11-4-2019

Cascaded Neural Networks for Identification for Identification and Posture-B Threat Assessment of Armed People

Benjamin Abruzzo
CCDC Armaments Center, U.S. Army ARDEC, Picatinny Arsenal, NJ

Kevin Carey

Christopher Lowrance
United States Military Academy

Eric Sturzinger
United States Military Academy

Ross Arnold
CCDC Armaments Center

See next page for additional authors

Authors

Benjamin Abruzzo, Kevin Carey, Christopher Lowrance, Eric Sturzinger, Ross Arnold, and Christopher Korpela

Cascaded Neural Networks for Identification and Posture-Based Threat Assessment of Armed People

Benjamin Abruzzo*, Kevin Carey†, Christopher Lowrance†,
Eric Sturzinger†, Ross Arnold*, and Christopher Korpela†
*Armament Graduate School Picatinny Arsenal, NJ 07806
{benjamin.a.abruzzo.civ, ross.d.arnold4.civ}@mail.mil
†Electrical Engineering and Computer Science United States Military Academy West Point, NY, 10996
{kevin.carey, christopher.lowrance, eric.sturzinger, christopher.korpela}@westpoint.edu

***Abstract*—This paper presents a near real-time, multi-stage classifier which identifies people and handguns in images, and then further assesses the threat-level that a person poses based on their body posture. The first stage consists of a convolutional neural network (CNN) that determines whether a person and a handgun are present in an image. If so, a second stage CNN is then used to estimate the pose of the person detected to have a handgun. Lastly, a feed-forward neural network (NN) makes the final threat assessment based on the joint positions of the persons skeletal pose estimate from the previous stage. On average, this entire pipeline requires less than 1 second of processing time on a desktop computer. The model was trained using approximately 2,000 images and achieved a pistol and person detection rate of 22% and 55%, respectively. The final stage NN correctly identified the severity of the threat with 84% accuracy. The images used to train each stage of our multi-classifier model are available online. With an expanded dataset the accuracy of detecting people and pistols can likely be improved in the future.**

(a) Low-threat (b) Medium-threat (c) High-threat

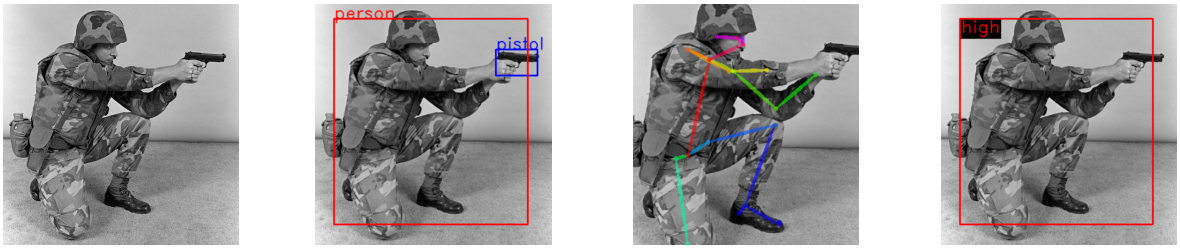
Figure 1: Examples of three different body postures corresponding to level of threat.

The views expressed are those of the authors and do not reflect the official policy or position of the US Army, Department of Defense, nor the US Government. Funding for this research was provided in part by the Combatting Terrorism Technical Support Office and the Office of Naval Research.

I. Introduction

Situational awareness (SA) is a fundamental security cornerstone. When presented with dangerous circumstances, the awareness of an individual to potential dangers is critical to timely and effective decision making. This importance of SA is accentuated in life-and-death scenarios during which a timely response may be key to survival. When considering the dangers posed by firearms, the identification of an armed threat followed by a rapid security response is crucial. In many situations, firearms are explicitly prohibited and the simple presence of a firearm elevates the situation to a critical status. For example, school zones, transportation hubs, stadiums, governmental offices, and most places of work are commonly gun-free zones. If a weapon has been identified within such a zone, a security response should be engaged to neutralize any threats. Using technology, the detection and identification of threats can reduce the amount of harm a bad-actor can inflict.

In contrast to gun-free zones, it can be acceptable or even expected for people to be armed in certain locations. Naturally, in states and countries with open carry laws, citizens are regularly armed but do not present a threat, presenting a challenge when differentiating between bad-actors and ordinary civilians. This has been especially common



(a) Original image (b) YOLO Object Detection (c) OpenPose skeleton (d) Final classification

Figure 2: Stages of the threat classification pipeline beginning with the original image.

during the US war on terror, where armed forces are often exposed to civilian populations intermixed with hostile agents. Similarly, a civil police patrol may be exposed to a population that contains members of organized crime, but also in which the open-carry of firearms is legal. Identifying and discriminating between potential and actual threats is very difficult in these ambiguous situations.

In this paper, we present a deep learning model for the real-time threat classification of people with pistols based on their body posture with the weapon. We developed four threat categories: no threat, low threat, medium threat, and high threat based on the presence of both a pistol and a person as well as the overall body posture of any individual associated with a pistol. Examples of the three latter postures (low, medium, and high) are shown in Fig. 1. The no-threat condition represents a category where a pistol is either not present or is not within reach of anyone. Low threat occurs when a gun is holstered or within reach but is not held/brandished. A subject is determined to be a medium threat if their hand is touching or holding a gun but that gun is not sighted in line with their eyes. High threat is determined to be the case where the gun is actively being aimed and a shot may be imminent. In any of these conditions, our system is intended to be a notification or alert system that draws the attention of security professionals and raises their level of awareness. The detection is designed to be used on a wide range of camera systems ranging from medium resolution security cameras to body cameras to surveillance drones.

The first stage in our procedure is a CNN trained to detect humans and pistols within an image or a scene. If no pistol is detected or is detected but cannot be associated with a person, then no further action is taken; this is considered a no-threat situation. If there is a pistol present in the scene and it can be associated with a person also in the scene, then the second stage of our method uses a another CNN to construct a notional skeleton of the person to determine their posture. After a skeleton is constructed, a reduced order neural network is used to classify the posture as either low-, medium-, or high-threat. This process is depicted in Figs. 2 and 3. The modularity of this methodology allows each stage to designed and calibrated separately, with all networks tied together serially.

The remainder of this article is organized as follows. In section II we discuss the state of the art related to this effort. Section III disusses our method in detail in addition to our process of developing datasets for training each network. The performance of the developed system is provided in Section IV. Finally, we analyze our results in terms of strength and weaknesses in Section V.

II. Related Work

Much of the literature related to assessing the threat level of people with hand guns can be divided into two areas of research. The first area involves detecting guns in images and video, while the second involves the characterization of human poses and body postures.

A. Hand Gun Detection

Hand gun, or pistol, detection has attracted significant attention in research due to its myriad of applications across various security domains, such as airport security and building surveillance. Tiwari et al. employed color-based object segmentation based on a template pistol feature set and an interest point detector to find similarities between objects; a pistol similarity threshold of 50% achieved 84% accuracy [1]. Martinez-Diaz et al. used a three-layer

NN to classify pistols in an image by generating a set of moments on the order of 15 milliseconds via a GPU for each detected object that were invariant to scale, rotation, and translation [2]. Another study compared two classification models: a sliding window and region-based approach - which found that an R(egion)-CNN was the better performing model, achieving an over 84% correct prediciton rate [3].

In the context of airport security, Damashek and Doherty developed a model to detect pistols through x-ray images using Chamfer matching, a parametric edge matching algorithm, which finds the minimal distance between each edge of the object in the x-ray and those in a template pistol image [4]. This model performed well but was severely degraded with slight occlusions, affecting the edge distance calculations. Lai and Maples compared the performance of the VGGNet model along with three instances of a Tensorflow-based implementation of the Overfeat network (each with a unique combination of learning rate and confidence threshold) to not only detect but also classify weapons in images. They were able to achieve 89% accuracy in one of the Overfeat models [5]. These works focus on detection and classification, without further regard to deriving context, such as the meaning of the presence of a gun in the image.

B. Human Pose Estimation

Human pose estimation has also attracted much attention in the past few years due to its wide spectrum of potential applications. Detecting a human in an image or video is now somewhat trivial using one of the previously mentioned object detectors. However, detecting a specific skeletal pose to infer a current action or intent is somewhat more complex. Takai and Miwa sought to infer the current action being taken by a human by first noting temporal changes in pose (assigning normalized values from 0 to 1) and if above a certain threshold, identify the person's action (e.g. picking up an item off the ground) [6]. Toshev and Szegedy were the first to apply deep NNs to human pose estimation. They employed multi-target joint localization regression CNNs in a cascading architecture to leverage higher resolution subimages, ultimately resulting in a much more accurate pose estimation [7]. Cao et al. defined a realtime multi-person 2D pose estimation model, resulting in the public release of OpenPose [8]. The model defines Part Affinity Fields, (PAFs), which represent the position and orientation of each major joint in the body, associating body parts to respective people in images. Shahroudy et al. designed a Recurrent Neural Network (RNN) to model long-term temporal correlation of body part features. They were able to achieve an accuracy of nearly 70% [9].

A different approach by Park et al. sought to classify a handheld device's "pose" or position with the respect to the body (located in the hand, ear, pocket, or backpack) along with the person's walking speed using a regularized SVM [10]. Wei et al. developed a model to predict the skeletal pose of an action from a video using relevant action selection that filtered out irrelevant training data so as to train on only high quality, pertinent image frames [11]. A survey summarized current literature in automatic-behavior-recognition models, focusing on human surveillance and covered different system components from raw image processing and object classification to recognizing abstract events like car theft and fighting [12]. Kelley et al. presented a pseudo two-stage system using a hidden Markov model (HMM) to identify activity and infer intent of another entity by training on the observing entity's own previous actions and intent [13]. Their ultimate objective was to infer human intent from quantitative measurements like relative angle of movement and distance between two entities.

While there are many studies on object detection, classification, and pose estimation, we are unaware of any studies considering a combined workflow of detection and intent classification based on the presence of the object.

III. Approach

We are considering a constrained situation in which the presence of firearms is ubiquitous enough that simply detecting a weapon is not immediately concerning. As stated earlier, this may arise in areas where people tend to be legally armed or in a military situation in which a patrol may encounter armed residents with an uncertain intent. Within these environments, we seek to establish increasing levels of alertness that correspond with the position

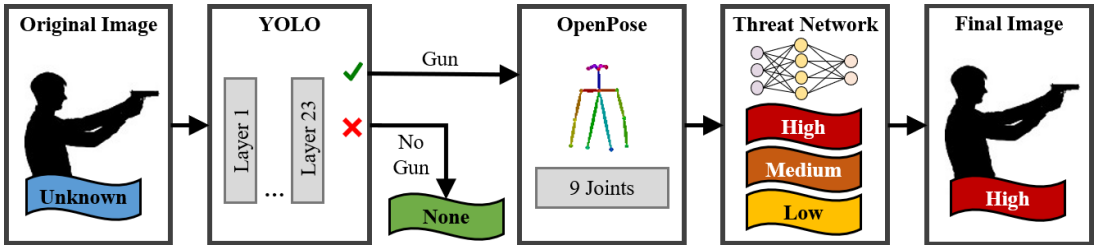


Figure 3: Top-level flow of threat classification

an armed actor's weapon hand. Pointing a gun forward likely represents a greater threat than pointing it down.Fig. 3 illustrates the flow of our method. The process begins with a digital image. The source of image is not specified and could be a security feed, a body-camera, or a camera on an autonomous agent such as a drone. For demonstrative purposes, we are limiting the search in this study to the detection of pistols, though these results are generalizable to other hand-held weapons.

The first step in the process is a person/pistol detection stage. Using a newly developed dataset, we trained a small version of the You-Only-Look-Once (YOLO-tiny) [14] neural network to classify pistols and people in images. The result of this process is a labeled array of bounding boxes corresponding to each detected person and pistol. An example of this result is shown in Figure 2b. After detecting the bounding boxes around all pistols and people present in the image, any pair of pistol/person bounding boxes that overlapped were considered to be associated. Without the presence of any overlaps, we consider the scene safe. If there are any associations, our method would extract a sub-image defined by the bounding box around the associated person for further evaluation.

The second stage uses the CMU *OpenPose* package to generate a notional skeleton as a simplified representation of the person [8]. This package was used without modification and was implemented using the provided pre-trained weights and skeleton models. Figure 2c illustrates the output of the OpenPose CNN package using the BODY25 model which is an augmented version of the 17-key-point Common Objects in COntext (COCO) model [15]. Alternative models and weights are included in the OpenPose repository, though each is associated with a decrease in speed, accuracy, or both.

Finally, the resulting skeleton was processed using a neural network to classify the body language of the person as a low-threat, medium-threat, or a high-threat. This network is discussed further in Section III-C.

A. Building and Compiling the Datasets

A significant challenge to implementing these neural networks robustly is a lack of high-quality, labeled datasets to train the pistol detector or the pose classifier. Raw images were gathered using a Google image search for terms such as "open-carry", "holding pistol", "firing pistol", etc. A human then evaluated those downloaded images to remove irrelevant images, label bounding boxes around pistols and people within the relevant images, and separated any humans holding guns into the High/Medium/Low categories based on their perceived threat. For a dataset of images of humans, we used the "person" tagged images from the Common Objects in Context (COCO) image dataset. The pistol and person images were then combined with images from the COCO dataset that contained neither to introduce negative training samples. The final dataset contained approximately 2,000 pistol images, 56,000 people images, and 60,000 with neither. After compiling and labeling the images, 15% of each class was randomly set aside for testing while the remaining 85% were used to train the pistol and person detection weights.

B. Training the Human & Pistol Detecting Network

The labeled dataset was used to train the detection weights of a YOLO network for only the two classes of people and pistols. The YOLO framework chosen because of the ease of implementation, speed, and reliability of the detection compared to other object detection networks (Fast R-CNN, R-FCN)[16]. The specific YOLO-tiny network was chosen due to its relative lower computational requirements when

compared to the full YOLOv3 network (23 layers vs 106 layers respectively). The size of the network was important because of the application base. If deployed on a drone, a small GPU is necessary due to the weight constraints of the drone, thereby decreasing the computational power accessible. YOLO operates using a 13x13 grid structure where each grid square is responsible for up to 5 objects. This structure allows for fast detection, but also allows the network to get overloaded when there are multiple classifications within a single cell leading to a failure point in the detection. For our application, we trained the YOLOv3 tiny network to detect both people and guns within approximately 300,000 iterations of training.

C. Training the Skeleton Network

The images that were previously sorted into high, medium, and low threat postures were used with the pre-trained CMU OpenPose detector to calculate the pixel locations of the joints of the person. This skeleton was then sliced to only evaluate certain joints determined to be key points; the hips, shoulders, elbows, wrists, neck. The pixel location of the pistol was also included as an additional joint location. These were thought to be the most important points as a shot can be fired from almost any leg position, but the torso, arms, and head have specific configurations for accurate shooting. The subset of skeletal points were then inputs to a small feed-forward neural network taking the 10x2 dimensional input to a 3 dimensional output (high threat, medium threat, low threat) with a single hidden layer of 8 nodes, a learning rate of 2.5e-3, and a training keep probability of 0.5.

There was significant pre-processing of the data before it was input into the neural network to decrease the number of training samples needed and to increase the robustness of the classification. First, the right elbow was set as the origin and then the rest of the values were normalized such that the magnitude of the right elbow to the right wrist was one. If the direction of the elbow to wrist vector has a negative x component (i.e. the right arm is pointed to the left of the picture), the picture is then flipped such that the elbow to wrist vector has a positive x component. These steps are taken to decrease the variance of possible positions thereby decreasing training time and samples necessary for robust results. The advantage to this system is it is very fast (0.07 seconds per prediction) and does not take significant computational overhead. The disadvantage is it is only able to predict effectively when all of the skeletal points described above are visible. With a larger dataset, partial skeletons could be evaluated as well.

Table I: Comparison of YOLO-tiny and YOLOv3 for detecting images with people and pistols in varying levels of threatening posture. Numbers reported are a count of images where at least one object was detected from the 300 source images.

	Object	High	Med.	Low
YOLO-tiny	Pistols	215	104	79
	People	244	272	288
YOLOv3	People	297	298	296

IV. Results

A. Object detection using YOLO

To evaluate the performance of the YOLO-tiny detector, 300 manually classified images were randomly selected from a database of high-, medium-, and low-threat images each, leading to a test set of 900 images. Table I presents the results of YOLO-tiny and YOLOv3 for detecting pistols and people detecting pistols and people from the 300 images of each threat-level class presented to the detectors. Detections using the YOLOv3 network were calculated using the pre-trained weights from the COCO object dataset. The weights for the YOLO-tiny network were generated using the method described in Section III-A. The results in Table I do not reflect the association of both people and pistols, just the ability of the detector to locate the presence of each object in the images. On average, the YOLOv3 network took 0.024 seconds longer to process each image. This was calculated on a desktop workstation with an Intel i7-8700K processor, and NVIDIA Quadro P4000 graphics card, and 16 Gigabytes of RAM.

B. Skeletal Threat

The OpenPose skeletal detection model was highly robust and was able to find any person that YOLO was able to detect. After 30 training epochs

Table II: Confusion matrix of skeletons classified by the feed-forward neural network.

		Predicted			
		High	Med	Low	Total
Actual	High	62	2	0	64
	Med.	6	26	17	49
	Low	0	2	52	54
	Total	68	30	69	167

Table III: Proportion of predicted values given the actual values (recall) rounded to the nearest whole number

		Predicted			
		High	Med.	Low	Total %
Actual	High	97%	3%	0%	38%
	Med.	12%	53%	35%	29%
	Low	0%	4%	96%	32%

with a batch size of 64, the conversion from skeletal pose to threat class had a validation accuracy of 84% overall. Table II shows the confusion matrix presenting true- and false-positives for the skeleton-based threat classification. This data was developed using the pre-classified images from the dataset developed for this study. The first row contains the classification results of the human designated high-threat images. In this case, 62 images were correctly identified as high-threat and 2 images were misclassified as either medium-threat. This shows that there is a high level of differentiation between low- and high-threat evaluations as no high-threats were classified as low-threat and vice-versa. The medium-threat skeletons were difficult to discern as they bridge between the two extremes, likely being more similar to low-threat skeletons than high-threat.

The proportion of predicted classes as a function of actual values (also known as recall) is given in Table III. This table displays the locations and degree of errors in prediction, which is effective for understanding the robustness of the system. From this, correctly predicting all classes were effective (97%, 53%, and 96% for high, medium, and low respectively). The medium threats had a moderate likelihood of being predicted as a low threat, however, they were more likely to be predicted as medium. The final column presents the percentage

Table IV: Proportion of actual values given the predicted values (precision) rounded to the nearest whole number

		Actual			
		High	Med.	Low	Total %
Predicted	High	91%	9%	0%	41%
	Med.	7%	87%	7%	18%
	Low	0%	25%	75%	54%

of each type of image of the set of images used (38% of the images were actually high-threat). Table IV shows how frequently a classification is actually correct, or the precision value. This gives an idea of how reliable a prediction is when the outcome is unknown. For example, of the 30 images that were classified as medium-threat, 26 of them actually were medium-threat (87%). These values approximately fit expectations as the recall was higher than random chance. Again, the final column is the percentage of images that were classified as each category out of all of the 167 images tested. This does show a slight bias to predict low as the proportion of low predictions is higher than the proportion of actual low values.

Using the same computer as described in Section IV-A, The entire pipeline was evaluated for computation time. 152 images were used which contained known high-threat individuals that would be detected by YOLO-tiny and successfully classified by the pose estimation network. On average, each image required 0.69 seconds to process the entire pipeline with a standard deviation of 0.12 seconds.

V. Conclusions and Future Work

In this paper, we presented a multi-stage threat-level classifier for the near real-time detection of people with handguns. Within the field, this study combines both context and pose to evaluate threat. Most other research in the field is focused on pose or context solely and not the other, allowing our system to become more nuanced through the combination of factors. Using this, our network was able to classify at a high rate of 84%, with both high and low threats being highly distinct from each other (97% and 96% recall rates respectively). The largest deficiency of these results is derived from the dataset available. The most important and

difficult piece of training the YOLO network was recognizing guns; the current configuration was only able recognize guns in particular orientations and struggled to locate holstered guns. We strongly believe this is because of the limits of our pistol dataset. By increasing the number of images as well as augmenting the dataset to rotate the pistols more, we should achieve an increased detection rate by increasing variability of the possible orientations of pistols.

With regard to classification of the threat severity, Table IV shows that when the system predicts the threat severity of a skeleton there is a high probability of being correct. However, these values also display a bias of the system. By comparing the total percentage of low-threat predictions, 54%, to the total percentage of low-threat images in Table III, 32%, the predicted percentage is significantly higher than the actual percentage. We believe that this bias to predict medium-threat as low-threat is due to the inherent similarities between a medium-threat and a low-threat, which caused even our human classifiers issues when labeling. The difference between having a hand near a holstered pistol, such as someone in mid-stride, and someone actively reaching for/touching their pistol is difficult to discern, likely causing this discrepancy in prediction.

Separately, we also constrained the classification to require the completely populated subset of skeleton joints. No attempt at classification was made on images without a complete torso, reducing the accuracy where a classification could have been tried. Rather than removing these images from our dataset, we intend to study how to best include partial skeletons into the training process to improve classification accuracy. Another method that may increase accuracy would be to add an additional class to account for the middle ground between medium- and low-threat postures. By adding a "moderate-threat" class for when the hand is in close proximity to the pistol, there might be a higher level of delineation between classes leading to a higher accuracy.

Additional Resources

For accessing our image dataset: tinyurl.com/threat-data. The code used for this article can be found here: github.com/westpoint-robotics/threat_detection

References

[1] R. K. Tiwari and G. K. Verma, "A Computer Vision based Framework for Visual Gun Detection Using Harris Interest Point Detector," in *Procedia Computer Science*, vol. 54. Elsevier, 2015, pp. 703–712.

[2] S. Martinez-Diaz, C. A. Palacios-Alvarado, and S. M. Chavelas, "Accelerated pistols recognition by using a GPU device," in *Proceedings of the 2017 IEEE 24th International Congress on Electronics, Electrical Engineering and Computing, INTERCON 2017*. Institute of Electrical and Electronics Engineers Inc., 10 2017.

[3] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66–72, 1 2018.

[4] A. Damashek and J. Doherty, "Detecting Guns Using Parametric Edge Matching," Tech. Rep.

[5] J. Lai and S. Maples, "Developing a Real-Time Gun Detection Classifier," Tech. Rep.

[6] M. Takai, "Extracting Method of Characteristic Posture From Human Behavior for Surveillance Camera," Tech. Rep., 2009.

[7] A. Toshev and G. Christian Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," Tech. Rep.

[8] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January. Institute of Electrical and Electronics Engineers Inc., 11 2017, pp. 1302–1310.

[9] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," 4 2016. [Online]. Available: http://arxiv.org/abs/1604.02808

[10] J.-g. Park, A. Patel, D. Curtis, S. Teller, and J. Ledlie, "Online pose classification and walking speed estimation using handheld devices." Association for Computing Machinery (ACM), 9 2012, p. 113.

[11] S.-E. Wei, N. C. Tang, Y.-y. Lin, M.-F. Weng, and H.-Y. M. Liao, "Skeleton-augmented Human Action Understanding by Learning with Progressively Refined Data." Association for Computing Machinery (ACM), 11 2014, pp. 7–10.

[12] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behavior-recognition algorithms," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 206–224, 3 2010.

[13] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, and G. Bebis, "Understanding human intentions via hidden markov models in autonomous mobile robots." Association for Computing Machinery (ACM), 3 2008, p. 367.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 6 2015. [Online]. Available: http://arxiv.org/abs/1506.02640

[15] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," 5 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[16] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 4 2018. [Online]. Available: http://arxiv.org/abs/1804.02767