of DMPHN, MPRNet, and HINet fail to remove the ghosting effect, whereas DeepDeblur, DeblurGAN-v2, and our model all address this detail. Our result is the most comparable to the DeepDeblur model of Nah et al. (2017), though details such as the woman's ear and eyes appear to be more accurately restored by their model. The DeblurGAN-v2 model of Kupyn et al. (2019) produces crisp restoration of the woman's eyes but does not restore detail to the hair as well as DeepDeblur or our model.

Figure 4.12 provides an example output to illustrate cases where the models in the prior literature are very effective. Namely, the text10 image from the dataset of Lai et al. (2016) is used to investigate the document restoration capabilities of the models. Although some of the state-of-the-art models did not adapt as well to natural images outside of the training distribution, the generalization capability of those models to documents is better than our model. Of the

(a) *Sharp*

(b) *Blurred*

(c) DeepDeblur (Nah et al. 2017)

(d) DeblurGAN-v2 (Kupyn et al. 2019)

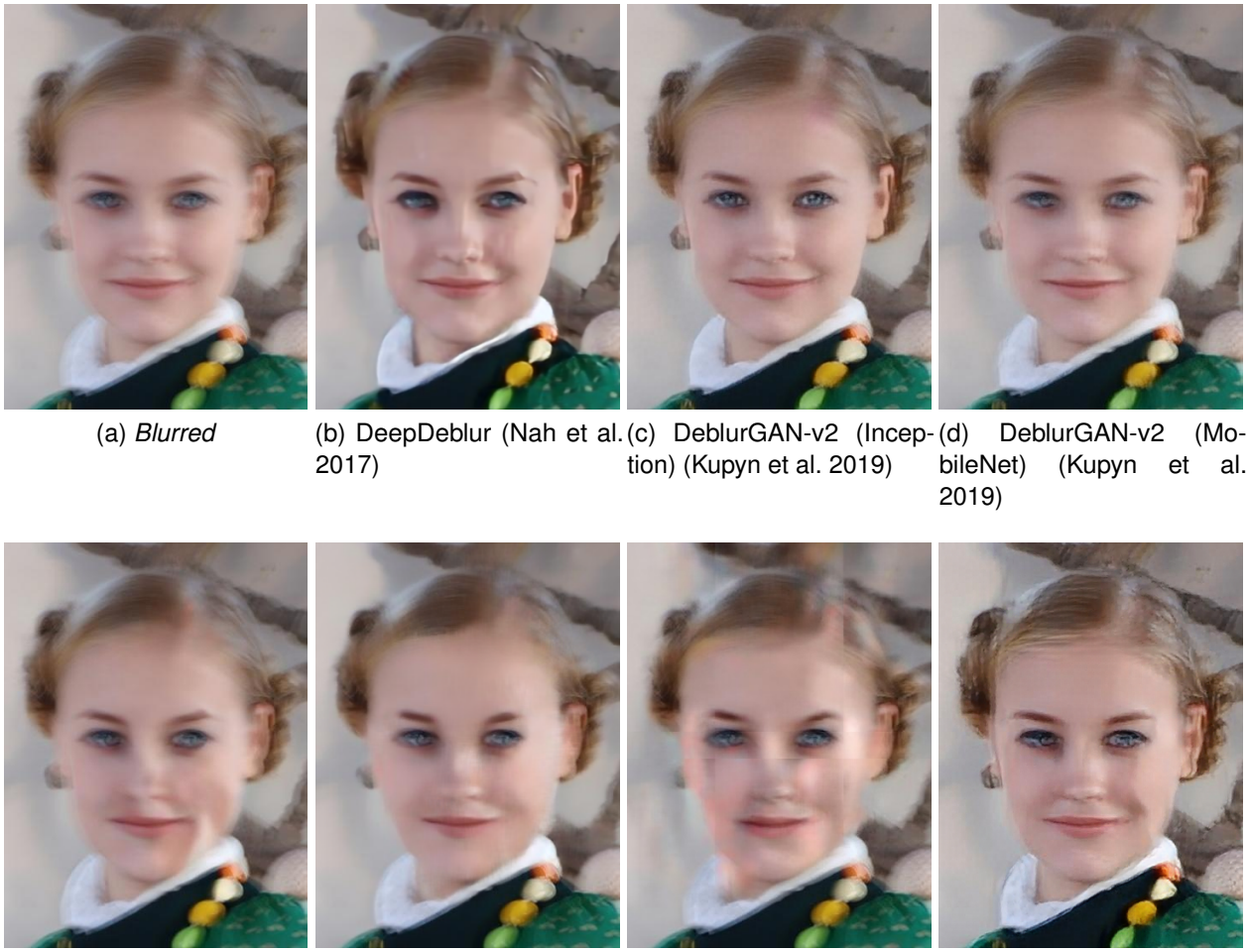
(e) DMPHN (Zhang et al. 2019)

(f) MPRNet (Zamir et al. 2021)

(g) HINet (Chen et al. 2021)

(h) ours

Figure 4.10: Examples restorations of "`385/11_01_003028`" from the dataset of Nah et al. (2017). Pre-trained models are used to evaluate existing methods.



(a) *Blurred* (b) DeepDeblur (Nah et al. 2017) (c) DeblurGAN-v2 (Inception) (Kupyn et al. 2019) (d) DeblurGAN-v2 (MobileNet) (Kupyn et al. 2019)

(e) DMPHN (Zhang et al. 2019) (f) MPRNet (Zamir et al. 2021) (g) HINet (Chen et al. 2021) (h) Ours

Figure 4.11: Examples restorations of "`face2`" from the dataset of Lai et al. (2016). Pre-trained models are used to evaluate existing methods.

compared methods, only the DMPHN model fails to restore the image to full quality. Notably, although MPRNet and HINet both produce poor results in Figure 4.11, their performance on this document example is unparalleled. Overall MPRNet appears to produce the output that is the most easily readable among all the comparisons. Although our model fails to restore the image and also introduces new artifacts, there is some evidence of restoration in the sharpening of ghosted text and lines.

(a) *Blurred*

(b) DeepDeblur (Nah et al. 2017)

(c) DeblurGAN-v2 (Inception) (Kupyn et al. 2019)

(d) DeblurGAN-v2 (MobileNet) (Kupyn et al. 2019)

(e) DMPHN (Zhang et al. 2019)

(f) MPRNet (Zamir et al. 2021)

(g) HINet (Chen et al. 2021)

(h) Ours

Figure 4.12: Examples restorations of "`text10`" from the dataset of Lai et al. (2016). Pre-trained models are used to evaluate existing methods.

4.3 Discussion

This study sheds light on the choice of loss function for deep learning-based image restoration techniques. We first provided a comprehensive review of the fields of image processing, deep learning, and adversarial networks, which have begun to overlap in contemporary research. As it specifically relates to deblurring, we showed a lack of concrete knowledge surrounding the choice of loss function for training deep image deblurring models. Based on this limitation in the literature, we proposed a comparative study phrased as three research questions:

1. Which content loss functions are the most effective for image deblurring?
2. Without using content losses, do adversarial losses stably converge?
3. How does the combination of content and adversarial losses affect deblurring performance relative to using adversarial loss or content loss in isolation?

To answer these questions, we first defined simple generator and discriminator architectures and a common training procedure. After training generators using each adversarial, content, and combined loss function, we reported quantitative and qualitative results to assess the deblurring performance of each generator model.

As it relates to content losses, the typical choice in the literature is the MSE which directly equates to maximizing PSNR. This can be a problematic loss function due to the lacking correlation between PSNR and perceptible quality to the HVS in some cases. As such, we proposed that MAE could be a viable alternative due to its decreased sensitivity to outliers and noise. We found that generator models trained using MAE produced higher validation metrics and higher quality qualitative results relative to the MSE loss function. Indeed, the qualitative results of the MSE loss exhibit a large number of artifacts relative to MAE suggesting that MAE is more robust to the noise and outlier errors during training. Noting also that spatial MSE and MAE place equal weight on all spectral bands in the image signal (Sims 2020), we believed using a frequency-domain representation could improve the generated image quality

to the HVS. This hypothesis was confirmed experimentally; the MAE loss computed over DCT coefficients resulted in a generator model that outperformed the model trained using spatial MAE in terms of the validation metrics. Notably, the MAE-DCT loss resulted in the generator model that produced the highest overall quantitative metrics over any other model trained in this study.

Following Johnson et al. (2016) and Kupyn et al. (2018), we also hypothesized that perceptual content losses based on pre-trained image classification models could be effective choices as a primary content loss function. This expectation was refuted by the quantitative results, where perceptual losses appeared to perform significantly worse than any of the simpler MAE or MSE loss functions. However, in qualitative assessment, the generator models trained on `block2_conv2`, `block3_conv3`, or the ensemble of VGG-19 activations, produced results that were of perceptibly higher quality than any of the models trained using MAE or MSE. We found that using the `block3_conv3` layer of VGG-19 indeed produces higher quality results than either `block1_conv1` or `block2_conv2`, but also that the ensemble of these three layers produces comparable results to `block3_conv3` in isolation.

Relative to the content losses, we showed that no adversarial loss produced generator models that could effectively restore blurry images. We found that the generator optimized using LSGAN loss produced the best quantitative results over all of the testing datasets, but failed to compare to the lowest quality content losses metrics. Although both the saturating and non-saturating version of the GAN loss collapsed during training, we found that the saturating GAN converged on a model that produces better metrics than did the non-saturating GAN. Notably, both losses collapsed around the same time suggesting a particular portion of the space caused instability in both variations of the vanilla GAN. In training, all but the LSGAN loss produced curves that decayed over the duration. The WGAN-GP loss is the only loss that produced stable metrics during training, the remaining losses produced a large amount of noise in the metrics. It is worth noting, this could be due to the discriminator to generator training ratio of five to one that may produce a more stable gradient signal by merit of a fully converged

discriminator. Qualitatively, the generators produced by adversarial losses contained more checkerboard artifacts from transposed convolutional layers than the generator models trained using content losses. The deblurring performance is also significantly worse as indicated by the poor performance on the edges of the woman's face, collar, necklace, and eyes. The empirical results suggest that adversarial loss alone is not a viable choice for image deblurring models.

Prior deblurring research based on adversarial losses uses combined losses consisting of adversarial and content components. In some cases, there could be more than one content loss (Kupyn et al. 2018) or more than one adversarial loss (Kupyn et al. 2019). We showed empirically that combined losses have a cross-regulatory effect that can help stabilize the training relative to using any constituent loss in isolation. Checkerboard and DC bias artifacts that appeared as a result of training with only adversarial losses and spatial distortions that appeared as a result of training with only content losses were both mitigated somewhat by using adversarial and content losses in combination. Although the MAE-DCT content loss produced the highest quantitative metrics in the study, the generator trained using non-saturating GAN adversarial loss and `block3_conv3` as a perceptual content loss produced a higher quality result subjectively in terms of the qualitative comparison.

4.3.1 Contributions to Theory and Research

This study makes three key contributions to research on deep image deblurring. First, we provide an empirical comparison of adversarial losses in isolation, which has not been done in the context of image deblurring to the best of our knowledge. The current literature applies different adversarial losses to different generator architectures, making it difficult to understand the contribution of the loss to the underlying problem. By holding the generator model, training data, and training parameters constant, we showed that generator models trained using adversarial losses alone fail to adequately learn the deblurring task when tested against real-world data.

We also provide evidence that MAE may be a better content loss function when applied to image restoration tasks than the current standard of MSE. Although the qualitative results were poor relative to other methods in this study, the quantitative results of MAE we persistently higher, both in terms of PSNR and SSIM. State-of-the-art methods trained using MSE over pixel illuminances may benefit from replacing the MSE loss function with MAE during training. Following the suggestion of Sims (2020), we also showed that using the MAE over DCT coefficients could further improve the quantitative performance. Qualitatively, using the MAE-DCT produced blurrier results than the spatial MAE, but also introduced fewer new degradations and artifacts. Ultimately, we showed that using a perceptual content loss as a sole content loss produces perceptibly better results than any of the simple MSE or MAE approaches despite marking lower in terms of quantitative metrics. To the best of our knowledge, no current work attempts to train generator models using solely perceptual content loss, but the results of this study provide a strong argument for further investigation.

Finally, we exhaustively confirm that indeed, the combination of adversarial and content loss has an ensemble effect that produces higher quality perceptive results relative to training generator models from solely content or adversarial loss. Notably, the combined losses did not achieve higher PSNR nor SSIM values on the testing benchmarks despite generalizing better to real-world data. Although prior research has suggested different adversarial loss functions under different mathematical premises, empirically, the vanilla non-saturating GAN produced the best quantitative and qualitative result in this study when combined with a perceptual loss, namely, from `block3_conv3` outputs from VGG-19. This suggests that the state-of-the-art results of prior work may attribute more to the novelties and differences in the generator architecture than the loss functions.

4.3.2 Implications for Practice

Although the models developed in academia frequently achieve state-of-the-art performance on benchmarks, the extension to practice is often less clear. Frequently, state-of-the-art

models can contain large parameter sets that are impractical for applications such as autonomous vehicle operation, mobile phones, and edge computing. Furthermore, the usage of complex methodologies makes implementing some cutting-edge models excessively expensive for practitioners in some cases. In this study, we showed that a simple model consisting only of convolutional layers, activation functions, and residual skip connections is capable of achieving results that are perceptibly good for certain tasks. Although the model in this work did not adapt well to fine-grained tasks, such as text and document deblurring (see Figure 4.12), it exhibited good capability for deblurring large objects, such as human faces (see Figure 4.11) and pedestrians in the GoPro data (see Figure 4.10). This is salient for application areas like autonomous vehicles where accurately being able to identify pedestrians and vehicles is paramount. Kupyn et al. (2018) have shown for instance that object-detection models such ac YOLO can better detect objects in blurry images that have been restored.

In practice, executing large grid searches over deep learning models can be impractically expensive or time-consuming. This work provides empirical evidence that practitioners can use to guide the development of their image restoration pipelines without incurring excessive expense or complexity. Practically, this work showed that the choice of the loss function can impact the learning of the generator model. In particular, we demonstrated that perceptual loss functions that are computed from pre-trained natural image priors result in models that produce perceptibly better results than simple MSE or MAE losses despite measurably worse benchmark results. Furthermore, we demonstrated that adversarial losses alone poorly adapt to the task of image restoration despite not producing as many artifacts as content losses are capable of. Finally, we showed that the combination of perceptual content losses with adversarial losses produces a perceptibly better result, but only marginally. Because the resources and time required to train adversarial models are far greater than that of the simpler content and perceptual losses, this is something practitioners should take into account.

4.3.3 Limitations and Implications for Future Work

Due to the stochastic nature of mini-batch gradient descent optimization algorithms, the choice of random number seed can potentially impact the performance of a training algorithm (Lucic et al. 2018). In this study we held a single random seed constant across all models, ensuring that the training data that the models observed would be the same for each experiment. However, due to computational limitations, results were only computed using one random number seed. To account for the noise in the optimization procedure and produce a more robust result, future work will execute the experiments using between ten and thirty random number seeds. Additionally, the generator and adversarial models studied in this work were kept simple due to the immense overhead of training factorial combinations of models with different losses. Future work based on state-of-the-art generator models can apply the method in this study to continue to advance the understanding of how different loss functions apply to image restoration tasks, specifically, deblurring.

This work investigated the effect of different choices of loss function for a single generator model similar to that of Kupyn et al. (2018) and Gao et al. (2019). The more bleeding-edge innovations in image restoration, such as recurrent models (Tao et al. 2018), patch-hierarchical models (Zhang et al. 2019), attentions mechanisms (Suin et al. 2020), and the like, all focus on architectural components of the generator and omit the usage of adversarial losses. Future work will investigate if (1) the usage of adversarial losses improves the ability of these models in terms of perceptual output, which was shown to be weak for some real-world cases relative to simpler models (see Section 4.2.4) and (2) what the effect of choosing different content losses has on these models. Current state-of-the-art approaches apply an MSE content loss for pixels, but this work provides evidence to suggest that MAE and frequency-domain losses may be worth considering when researching deep image restoration models.

Chapter 5

Conclusion

In this dissertation, we first developed an understanding of how trust influences the perceived risk and benefits from using an AV and the choice to adopt one. We then performed a study to improve image deblurring models by determining which losses are effective on a simple model. This chapter concludes the dissertation by briefly summarizing the contributions of both studies.

5.1 The influence of trust on autonomous vehicle adoption

The study in Chapter 3 aimed to provide an understanding of how trust and enjoyment interact with the adoption of ADS and whether the introduction of a human-computer interface between the ADS and the driver improved the trust in the supporting AI technology. The goal was to develop a system that relied on the AI present in an AV to augment the intelligence of the human driver and improve the driver's trust in the AI technology. The results of this study confirm ten of eleven proposed hypotheses related to ADS. The presented perception augmentation module is also shown to increase the driver's trust in the AI-based ADS. As a result, this increased trust positively impacts the perceived benefits and negatively impacts the perceived risks of using the ADS. Trust also has a positive impact on the intention to adopt autonomous features, as does the propagation of the effects on perceived risks and benefits. As such, there is a strong argument for the development of components like the perception augmentation module alongside consumer ADS to improve human trust in– and adoption of AVs.

5.2 The choice of loss function for deep image restoration

In Chapter 4 we provide a comparative study of the different content and adversarial loss functions applied to the task of deep image deblurring. We show that despite the prevalence of MSE in the literature, MAE may be a stronger alternative as a content loss for non-adversarial methods. We also demonstrate that perceptual losses are effective content losses in the absence of an adversarial loss despite poor performance on testing metrics relative to MAE and MSE content losses. We find consistency with the literature in terms of `block3_conv3` of VGG-19 producing the best perceptive results when applied as a primary or auxiliary perceptual content loss. When combined with an adversarial loss, we notice that the two independent loss functions have a cross-regulatory effect and produce better perceptive results than a generator trained with either loss in isolation. We showed that the simple model in this study generalizes better to real-world natural image blurs than state-of-the-art models, but fails to produce quality results on sparse images like text, fine details, or documents.

5.3 Limitations and Future Work

Limitations of the study in Chapter 3 point out directions for future research. In future projects, the quality of the simulated AV can be improved by replacing our software-in-the-loop simulation with a hardware-in-the-loop simulation by utilizing a model of a vehicle in a physical space. This allows both for deeper immersion from the passenger and more realistic implementations of potential perception augmentation modules. Additionally, the subjective nature of the participants' responses is a limitation of the psychological constructs used in the study. In future work, objective measures can be collected using technology such as eye tracking and brain-computer interfaces that objectively measure visual attention and neural activation from the participants, respectively. These objective measures can provide insight into where the participant is looking and how surprising certain events are to the participant.

Although in Chapter 4 we shed light on how different loss functions affect the training of deep image generators, the study embodies two limitations. First, due to computational overhead, each model was trained using a single random number seed. In the future, results can be generated by training models with 30 different random seeds to filter noise in the metrics and ensure that the current result is not anomalous. Second, the generator model used in the comparative study was kept simple to reduce the impact of over-parameterization on the results. In future work, we will investigate whether the results shown in this study are consistent across different state-of-the-art generator architectures.

Bibliography

Abraham H, Lee C, Brady S, Fitzgerald C, Mehler B, Reimer B, Coughlin JF (2017) Autonomous vehicles, trust, and driving alternatives: A survey of consumer preferences. *Transportation Research Board 96th Annual Meeting, Washington, DC*, 8–12.

Agarwal R, Karahanna E (2000) Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly* 24(4):665–694, ISSN 02767783, URL `http://www.jstor.org/stable/3250951`.

Agarwal R, Prasad J (1998) A conceptual and operational definition of personal innovativeness in the domain of information technology. *Information Systems Research* 9(2):204–215.

Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN, Inkpen K, et al. (2019) Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Arjovsky M, Bottou L (2017) Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862* .

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. *International conference on machine learning*, 214–223 (PMLR).

Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. *arXiv preprint arXiv:1607.06450* .

Badue C, Guidolini R, Carneiro RV, Azevedo P, Cardoso VB, Forechi A, Jesus L, Berriel R, Paixao TM, Mutz F, et al. (2020) Self-driving cars: A survey. *Expert Systems with Applications* 113816.

Bagozzi RP, Yi Y (1988) On the evaluation of structural equation models. *Journal of the Academy of Marketing Science* 16(1):74–94.

Bansal P, Kockelman KM (2018) Are we ready to embrace connected and self-driving vehicles? a case study of texans. *Transportation* 45(2):641–675.

Bayer BE (1976) Color imaging array. US Patent 3,971,065.

Bazilinskyy P, de Winter J (2015) Auditory interfaces in automated driving: an international survey. *PeerJ Computer Science* 1:e13.

Bhattacharya B, Ghosh A, Basu Roy Chowdhury S (2018) Training autoencoders in sparse domain. *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1):8049–8050, URL `https://ojs.aaai.org/index.php/AAAI/article/view/12155`.

Bimbraw K (2015) Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology. *2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, volume 1, 191–198 (IEEE).

Bojarski M, Testa DD, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang J, Zhang X, Zhao J, Zieba K (2016) End to end learning for self-driving cars. *CoRR* abs/1604.07316.

Bollen K, Lennox R (1991) Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin* 110:305–314.

Bonnefon JF, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576.

Bosch LLC (2019) Automated vehicle perceptions study. `https://www.bosch-mobility-solutions.us/us/highlights/automated-mobility/automated-vehicle-perceptions-study/`.

Brown B, Laurier E (2017) The trouble with autopilots: Assisted and autonomous driving on the social road. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 416–429, CHI '17 (New York, NY, USA: ACM).

Carlson A, Skinner KA, Vasudevan R, Johnson-Roberson M (2018) Modeling camera effects to improve visual learning from synthetic data. *Proceedings of The European Conference on Computer Vision (ECCV) Workshops*.

Casner SM, Hutchins EL, Norman D (2016) The challenges of partially automated driving. *Communications of the ACM* 59(5).

Chan CY (2017) Advancements, prospects, and impacts of automated driving systems. *International Journal of Transportation Science and Technology* 6(3):208–216.

Chen L, Lu X, Zhang J, Chu X, Chen C (2021) HINet: half instance normalization network for image restoration.

Choi JK, Ji YG (2015) Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction* 31(10):692–702.

Czolbe S, Krause O, Cox I, Igel C (2020) A loss function for generative neural networks based on watson's perceptual model. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds., *Advances in Neural Information Processing Systems*, volume 33, 2051–2061 (Curran Associates, Inc.), URL https://proceedings.neurips.cc/paper/2020/file/165a59f7cf3b5c4396ba65953d679f17-Paper.pdf.

D G, D S (2005) A practical guide to factorial validity using PLS-graph: Tutorial and annotated example. *Communications of the Association for Information Systems* 16(5):91–109.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. *CVPR09*.

Dikmen M, Burns CM (2016) Autonomous driving in the real world: Experiences with tesla autopilot and summon. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 225–228, Automotive'UI 16 (New York, NY, USA: ACM).

Donaldson T, Dunfee TW (1994) Toward a unified conception of business ethics: Integrative social contracts theory. *The Academy of Management Review* 19(2):252–284, ISSN 03637425, URL http://www.jstor.org/stable/258705.

Dunfee TW, Smith NC, Ross WT (1999) Social contracts and marketing ethics. *Journal of Marketing* 63(3):14–32, ISSN 00222429, URL http://www.jstor.org/stable/1251773.

Endsley MR (2017) Autonomous driving systems: A preliminary naturalistic study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making* 11(3):225–238.

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision* 88(2):303–338.

Fleetwood J (2017) Public health, ethics, and autonomous vehicles. *American journal of public health* 107(4):532–537.

Forgas J (1995) Mood and judgment: The affect infusion model (AIM). *Psychological bulletin* 117:39–66, URL http://dx.doi.org/10.1037/0033-2909.117.1.39.

Fornell C, Larcker DF (1981) Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research* 18(1):39–50.

Gao H, Tao X, Shen X, Jia J (2019) Dynamic scene deblurring with parameter selective sharing and nested skip connections. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gefen D, Karahanna E, Straub DW (2003) Trust and TAM in online shopping: An integrated model. *MIS Quarterly* 27(1):51–90, ISSN 02767783, URL http://www.jstor.org/stable/30036519.

Girshick RB, Donahue J, Darrell T, Malik J (2013) Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR* abs/1311.2524.

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. Teh YW, Titterington M, eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, 249–256 (Chia Laguna Resort, Sardinia, Italy: PMLR), URL http://proceedings.mlr.press/v9/glorot10a.html.

Goodall NJ (2016) Can you program ethics into a self-driving car? *IEEE Spectrum* 53(6):28–58.

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. *arXiv preprint arXiv:1406.2661* .

Goodrich MA, Schultz AC, et al. (2008) Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction* 1(3):203–275.

Gowda N, Ju W, Kohler K (2014) Dashboard design for an autonomous car. *Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 1–4, AutomotiveUI '14 (New York, NY, USA: ACM).

Grudin J (2009) AI and HCI: Two fields divided by a common focus. *Ai Magazine* 30(4):48–48.

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A (2017) Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028* .

Gunning D (2017) Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)* 2.

Guo J, Kurup U, Shah M (2019) Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* 21(8):3135–3151.

Haspiel J, Du N, Meyerson J, Robert Jr LP, Tilbury D, Yang XJ, Pradhan AK (2018) Explanations and expectations: Trust building in automated vehicles. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 119–120, HRI '18 (New York, NY, USA: ACM).

He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 37(9):1904–1916.

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* .

Howard D, Dai D (2014) Public perceptions of self-driving cars: The case of Berkeley, California. *Transportation Research Board 93rd Annual Meeting*, volume 14, 1–16.

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, 448–456 (PMLR).

Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. *European conference on computer vision*, 694–711 (Springer).

Jolicoeur-Martineau A (2018) The relativistic discriminator: a key element missing from standard GAN.

Kalra N, Paddock SM (2016) Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice* 94:182–193.

Kauten C, Gupta A, Qin X, Li H, Bevly D, Jenkins A (2018) A perception augmentation system for autonomous vehicles. *Proceedings of the 2018 Pre-ICIS SIGDSA Symposium* (San Francisco, CA,

USA), URL https://aisel.aisnet.org/sigdsa2018/4.

Kendall A, Gal Y (2017) What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems*, 5574–5584.

Khan SH, Hayat M, Porikli F (2019) Regularization of deep neural networks with spectral dropout. *Neural Networks* 110:82–90.

Khandelwal P, Zhang S, Sinapov J, Leonetti M, Thomason J, Yang F, Gori I, Svetlik M, Khante P, Lifschitz V, et al. (2017) BWIBots: A platform for bridging the gap between AI and human–robot interaction research. *The International Journal of Robotics Research* 36(5-7):635–659.

Kiani Galoogahi H, Fagg A, Huang C, Ramanan D, Lucey S (2017) Need for speed: A benchmark for higher frame rate object tracking. *Proceedings of the IEEE International Conference on Computer Vision*, 1125–1134.

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Koenig-Lewis N, Marquet M, Palmer A, Zhao AL (2015) Enjoyment and social influence: predicting mobile payment adoption. *The Service Industries Journal* 35(10):537–554, URL http://dx.doi.org/10.1080/02642069.2015.1043278.

Koh J, Lee J, Yoon S (2021) Single-image deblurring with neural networks: A comparative survey. *Computer Vision and Image Understanding* 203, ISSN 1077-3142, URL http://dx.doi.org/https://doi.org/10.1016/j.cviu.2020.103134.

Köhler R, Hirsch M, Mohler B, Schölkopf B, Harmeling S (2012) Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. *European conference on computer vision*, 27–40 (Springer).

König M, Neumayr L (2017) Users' resistance towards radical innovations: The case of the self-driving car. *Transportation Research Part F: Traffic Psychology and Behaviour* 44:42–52.

Koo J, Kwac J, Ju W, Steinert M, Leifer L, Nass C (2015) Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9(4):269–275.

Koo J, Shin D, Steinert M, Leifer L (2016) Understanding driver responses to voice alerts of autonomous car operations. *International Journal of Vehicle Design* 70:377.

Kupyn O, Budzan V, Mykhailych M, Mishkin D, Matas J (2018) Deblurgan: Blind motion deblurring using conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8183–8192.

Kupyn O, Martyniuk T, Wu J, Wang Z (2019) Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8878–8887.

Kyriakidis M, Happee R, de Winter JC (2015) Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology and behaviour* 32:127–140.

Lai WS, Huang JB, Hu Z, Ahuja N, Yang MH (2016) A comparative study for single image blind deblurring. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1709.

Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al. (2017) Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Lee JG, Kim KJ, Lee S, Shin DH (2015) Can autonomous vehicles be safe and trustworthy? effects of appearance and autonomy of unmanned driving systems. *International Journal of Human-Computer Interaction* 31(10):682–691.

Li H, Gupta A, Zhang J, Sarathy R (2014) Examining the decision to use standalone personal health record systems as a trust-enabled fair social contract. *Decision Support Systems* 57:376–386, ISSN 0167-9236, URL `http://dx.doi.org/https://doi.org/10.1016/j.dss.2012.10.043`.

Li H, Sarathy R, Xu H (2010) Understanding situational online information disclosure as a privacy calculus. *Journal of Computer Information Systems* 51.

Li H, Sarathy R, Xu H (2011) The role of affect and cognition on online consumers' decision to disclose personal information to unfamiliar online vendors. *Decision Support Systems* 51(3):434–445, ISSN 0167-9236, URL `http://dx.doi.org/https://doi.org/10.1016/j.dss.2011.01.017`.

Li Y, Schwing A, Wang KC, Zemel R (2017) Dualing gans. *Advances in Neural Information Processing Systems* 30:5606–5616.

Lin SC, Zhang Y, Hsu CH, Skach M, Haque ME, Tang L, Mars J (2018) The architectural implications of autonomous driving: Constraints and acceleration. *Proceedings of the Twenty-Third International*

Conference on Architectural Support for Programming Languages and Operating Systems, 751–766.

Lin T, Maire M, Belongie SJ, Bourdev LD, Girshick RB, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context. *CoRR* abs/1405.0312.

Lin Z, Khetan A, Fanti G, Oh S (2020) PacGAN: The power of two samples in generative adversarial networks. *IEEE Journal on Selected Areas in Information Theory* 1(1):324–335.

Lindell M, Whitney D (2001) Accounting for common method variance in cross-sectional research design. *The Journal of applied psychology* 86:114–21.

Loshchilov I, Hutter F (2016) SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* .

Lowry P, Gaskin J, Twyman N, Hammer B, Roberts T (2013) Taking "fun and games" seriously: Proposing the hedonic-motivation system adoption model (HMSAM). *Journal of the Association for Information Systems* 14:617–671.

Lowry PB, Gaskin J (2014) Partial least squares (PLS) structural equation modeling (SEM) for building and testing behavioral causal theory: When to choose it and how to use it. *IEEE Transactions on Professional Communication* 57(2):123–146.

Lucic M, Kurach K, Michalski M, Gelly S, Bousquet O (2018) Are GANs created equal? a large-scale study. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds., *Advances in Neural Information Processing Systems*, volume 31 (Curran Associates, Inc.), URL `https://proceedings.neurips.cc/paper/2018/file/e46de7e1bcaaced9a54f1e9d0d2f800d-Paper.pdf`.

Lucy LB (1974) An iterative technique for the rectification of observed distributions. *The Astronomical Journal* 79:745.

Lugano G (2017) Virtual assistants and self-driving cars. *2017 15th International Conference on ITS Telecommunications (ITST)*, 1–5 (IEEE).

Luo X, Li H, Zhang J, Shim J (2010) Examining multi-dimensional trust and multi-faceted risk in initial acceptance of emerging technologies: An empirical study of mobile banking services. *Decision Support Systems* 49(2):222–234.

Lutin JM, ITE F, Kornhauser AL (2013) The revolutionary development of self-driving vehicles and implications for the transportation engineering profession. *Cell* 215:630–4125.

MacKenzie SB, Podsakoff PM, Jarvis CB (2005) The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *The Journal of applied psychology* 90 4:710–30.

Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Mathieson K, Peacock E, Chin WW (2001) Extending the technology acceptance model: The influence of perceived user resources. *SIGMIS Database* 32(3):86–112, ISSN 0095-0033, URL `http://dx.doi.org/10.1145/506724.506730`.

McAllister R, Gal Y, Kendall A, Van Der Wilk M, Shah A, Cipolla R, Weller A (2017) Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4745–4753.

McKnight DH, Choudhury V, Kacmar C (2002) Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research* 13(3):334–359.

Metz L, Poole B, Pfau D, Sohl-Dickstein J (2016) Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163* .

Mirza M, Osindero S (2014) Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Miyato T, Kataoka T, Koyama M, Yoshida Y (2018) Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* .

Morris DM, Erno JM, Pilcher JJ (2017) Electrodermal response and automation trust during simulated self-driving car use. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, 1759–1762 (SAGE Publications Sage CA: Los Angeles, CA).

Muensterer OJ, Lacher M, Zoeller C, Bronstein M, Kübler J (2014) Google glass in pediatric surgery: An exploratory study. *International Journal of Surgery* 12(4):281–289.

Nah S, Baik S, Hong S, Moon G, Son S, Timofte R, Lee KM (2019) NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Nah S, Hyun Kim T, Mu Lee K (2017) Deep multi-scale convolutional neural network for dynamic scene deblurring. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.

Nees MA (2016) Acceptance of self-driving cars: An examination of idealized versus realistic portrayals with a self-driving car acceptance scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, 1449–1453 (SAGE Publications Sage CA: Los Angeles, CA).

Nguyen TD, Le T, Vu H, Phung D (2017) Dual discriminator generative adversarial nets. *arXiv preprint arXiv:1709.03831* .

Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier gans. *International conference on machine learning*, 2642–2651 (PMLR).

Pfleging B, Rang M, Broy N (2016) Investigating user needs for non-driving-related activities during automated driving. *Proceedings of the 15th international conference on mobile and ubiquitous multimedia*, 91–99 (ACM).

Politis I, Brewster S, Pollick F (2017) Using multimodal displays to signify critical handovers of control to distracted autonomous car drivers. *Int. J. Mob. Hum. Comput. Interact.* 9(3):1–16.

Purohit K, Rajagopalan AN (2020) Region-adaptive dense network for efficient motion deblurring. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(7):11882–11889, URL `http://dx.doi.org/10.1609/aaai.v34i07.6862`.

Raue M, D'Ambrosio L, Ward C, Lee C, Jacquillat C, Coughlin J (2019) The influence of feelings while driving regular cars on the perception and acceptance of selfâĂŘdriving cars. *Risk Analysis* 39, URL `http://dx.doi.org/10.1111/risa.13267`.

Redmon J, Divvala SK, Girshick RB, Farhadi A (2015) You only look once: Unified, real-time object detection. *CoRR* abs/1506.02640.

Redmon J, Farhadi A (2016) YOLO9000: better, faster, stronger. *CoRR* abs/1612.08242.

Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. *CoRR* abs/1804.02767.

Reeves R, Kubik K (2006) Shift, scaling and derivative properties for the discrete cosine transform. *Signal processing* 86(7):1597–1603.

Reijers W, O'Brolcháin F, Haynes P (2016) Governance in blockchain technologies; social contract theories. *Ledger* 1:134–151, URL http://dx.doi.org/10.5195/ledger.2016.62.

Richardson WH (1972) Bayesian-based iterative method of image restoration. *JoSA* 62(1):55–59.

Ringle C, Da Silva D, Bido D (2015) Structural equation modeling with the SmartPLS. *Brazilian Journal Of Marketing* 13(2).

Rosenzweig J, Bartl M (2015) A review and analysis of literature on autonomous driving. *E-Journal Making-of Innovation* .

Rosique F, Navarro PJ, Fernández C, Padilla A (2019) A systematic review of perception system and simulators for autonomous vehicles research. *Sensors* 19(3):648.

Rouibah K, Lowry PB, Hwang Y (2016) The effects of perceived enjoyment and perceived risks on trust formation and intentions to use online payment systems: New perspectives from an Arab country. *Electronic Commerce Research and Applications* 19:33–43, ISSN 1567-4223, URL http://dx.doi.org/https://doi.org/10.1016/j.elerap.2016.07.001.

Salimans T, Kingma DP (2016) Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv preprint arXiv:1602.07868* .

Schoettle B, Sivak M (2014) Public opinion about self-driving vehicles in China, India, Japan, the US, the UK, and Australia. Technical report, University of Michigan, Ann Arbor, Transportation Research Institute.

Schuler CJ, Hirsch M, Harmeling S, Schölkopf B (2015) Learning to deblur. *IEEE transactions on pattern analysis and machine intelligence* 38(7):1439–1451.

Schwarting W, Alonso-Mora J, Rus D (2018) Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems* 1(1):187–210.

Schweitzer F, van den Hende EA (2016) To be or not to be in thrall to the march of smart products. *Psychology and Marketing* 33:830–842.

Shahrdar S, Park C, Nojoumian M (2019) Human trust measurement using an immersive virtual reality autonomous vehicle simulator. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 515–520.

Shariff A, Bonnefon JF, Rahwan I (2017) Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour* 1(10):694.

Shen Z, Wang W, Lu X, Shen J, Ling H, Xu T, Shao L (2019) Human-aware motion deblurring. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5572–5581.

Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *Journal of Big Data* 6(1):60.

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Sims SD (2020) Frequency domain-based perceptual loss for super resolution. *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6 (IEEE).

Srivastava A, Valkov L, Russell C, Gutmann MU, Sutton C (2017) VeeGAN: Reducing mode collapse in GANs using implicit variational learning. *arXiv preprint arXiv:1705.07761* .

Steffens CR, Messias LR, Drews-Jr PJ, Botelho SSdC (2020) Cnn based image restoration. *Journal of Intelligent & Robotic Systems* 1–19.

Su S, Delbracio M, Wang J, Sapiro G, Heidrich W, Wang O (2017) Deep video deblurring for hand-held cameras. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1279–1288.

Suin M, Purohit K, Rajagopalan A (2020) Spatially-attentive patch-hierarchical network for adaptive motion deblurring. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3606–3615.

Sun J, Cao W, Xu Z, Ponce J (2015) Learning a convolutional neural network for non-uniform motion blur removal. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 769–777.

Surden H, Williams MA (2016) Technological opacity, predictability, and self-driving cars. *Cardozo L. Rev.* 38:121.

Tao X, Gao H, Shen X, Wang J, Jia J (2018) Scale-recurrent network for deep image deblurring. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8174–8182.

Teo T, Noyes J (2011) An assessment of the influence of perceived enjoyment and attitude on the intention to use technology among pre-service teachers: A structural equation modeling approach. *Computers & education* 57(2):1645–1653.

Tsai FJ, Peng YT, Lin YY, Tsai CC, Lin CW (2021) BANet: blur-aware attention networks for dynamic scene deblurring.

Udovicic K, Jovanovic N, Bjelica MZ (2015) In-vehicle infotainment system for Android OS: User experience challenges and a proposal. *2015 IEEE 5th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, 150–152.

Ulyanov D, Vedaldi A, Lempitsky V (2016) Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* .

van der Heijden H (2004) User acceptance of hedonic information systems. *MIS Quarterly* 28(4):695–704, ISSN 02767783, URL `http://www.jstor.org/stable/25148660`.

Various (2018) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard, Society of Automotive Engineers, New York, NY, URL `https://www.sae.org/standards/content/j3016_201806/`.

Venkatesh V (2000) Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research* 11(4):342–365, ISSN 10477047, 15265536, URL `http://www.jstor.org/stable/23011042`.

Venkatesh V, Speier C, Morris M (2002) User acceptance enablers in individual decision making about technology: Toward an integrated model. *Decision Sciences - DECISION SCI* 33:297–316, URL `http://dx.doi.org/10.1111/j.1540-5915.2002.tb01646.x`.

Wallace GK (1992) The jpeg still picture compression standard. *IEEE transactions on consumer electronics* 38(1):xviii–xxxiv.

Wang Y, Zhang L, Van De Weijer J (2016) Ensembles of generative adversarial networks. *arXiv preprint arXiv:1612.00991* .

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4):600–612.

Wiener N (1949) *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications* (The MIT Press), ISBN 978-0262730051.

Xie L, Wang J, Wei Z, Wang M, Tian Q (2016) Disturblabel: Regularizing CNN on the loss layer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4753–4762.

Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* .

Xu L, Zheng S, Jia J (2013) Unnatural L0 sparse representation for natural image deblurring. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1107–1114.

Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH, Shao L (2021) Multi-stage progressive image restoration.

Zhang H, Dai Y, Li H, Koniusz P (2019) Deep stacked hierarchical multi-patch network for image deblurring. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5978–5986.

Zhang K, Luo W, Zhong Y, Ma L, Stenger B, Liu W, Li H (2020) Deblurring by realistic blurring. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2737–2746.

Zhou R, Feng C (2017) Difference between leisure and work contexts: The roles of perceived enjoyment and perceived usefulness in predicting mobile video calling use acceptance. *Frontiers in Psychology* 8, URL `http://dx.doi.org/10.3389/fpsyg.2017.00350`.

Zhu JY, Park T, Isola P, Efros AA (2020) Unpaired image-to-image translation using cycle-consistent adversarial networks.