Classifying Images of Unseen Classes with Zero-Shot Learning
(Technical Paper)

An Analysis of Facial Recognition in Black Lives Matter Protests
(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Nikash Sethi
Fall, 2020

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signed: Nikash Sethi Date: November 2, 2021

Approved Date: November 2, 2021
Vicente Ordoñez-Roman, Department of Computer Science

Approved Date November 2, 2021
Catherine Baritaud, Department of Engineering and Society

Perhaps the most promising and fascinating innovation within artificial intelligence in the past few decades is computer vision. State-of-the-art models and algorithms are able to classify images they have never seen before with incredibly high accuracy, and researchers continue to implement and deploy interesting applications of the technology into society. A particularly interesting problem within computer vision is zero-shot, in which models must classify categories of images that have not been trained at all (Wang et al., 2019, p. 1). Traditional computer vision relies on training models on large sets of images, to extract and represent features that they can later match with novel images. In zero-shot learning, though, test images may be from classes that models have never seen before, which presents a unique and open-ended problem that more closely encompasses real-world applications of computer vision. The field of zero-shot learning has seen fascinating advancement in the past few years, and I find the topic futuristic and applicable to a wide range of problems in image classification.

Few-shot learning is a technique closely related to zero-shot learning, and involves scenarios in which only a few image features are available as labeled examples during training (Rahman et al., 2018, p. 5655). Dickson (2020) argues that the most influential application of few-shot learning is facial recognition (para. 4). In this application, algorithms learn representations of faces with very few data points, and are able to classify faces based on these embeddings.

Facial recognition has come under heavy scrutiny over the past few months, as police departments around the country started using the technology to track down activists who participated in Black Lives Matter protests following the death of George Floyd (Rector & Winton, 2020, para. 1). The relationships between police departments, activists, legislators, and technology companies that develop facial recognition algorithms have become increasingly

complicated, and tensions surrounding whether facial recognition should be used in policing have grown considerably. Therefore, it is important to objectively analyze the parties involved and how they interact with each other in order to determine potential solutions to the controversy.

The technical and STS research both look to advance the space of computer vision and specifically zero-shot learning. On the technical side, I will work with Professor Vicente Ordóñez-Roman, an Assistant Professor in the Department of Computer Science, to explore a niche space within zero-shot learning in which classes are paired with and trained on natural language textual descriptions. These text descriptions allow models to extract features without being trained on images that belong to every class, and present a far more extendable and generalizable approach to zero-shot learning than current methods offer. Tightly coupled, the STS research focuses on analyzing the various actors surrounding facial recognition in policing through an Actor-Network Theory framework that Callon (1984) and Latour (1996) pioneered. Through the STS research, I will present an objective overview of the controversy of deploying facial recognition in police departments, which will point toward potential solutions or regulations we can apply on the technology that benefit society as a whole.

Figure 1 depicts a timeline of the work that I have completed and will conduct for the technical and STS research projects. Major deliverables for the STS research project include the statement of topics and the prospectus for STS 4500, and the undergraduate thesis for STS 4600. The undergraduate thesis is composed of an executive summary, a technical report, an STS research paper, and this prospectus. The technical work will begin in Spring of 2021, and will culminate in a journal style paper overviewing my implementation and findings.

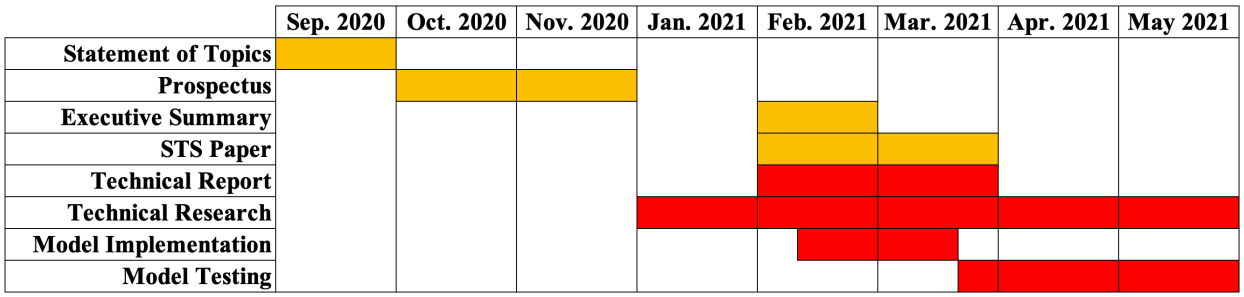

Figure 1: Timelines and Deliverables. The STS research project spans the entire 2020-21 course year, and the technical research will be completed in Spring of 2021. STS work is represented in orange, and technical work is represented in red. (Sethi, 2020).

CLASSIFYING UNSEEN CATEGORIES WITH ZERO-SHOT LEARNING

The field of computer vision has progressed tremendously over the past decade, offering applications in a wide variety of industries. One of the most recent and revolutionary innovations in the space is zero-shot learning, in which models are trained on a set of classes and then tested on images from a completely disjoint set of unseen classes (Wang et al., 2019, p. 1). My research project in zero-shot learning will focus on classifying novel, unseen images of certain categories based on textual descriptions that describe each class.

Computer vision technologies have achieved incredibly high accuracies in classifying known images. State of the art models today are excellent at identifying features of certain images and corresponding them with images from classes they have been trained on in the past. That said, this problem space does not solve many important real-world applications, as it is impossible to train models on images of every possible class. Zero-shot learning accounts for this lack of generalizability by no longer requiring models to be trained on images from each class, and instead relying on attributes or other information that describes classes to which images can belong (Xian et al., 2019, p. 2253). Wang et al. (2019) find that zero-shot learning models are

perfect for scenarios where target classes are large, rare, changing, or bottlenecked in amount of labeled data for training (p. 2).

ZERO-SHOT LEARNING WITH TEXTUAL CLASS DESCRIPTIONS

Many current zero-shot learning approaches require structured lists of attributes that represent important features and qualities of image classes (Xian et al., 2019, p. 2253). However, very few large-scale datasets provide these attribute lists with their images. Using natural language textual descriptions to describe classes offers far more flexibility in conveying important features in classes without having to standardize attributes. These textual descriptions are much more common and attainable than lists of attributes, which is a necessity for scalable zero-shot learning innovation. According to Reed et al. (2016), natural language text is especially advantageous when classifying images with subtle differences across classes (p. 50). In these scenarios, attribute lists would not be able to capture and emphasize slight visual differences between different categories of images.

Over the next semester, I will develop zero-shot learning models and approaches to achieve high accuracy in classifying images from unseen classes, where classes are described with natural language textual descriptions. This specific, niche application of zero-shot learning is fascinating to me because it offers an opportunity to work within a relatively unexplored field of artificial intelligence and computer vision. Through my research, I will be able to gain an understanding of the current state of computer vision in academia and progress the rapidly-growing advancement of zero-shot learning.

In implementing a zero-shot learning approach that uses textual class descriptions, I will be using two key machine learning concepts. First, I will build a natural language component to extract features from textual descriptions of classes. Rahman et al. (2018) offer an approach to

embed vectors of image features and achieve high performance in a generalized zero-shot learning task where test images might be from seen or unseen classes (p. 5652). Second, I will implement a semantic embedding component, which will embed vectors of features of each image with features derived from the natural language component into a semantic space. Li et al. (2019) take a similar approach to this embedding problem, and perform zero-shot learning with a language component that understands the meaning of textual descriptions (p. 8690).

As described in Figure 2 below, my model's training phase will extract features from textual descriptions and learn semantic embeddings of each class. In the testing phase, vector representations of test images will be compared with the semantic space and the image will be classified accordingly. Through this process, the model will be able to categorize images from classes it has not been trained on using embeddings generated solely from textual descriptions.

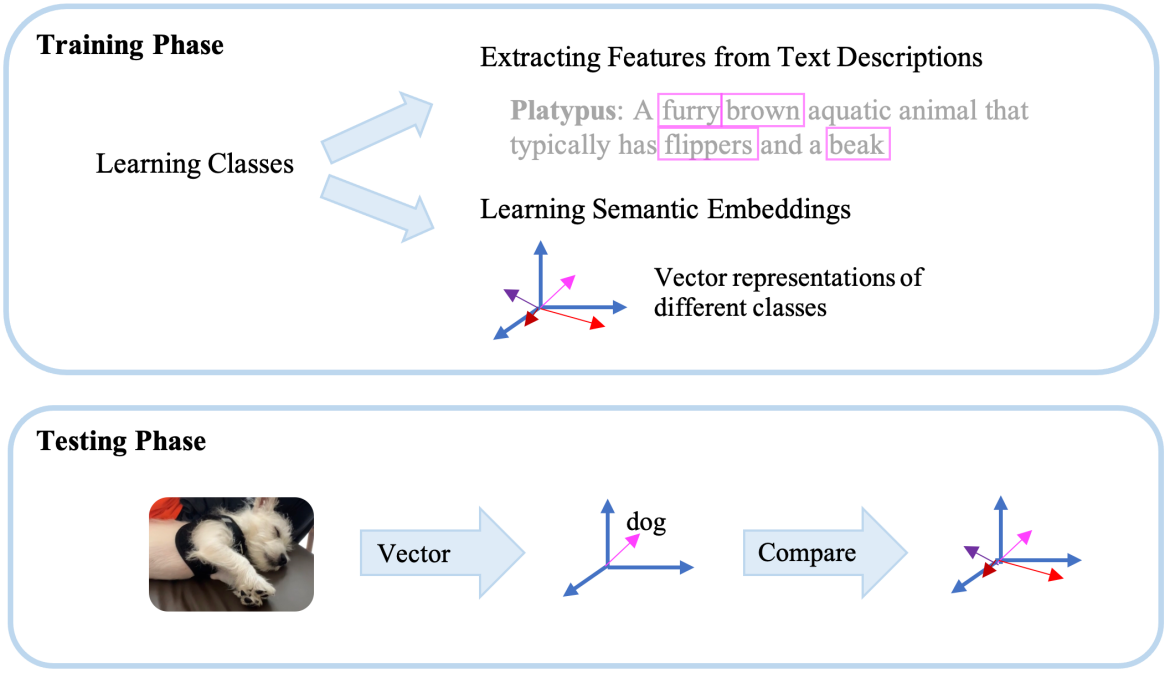

Figure 2: Phases of Text-Based Zero-Shot Learning. The approach to zero-shot learning that I will take will extract features from natural language descriptions and generate semantic vectors in the training phase. In the testing phase, vectors for images will be generated and compared against learned embeddings. (Sethi, 2020).

Throughout this project, I will work with Professor Vicente Ordóñez-Roman in writing a technical report in the form of a journal article. I will primarily be using the ImageNet dataset, which is a hierarchal collection of millions of images from a wide range of classes (Stanford Vision Lab, 2016). In addition, I will test my model's performance on the Animals with Attributes 2 dataset, which is a benchmark dataset that Xian et al. (2019) proposed for zero-shot learning (p. 2251). While this dataset only has 37,000 images, it is accompanied with lists of attributes corresponding to each class of animal (Xian et al., 2019, p. 2256).

After implementing the model, I will test the performance of my approach on the ImageNet dataset according to a series of widely accepted benchmarks that Xian et al. (2019) proposed for zero-shot learning innovation (p. 2260). Xian et al. compute the performance and accuracies of various state of the art methods on small scale datasets like Animals with Attributes 2, and more realistic and practical datasets like ImageNet (p. 2260). By comparing my model's performance with these accuracies, I will be able to gauge the success and applicability of my approach to solving the zero-shot learning problem with textual class descriptions. I hope to achieve classification accuracies that are comparable to or even greater than many of the state-of-the-art implementations of zero-shot learning. Further work that I may explore could involve mitigating bias toward images of seen classes by using generative adversarial networks. Paul et al. (2019) pursue this problem by generating embeddings for unseen classes, and it would be interesting to apply their results to my approach (p. 7050).

FACIAL RECOGNITION IN BLACK LIVES MATTER PROTESTS

Advancements in computer vision technologies, especially those in zero-shot and few-shot learning, have resulted in highly accurate facial recognition technologies that have a wide

range of industrial applications. One of the most popular and recently most controversial applications of this technology is in policing. During Black Lives Matter protests, police departments across the country used facial recognition technologies to track potential criminals (Rector & Winter, 2020, para. 1). This practice has angered activists and propelled legislators to advocate for stricter regulation surrounding the industrial use of facial recognition (Vincent, 2020, para. 5). The STS research project seeks to analyze the use of facial recognition in policing through the Actor-Network Theory framework that Latour (1996) helped define, in order to explore the complex and intricate relationships between various actors and entities involved in this controversy (p. 380).

Industrial applications of facial recognition first became prevalent when American military and intelligence used facial recognition in Iraq and Afghanistan to identify potential terrorists (Williams, 2015, para. 1). Williams (2015) found that police departments across America began to adopt and deploy facial recognition around 2010 in order to identify and track criminals (para. 10). More recently, though, these facial recognition technologies have been found to disproportionately target and harm minorities (Rector & Winton, 2020, para. 20). Fitch (2019) reports on research done by the National Institutes of Standards and Technology that found that a majority of industry-leading facial recognition technologies misidentify Asian and African Americans far more than Caucasians (para. 3). Given recent events surrounding Black Lives Matter movements, many activists are lashing out against unfair policing tactics, which has only compounded opposition against the use of facial recognition technologies in policing.

UNRAVELING FACIAL RECOGNITION IN POLICING

A deeper look into this controversial application of facial recognition reveals a wide range of actors and entities with different perspectives and motivations toward the use of the

technology. Parties involved in this intricate network demonstrate varying levels of support for facial recognition applications in policing. Further, some actors publicly display perspectives contradictory to their practices and actions. The complexity of relationships between different groups that are involved in the controversy of deploying facial recognition for policing purposes demands the need for an in-depth analysis of the network that these actors form.

Through the STS research, I aim to analyze this intricately related network. Beyond understanding the perspectives, attitudes, and goals of each actor, I will analyze how the actors in this space interact, and lay out public and hidden relationships that they have with each other. Ultimately, I hope to unravel these complicated relationships and present a clear view on the environment surrounding the application of facial recognition technologies in policing. To do so, I will be using an Actor-Network Theory framework. This framework was pioneered by Latour (1996) and Callon (1984), and has been applied on a wide range of societal problems and dilemmas related to advancements in technology. Actor-Network Theory is used to understand the role of science and technology in structuring power relationships (Callon, 1984, p. 197). The framework aims to describe societies in a structured and systematic way, by delimiting the identity of actors and the interactions they have (Latour, 1996, p. 370). By interpreting facial recognition technology and its applications in policing through an Actor-Network Theory perspective, we can begin to identify potential solutions and regulations that mutually benefit all involved parties. Latour (1996) argues that applications of Actor-Network Theory offer an objective, black-and-white approach to analyzing complex and intricate networks, which makes it perfectly applicable to study the use of facial recognition in policing (p. 380).

Once we apply the Actor-Network Theory framework to the application of facial recognition in policing, we can begin to understand the perspectives and relationships that each

entity maintains in the larger networks they are a part of. Figure 3 overviews four of the major actors in this network, their primary goals, and their support for or opposition to the use of facial recognition technologies in policing. Police departments are the primary users of facial recognition technologies in this context and have therefore come under heavy scrutiny from activists and lawmakers. Activists advocate for stricter regulation against facial recognition from legislators. These legislators seek to regulate the use and misuse of facial recognition technologies in applications both inside and outside policing. Finally, the technology companies who produce and sell facial recognition algorithms to police departments influence and react to the actions of legislation around facial recognition. The following section of this paper will dive deeper into each of these actors to clarify this complex and intricate network.

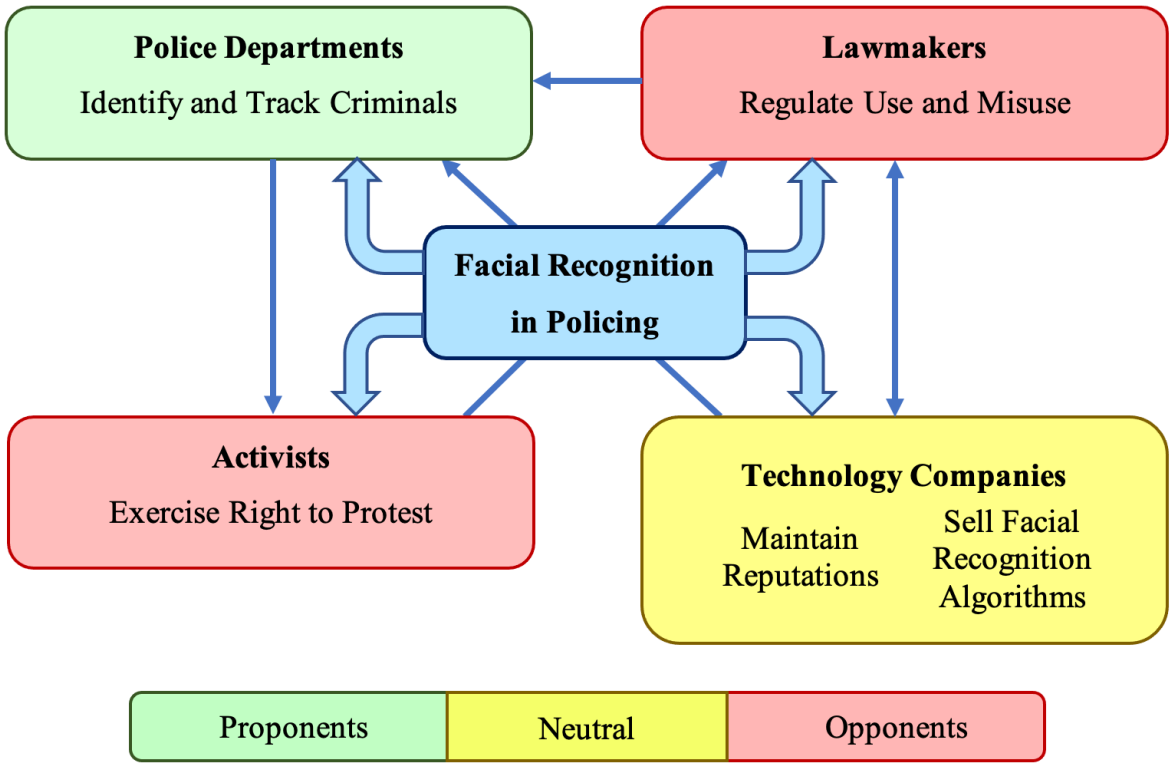

Figure 3: Actor Perspectives on Facial Recognition. Various actors have conflicting motivations and main goals related to facial recognition technologies in policing, which has deterred regulation on this practice from being widely accepted. (Sethi, 2020).

DEFINING THE ACTORS IN FACIAL RECOGNITION

Police departments across the country rely on facial recognition for identifying and tracking suspects. The Los Angeles Police Department has used facial recognition software almost 30,000 times since November of 2009 (Rector & Winton, 2020, para. 1). This reliance continued into the Black Lives Matter protests during the summer of 2020. In one incident, Vincent (2020) reported that the New York Police Department used facial recognition to track down the address of Derrick Ingram, an activist who was accused of assault during a Black Lives Matter protest (para. 2). When using facial recognition, police departments run images of suspects from surveillance cameras against a database of legally obtained mugshots (Rector & Winton, 2020, para. 1). Still, many believe this practice crosses ethical boundaries, and some police departments such as the Boston Police Department decided not to deploy facial recognition practices altogether (Williams, 2015, para. 21).

Police departments obtain these facial recognition products from technology companies. Amazon, Microsoft, and IBM, each large-scale technology companies that research cutting-edge facial recognition algorithms, have all publicly stated they will stop selling facial recognition technologies to police departments (Fowler, 2020, para. 11). According to Fitch (2019), IBM even released a collection of images of people of different races and genders to help train algorithms in less biased ways (para. 9). That said, legislators argue that these public statements hold little promise and are merely tactics to maintain a good reputation. While these technology companies may appear to stand against facial recognition, Fowler (2020) believes they are hesitant to deny the business opportunities of selling the technologies to police departments (para. 22). Despite their voluntary commitments, large technology companies may actually slow legislative efforts to ban the use of facial recognition in policing (Fowler, 2020, para. 4).

It is important to note that large-scale, household name technology companies like Amazon and Microsoft only account for a minor portion of the facial recognition products used by police departments (Fowler, 2020, para. 11). The companies that develop and deploy most of the facial recognition technologies in police departments in America, including NEC Corporation and Clearview AI, have not publicly stated that they will make the same commitments as Amazon and Microsoft (Fowler, 2020, para. 11). Fowler (2020) argues that if these companies continue providing the technology to police departments, new legislation is the only way to stop facial recognition in policing (para. 14).

Advocators push for broad federal regulations to govern applications of facial recognition. Figure 4 outlines the current state of legislation on facial recognition technologies and the regulation that lawmakers are advocating for. Recent legislative efforts are limited to local and state legislation, and do not cover the full scope of facial recognition technologies (Learned-Miller et al., 2020, p. 3). Learned-Miller et al. (2020) propose the creation of a new federal office to regulate facial recognition, hold developers and deployers accountable, and standardize the use of the technology (p. 3). Still, some lawmakers believe facial recognition is far more effective than fingerprinting at identifying criminal suspects (Williams, 2015, para. 2).

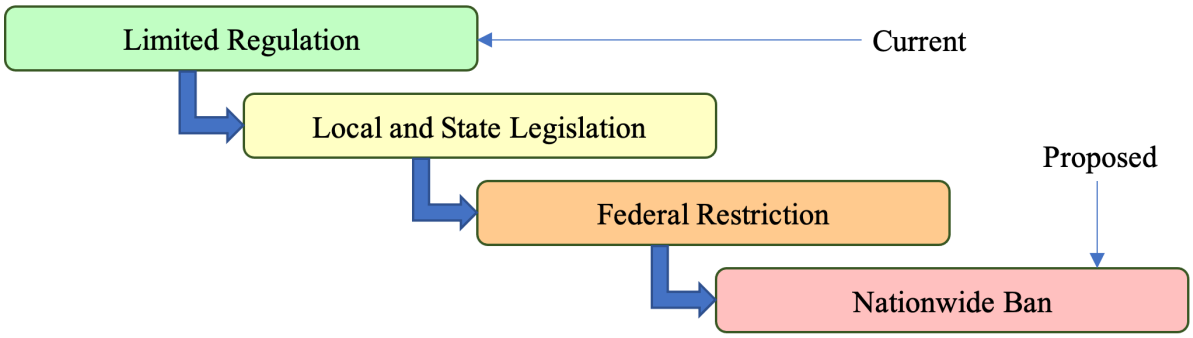

Figure 4: Levels of Facial Recognition Regulation. While current regulation on facial recognition is limited and varies across different regions, activists and lawmakers advocate for stricter laws that ban the technology from being used in policing altogether. (Sethi, 2020).

By viewing each of the above actors through an Actor-Network Theory framework, we can begin to unravel the intricate relationships in the space of using facial recognition for policing. The actors themselves are often far more nuanced than they may appear to the general public, which emphasizes the need for an objective and in-depth analysis of the network. I will present my findings in an STS research paper which overviews each of these actors and the relationships they have with each other, so that readers may raise their awareness on the topic and begin to hypothesize regulations and solutions that best benefit all parties involved.

ETHICALLY ADVANCING APPLICATIONS OF COMPUTER VISION

With the development of any technology, we must consider social and ethical implications in parallel with innovation. Zero-shot learning is the cutting edge of the field of computer vision, and advancing techniques and algorithms used to classify images from unseen categories has far-reaching applications in a variety of different fields. My research in zero-shot learning will explore a niche in computer vision that has not been studied in depth before. As an engineer, though, I cannot disregard the non-technical aspects and implications of this research.

The field of zero-shot learning, and more so few-shot learning, has led to large-scale social controversies in the application of facial recognition in policing. Various actors and relationships are at play in this dilemma, which demonstrates the need for an objective analysis of the goals and perspectives of police departments, technology corporations, activists, and legislators. The tight coupling between my computer vision research in zero-shot learning and my STS research in facial recognition applications in policing will serve as a testament to true engineering, and allow me to innovate in the right way.

References

Callon, M. (1984). Some elements of a sociology of translation: domestication of the scallops and the fisherman of St Brieuc Bay. *The Sociological Review*, *32*(1), 196-233. doi:10.1111/j.1467-954X.1984.tb00113.x

Dickson, B. (2020, August 12). What is one-shot learning? TechTalks. Retrieved from https://bdtechtalks.com/

Fitch, A. (2019, December 19). Facial-recognition software suffers from racial bias. The Wall Street Journal. Retrieved from https://www.wsj.com/

Fowler, G. A. (2020, June 12). Black Lives Matter could change facial recognition forever – if Big Tech doesn't stand in the way. The Washington Post. Retrieved from https://www.washingtonpost.com/

Latour, B. (1996). On actor-network theory: a few clarifications. Soziale Welt, 47(4), 396-381. Retrieved from https://www.jstor.org/

Learned-Miller, E., Ordóñez, V., Morgenstern, J., & Buolamwini, J. (2020). Facial recognition technologies in the wild: a call for a federal office. Retrieved from The Algorithmic Justice League website: https://bit.ly/34GOzou

Li, Z., Yao, L., Zhang, X., Wang, X., Kanhere, S., & Zhang, H. (2019). Zero-shot object detection with textual descriptions. AAAI Conference on Artificial Intelligence, 33(1), 8690-8697. doi:10.1609/aaai.v33i01.33018690

Paul, A., Krishnan, N. C., & Munjal, P. (2019). Semantically aligned bias reducing zero shot learning. 2019 IEEE Conference on Computer Vision and Pattern Recognition, 7049-7058. doi:10.1109/CVPR.2019.00722

Rahman, S., Khan, S., & Porikli, F. (2018). A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. IEEE Transactions on Image Processing, 27(11), 5652-5667. doi:10.1109/TIP.2018.2861573

Rector, K. & Winton, R. (2020, September 21). Despite past denials, LAPD has used facial recognition software 30,000 times in the last decade, records show. The Los Angeles Times. Retrieved from https://www.latimes.com/

Reed, S., Ataka, Z., Lee, H., & Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 49-58. doi:10.1109/CVPR.2016.13

Sethi, N. (2020). Actor perspectives on facial recognition. [Figure 3]. Prospectus (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Sethi, N. (2020). Levels of facial recognition regulation. [Figure 4]. Prospectus (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Sethi, N. (2020). Phases of text-based zero-shot learning. [Figure 2]. Prospectus (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Sethi, N. (2020). Timelines and deliverables. [Figure 1]. Prospectus (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Stanford Vision Lab. (2016). ImageNet. Retrieved from http://www.image-net.org/

Vincent, J. (2020, August 18). NYPD used facial recognition to track down Black Lives Matter activist. The Verge. Retrieved from https://www.theverge.com/

Wang, W., Zheng, V., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: settings methods, and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1-37. doi:10.1145/3293318

Williams, T. (2015, August 12). Facial recognition software moves from overseas wars to local police. The New York Times. Retrieved from https://www.nytimes.com/

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2019). Zero-shot learning – a comprehensive evaluation of the good, the bad, and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9), 2251-2265. doi:10.1109/TPAMI.2018.2857768