Network-Aware 5G Edge Computing for Object Detection: Augmenting Wearables to “See” More, Farther and Faster

ZHONGZHENG YUAN1, (Student Member, IEEE), TOMMY AZZINO1, (Student Member, IEEE), YU HAO1, YIXUAN LYU1, HAOYANG PEI1, ALAIN BOLDINI2, MARCO MEZZAVILLA1 (Senior Member, IEEE), MAHYA BEHESHTI2,3 (Student Member, IEEE), MAURIZIO PORFIRI2,4 (Fellow, IEEE), TODD HUDSON5,6,7, WILLIAM SEIPLE5,6, YI FANG1, SUNDEEP RANGAN1 (Fellow, IEEE), YAO WANG1,4 (Fellow, IEEE), J. R. RIZZO2,3,4,7

1Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201, USA
2Department of Mechanical and Aerospace Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201, USA
3Department of Rehabilitation Medicine, NYU Langone Health, New York, NY 10016, USA
4Department of Biomedical Engineering, Tandon School of Engineering, New York University, NY 11201, USA
5Lighthouse Guild, New York, NY 10023, USA
6Department of Ophthalmology, NYU Grossman School of Medicine, New York, NY 10016, USA
7Department of Neurology, NYU Langone Health, New York, NY 10016, USA

∗Equal contribution

Corresponding author: Zhongzheng Yuan (e-mail: zy740@nyu.edu).

This work was supported in part by the NSF grant 1952180 under the Smart and Connected Community program as well as the industrial affiliates of NYU WIRELESS. In addition, Azzino, Mezzavilla and Rangan were supported by NSF grants 1925079, 1564142, and 1547332 and the Semiconductor Research Corporation (SRC). Yuan and Wang were also supported by NSF grant 2003182. Yu Hao was also partially supported by NYUAD Institute (Research Enhancement Fund - RE132).

ABSTRACT Advanced wearable devices are increasingly incorporating high-resolution multi-camera systems. As state-of-the-art neural networks for processing the resulting image data are computationally demanding, there has been growing interest in leveraging fifth generation (5G) wireless connectivity and mobile edge computing for offloading this processing to the cloud. To assess this possibility, this paper presents a detailed simulation and evaluation of 5G wireless offloading for object detection within a powerful, new smart wearable called VIS4ION, for the Blind-and-Visually Impaired (BVI). The current VIS4ION system is an instrumented book-bag with high-resolution cameras, vision processing and haptic and audio feedback. The paper considers uploading the camera data to a mobile edge cloud to perform real-time object detection and transmitting the detection results back to the wearable. To determine the video requirements, the paper evaluates the impact of video bit rate and resolution on object detection accuracy and range. A new street scene dataset with labeled objects relevant to BVI navigation is leveraged for analysis. The vision evaluation is combined with a detailed full-stack wireless network simulation to determine the distribution of throughputs and delays with real navigation paths and ray-tracing from new high-resolution 3D models in an urban environment. For comparison, the wireless simulation considers both a standard 4G-Long Term Evolution (LTE) carrier and high-rate 5G millimeter-wave (mmWave) carrier. The work thus provides a thorough and realistic assessment of edge computing with mmWave connectivity in an application with both high bandwidth and low latency requirements.

INDEX TERMS mobile edge computing, millimeter wave, 5G wireless, smart wearables, mobile machine vision, deep learning, object detection
I. INTRODUCTION

The technology in smart wearables is advancing rapidly with an increasing integration of rich camera and sensor data [1], [2]. At the same time, there has been remarkable progress in machine vision technologies for processing this visual information. A key challenge of deploying advanced machine vision algorithms in the wearable setting is that state-of-the-art deep neural networks are computationally demanding, particularly for mobile devices that are limited in power and processing constraints for high-resolution images [3].

Mobile edge computing combined with massive mobile broadband capabilities of fifth generation (5G) cellular wireless systems offers the possibility of offloading these computationally intensive vision processing tasks to the cloud [4], [5]. Importantly, 5G systems can leverage the millimeter-wave (mmWave) bands which afford vastly greater spectrum for higher-rate and lower-latency connectivity compared to standard 4G ones [6]–[8]. With mmWave connectivity, a mobile device or wearable can upload high-resolution video data to edge servers, where much greater computational processing can be performed. Wireless offloading can thereby reduce the power consumption and weight of the wearable and also enable support for multiple cameras for an enlarged field-of-view. Cloud connectivity may also provide real-time access to data from other users, converging on new cooperative service strategies.

In this work, we study the potential of wireless offloading of machine vision processing within a powerful, smart wearable for the Blind-and-Visually Impaired (BVI). The platform, called VIS^4ION (Visually Impaired Smart Service System for Spatial Intelligence and Navigation) [9]–[16] is a human-in-the-loop, sensing-to-feedback advanced wearable system that supports a host of micro-services during BVI human-in-the-loop, sensing-to-feedback advanced wearable System for Spatial Intelligence and Navigation) [9]–[16] is a platform, called VIS^4ION wearable: of the mmWave offloading system for the smart VIS^4ION wearable.

Our analysis follows four main steps, each of which bears significant new contributions to handle the unique nature of the mmWave offloading system for the smart VIS^4ION wearable:

• Creation of the NYU-NYC StreetScene dataset: First, to evaluate object detection, we curated a custom dataset, NYU-NYC StreetScene, of high-resolution videos taken during the ‘Last Mile’ pedestrian segment of commuting in NYC. The videos were manually annotated with objects specific to BVI navigation. A depth estimation method was developed for selected objects (standing people) so that the distance of the objects could also be estimated – key to assessing detection range. The dataset is made public, and is itself a contribution of the work [27].

• Evaluation of impact of video resolution and bit rate on object detection: Using the dataset, we conducted an extensive study to evaluate the impact of video...
resolution and bit rate on the object detection accuracy and the reliable detection range. As discussed in the prior work section, previous analyses such as [21], [22] considered only low-resolution images, and did not explicitly study detection range. Also, work such as [28] did not consider the effect of compression.

Wireless network evaluation: We next conducted detailed, realistic wireless network simulations of end users engaged in ’Last Mile’ pedestrian commuting similar to those from which the video was captured. To assess the unique capabilities of 5G, we simulated both a 5G mmWave carrier at 28 GHz, and a traditional 4G Long Term Evolution (LTE) carrier at 1.9 GHz. To accurately predict propagation at both frequencies, we used state-of-the-art ray tracing [29] combined with new highly detailed 3D models acquired from GeoPipe [30]. The use of such detailed models in wireless simulation is the first of its kind. The channel data was integrated into a widely-used end-to-end, full-stack network simulator, ns-3 [31], that captures blocking, beam tracking, the full 4G and 5G protocol stacks as well as realistic delays in the core network and edge cloud.

Performance analysis and availability: Combining the video and wireless analysis with inference times, we compared the performance of three scenarios: local processing only, offloading with LTE, and offloading with 5G mmWave + LTE. For each option, we were able to determine key performance numbers such as the object detection accuracy, range of detection, number of cameras that can be supported, and end-to-end latency. Also, since channel quality is variable, we determined the percentage of time that the performance values can be obtained for a given scenario. We further evaluated the performance achievable with an adaptive offloading scheme, which switches between edge and local computing and varies the video resolution based on the wireless throughput. In summary, we completed a thorough assessment of combining state-of-the-art machine vision and mobile edge computing for contemporary advanced wearables.

RELATED PRIOR WORK

With the growing use of computationally intensive deep learning methods for vision tasks, there has been significant work in studying offloading of this computation via edge computing; see, e.g., [18] for an excellent recent survey. However, very few consider the unique challenges of high-
resolution images transmitted over massive broadband 5G links as needed by the VIS4ION system. For example, some of the most recent works are as follows:

- Edge-AI [19] studies edge computing for image classification. Since the study is on relatively limited data rate 4G links, the work studies dynamically partitioning the layers in a CNN for vision classification. In this work, we assume the processing is entirely done at the edge or local device.
- mVideo [20] considers offloading large batches of surveillance images to the cloud for face detection. This work also only considers 4G.
- Hochsteetler et al. [32] studies inference time on a very low-power edge device without access to cloud resources at all.
- Liu et al. [21] consider edge assisted object detection in mobile Augmented Reality (AR) applications. To meet the stringent delay requirement, they combine edge computing for object detection and fast local object tracking, which we adopt in our work as well. They also propose slice-based processing so that video transmission and object detection can be run sequentially over successive slices to reduce the edge computing delay. Their simulations only consider indoor WiFi connections between the VR headset and the nearby server, while we consider users walking through urban streets using 4G and 5G cellular networks. They also investigated the effect of video resolution and bit rate on the object detection accuracy and processing delay. However, they only examine resolutions up to 720P and consider the faster R-CNN object detector, while we consider resolutions up to 2.2K and focus on the more popular YOLO detector.
- Ran et al. [22] also study edge computing for AR applications. They assume the local processor (smartphone) can run either a tiny-YOLO or a big-YOLO model, while the edge server only runs the big-YOLO model. They consider the trade-off among decision variables including spatial resolution, frame rate, mobile power consumption, edge vs. local processing, object detection model through measurement studies and propose a measurement-driven optimization framework to determine the optimal setting for these decision variables to optimize a weighted average of the detection accuracy and frame processing rate, under the delay, bandwidth, and power constraint. However, the resolution range considered in their study is very low (only up to 480×480). They also do not simulate real wireless networks.
- Jiang et al. [33] propose methods to adapt frame size and frame rate as well as detection models to meet a target detection accuracy, based on video content. They leverage the spatial (cross cameras) and temporal correlation of the optimal configuration to reduce the computation cost of profiling. This work however does not consider the impact of video compression (and hence the bit rate).
- Huang et al. [28] consider the impact of frame size and frame rate (as well as detection model configuration) on the object detection accuracy. This work is not evaluated in the context of edge computing and hence does not consider the effect of compression.

II. THE VIS4ION SYSTEM AND OFFLOADING ARCHITECTURE

A. MOTIVATION

Immobility is a fundamental challenge for persons with BVI [34]. Loss of sight leads to loss of spatial cognition [35]. Spatial cognition can be defined as the knowledge or cognitive representation of the structure, entities/objects, and relationships within space [36]. The overwhelming majority of physical spaces are not visually accessible or do not allow for safe and efficient travel [37], [38]. This limits spatial cognition for the blind and leads to inefficiencies and peril during navigation. This gap requires new tools to bridge such accessibility barriers and to promote independence in daily tasks. A host of assistive technologies for citizens with BVI have been proposed. The white cane [10] is arguably the most widely used [39], [40] and affordable tool, but its short perceptive range limits its function as a direct extension of physical touch [41]–[46]. High-tech hardware-based wearable devices have been developed to provide assistive features such as outdoor navigation [47]–[49]. However, they are generally either high cost (often exceeding $1,000 for advanced devices [50]–[52]) or overly cumbersome. In contrast, software-based solutions that run on ordinary smartphones are more affordable and accessible for BVIs. For example, Microsoft Seeing AI [43] and Blind Square [44] are widespread sensory substitution and navigation apps. However, these applications are not capable of offering advanced computer-vision-based assistive services or features due to the smartphones’ limited on-board computing power.

B. THE CURRENT VIS4ION SYSTEM AND ITS LIMITATIONS

To address these challenges, we recently developed VIS4ION, a Visually Impaired Smart Service System for Spatial Intelligence and Navigation [9]–[11], [13], [14], [53]. The system is implemented as a mobile sensor-to-feedback wearable device in the form of an instrumented book-bag – See Fig. 2]. This smart service system is capable of real time scene understanding with human-in-the-loop navigation assistance, supporting both mobility and orientation [12], [17]. VIS4ION has four components: (1) distance and ranging/image sensors scaffold-ed into the shoulder straps of a backpack; these sensors (including a stereo camera and an Inertial Measurement Unit) extract pertinent information about the environment; (2) an embedded system (micro-computer) with both computing and communication capability (inside backpack); (3) a haptic
FIGURE 2: VIS$^4$ION backpack: The current version of the VIS$^4$ION wearable jacket from [10]–[13], [16], [17] is an instrumented book bag with a single ZED camera, NVIDIA Jetson GPU for local vision processing, and haptic and audio feedback. In this work, we consider augmenting it (proposed augmentation shown in blue) with three additional cameras for rear, left and right visibility and wireless connectivity for edge processing.

interface (waist strap) that communicates spatial information computed from the sensory data to the end-user in real time via an intuitive, torso-based, ergonomic, and personalized vibrotactile scheme; and (4) a headset that contains both binaural bone conduction speakers and a noise-cancelling microphone for oral communication [10]. [11]. [13]. [15]. The system leverages stereo cameras as its primary sensory input and employs advanced computer vision algorithms on Nvidia Jetson processing boards. The goal of the embedded system is to enable continuous mapping, localization, and surveillance within a dynamically changing environment [10]–[13]. [15]. [17].

The key limitation in the current system is that the video processing is performed entirely locally and video-based, which is computationally and power intensive and limits the performance of visual analytics. Indeed, the wearable runs off a laptop-battery with approximately 66 Wh at 0.5 kg yielding 2–3 hours of function if continuously running the vision processing. To stay within these power limits, the wearable uses the Jetson Xavier NX to perform object detection. With a standard YOLO model the system is able to process only WVGA resolution video at a rate of 10-13 frames per second (FPS) — See Table 3; results are often even poorer in mobile phone applications employing similar approaches [55]. As we discuss below, the low resolution results in poor detection accuracy and limited range.

Another limitation of the current system is that it deploys only a single stereo camera providing a field of view of approximately 90 degrees horizontally and 60 degrees vertically. At about 3 meters of distance or range from the end user, a 90-degree field of view is very restrictive, leaving potentially pertinent spatial obstacles out of the perceptive capabilities of the system, ones that may be encountered with even slight orientation shifts in forward paths. While this may be circumvented by simply using ultra-wide angle cameras, the geometric distortion in such cameras can degrade performance of visual analytics [55]. In order to address these shortcomings, we propose to embed multiple cameras in the VIS$^4$ION backpack to provide omnidirectional coverage.

C. WIRELESS OFFLOADING SYSTEM STUDIED

Wireless offloading of the vision processing to a mobile edge cloud offers two key potential benefits: (1) greater processing capability in the edge cloud can enable analysis of multiple high-resolution camera streams for fast and more accurate object detection, over a wider field of view and greater range (distance); and (2) reducing processing on the wearable can prolong the battery life and/or reduce the battery weight.

To assess these potential gains, we consider two augmentations to the backpack, depicted in blue in Fig. 2. First, we consider a version of the wearable with four stereo cameras, for example, to cover four sectors of 90 degrees each. Second, to process the multiple cameras at higher resolution, we consider adding cellular connectivity to an edge cloud, with higher process capabilities. The system will adapt the number of camera streams to be uploaded to the edge server and their target bit rates based on the estimated uplink network capacity. Furthermore, the compression configuration (e.g., frame rate and spatial resolution) of each video stream will also be adjusted based on target rates (adaptive). The mobile edge cloud will analyze the videos from multiple cameras using a deep-learning network for object detection (and other tasks such as environmental mapping) and send the results back to the wearable over the downlink; see also Fig. 1.

When the wireless connection is temporarily down (e.g., due to blockage), the local processor can analyze one video stream at lower resolution, while storing the captured high-resolution video within its on-board memory. The high-resolution video can then be opportunistically uploaded to the edge server when the user re-establishes a high-bandwidth wireless connection, to enable the mapping of various environments, perform post-hoc behavioral analysis, etc.

D. DELAY REQUIREMENTS

For real-time pedestrian navigation, there is no generally agreed upon requirement for the tolerable total delay between the time an object appears in the environment of a pedestrian and the time it should be detected and reported to the pedestrian. In our previous work, we suggested 100 ms [56]. This benchmark is predicated on a physiologic marker, the high-end of the duration range for a large-amplitude saccade (fast eye movement); such an eye movement would be used by a normal-sighted pedestrian to identify a potential hazard. This stringent delay requirement enables
III. IMPACT OF VIDEO COMPRESSION AND SPATIAL RESOLUTION ON OBJECT DETECTION ACCURACY AND RANGE

A. OVERVIEW

Due to the high bit rate of raw high-resolution video, compression is needed to stream image frames to the edge servers via a throughput-constrained link. In this section, we analyze how video compression (including reducing video resolution) impacts object detection performance and inference time. The analysis will provide the bit rate and delay requirements for the wireless uploading described in Section IV.

Although the augmented VIS4ION system will use multiple stereo cameras to enable wider field of view and distance estimation, the analysis in this section focuses on a single monocular video. How to combine detection results from multiple views or make use of depth information in object detection is an interesting subject of future work. Thus, in the wireless evaluation in the next section, we will consider only uploading of multiple monocular streams.

Given a target rate for a camera, the video can be compressed at different spatial resolutions (frame size in terms of pixels) and temporal resolutions (frame rate), as illustrated in Fig 3. With the chosen spatiotemporal resolution, the bit rate is controlled by the quantization stepsize, which controls the amplitude resolution and affects the pixel quality. While there has been significant work in relating spatial, temporal, and amplitude resolution (STAR) to perceptual video quality [57]–[59], the effect of STAR on object detection accuracy is less understood. As mentioned in the Introduction, most prior works have only studied relatively low-resolution images.

Here, we conduct a study to systematically evaluate the impact of spatial and amplitude resolution on the object detection accuracy using a popular object detection deep learning model (YOLO [60]). We leave out the consideration of the temporal resolution at this time because the YOLO model works on video frames independently. This study enables us to determine the optimal spatial resolution for a given bit rate, and the achievable detection accuracy under the optimal resolution at this rate. We will further characterize the effect of the object distance (from the camera) on the detection accuracy under different spatial and amplitude resolutions, to provide recommendations/guidelines on the necessary spatial resolution and bit rate to meet the desired detection range for wearables that support pedestrian navigation application. Finally, we characterize the computational cost of YOLO (including inference time) at different spatial resolutions, which provides guidance on the tolerance for the roundtrip delay with wireless offloading.

B. CREATION OF THE NYU-NYC STREETSCENE DATA SET

Currently, there are no public datasets containing high-resolution videos captured from the perspective of a typical pedestrian. A significant effort in this work is the creation of a new, manually-annotated ‘StreetScene’ video dataset for this purpose.

Video collection

To test the performance of the YOLO model for detecting objects of interest for pedestrian navigation, we recorded a set of videos while wearing the current VIS4ION backpack which has a single stereo camera on the front shoulder strap. The camera model is the ZED camera from StereoLabs [61] – a lightweight, powerful recording device, ideal for wearables requiring spatial intelligence. A total of 9 videos were captured with the ZED at the 2.2K spatial resolution and 15 Hz temporal resolution, with a total video length of 43 minutes.

Object annotation

We manually annotated the bounding boxes for 15 objects of interest, listed in Table 1, along with the number of occurrence for each object. We annotated every 30th frame of each video (only the left view). This annotated dataset is publicly available at [27]. Because YOLO was trained where the ‘traffic light’ includes both ‘vehicle traffic light’ and ‘pedestrian signal’, we grouped these two separately annotated objects into the same object type when applying the YOLO model. Furthermore, because the detection performance of YOLO on ‘bench’, ‘stop sign’ and ‘dining table’ is very poor, we only report the detection performance for detecting the remaining 11 objects.
ON OBJECT DETECTION ACCURACY

Table 1: Object types annotated in the NYU-NYC StreetScene Dataset.

| Object       | person | car | traffic light | vehicle | traffic light | pedestrian | signal | potted plant | bicycle | truck | chair | fire hydrant | bus | umbrella | motor cycle | Not reported |
|--------------|--------|-----|---------------|---------|---------------|------------|--------|--------------|---------|-------|-------|--------------|-----|----------|-------------|--------------|
| Occurrences | 9783   | 5442| 1977          | 681     | 1122          | 704        | 459    | 480          | 370    | 349   | 162   | 247          | 277 | 180      |

* Defined relative to the 2.2K resolution (same reduction in both width and height).

Table 2: Spatial resolutions considered

| Resolution | Width | Height | Reduction factor* |
|------------|-------|--------|-------------------|
| 2.2K       | 2208  | 1232   | 1x                |
| 1080P      | 1920  | 1080   | 0.87x             |
| 720P       | 1280  | 720    | 0.58x             |
| WVGA       | 672   | 378    | 0.30x             |

Video compression

We compressed all the videos (left view only) in the StreetScene dataset using the FFmpeg software with the x265 codec [62], [63], which follows the latest international video coding standard H.265/HEVC [64]. We kept the same temporal resolution, and compressed the video either at the original 2.2K spatial resolution or reduced spatial resolutions (See Table 2) under different quantization parameters (QPs). Default down-sampling filters (‘bicubic’) in FFmpeg were used for the spatial down-sampling. Considering the low-delay requirement of the navigation application, we used a Group of Picture (GOP) length of 60 frames, without B-frames, i.e., each GOP starts with one I-frame, followed by 59 P-frames.

Distance estimation for standing people

To examine how distance affects detection accuracy in the StreetScene dataset, which does not have accurate distance measurements [1], we developed a method to estimate the distance of standing people in our ‘StreetScene’ dataset. We focus on distance estimation for this object type since this is a relatively small object whose detection can be greatly affected by the object distance. Given that the variation of the physical size of standing people is relatively small, the size of the box bounding a standing person is mainly determined by the distance of the person from the camera. Based on this observation, we trained a distance estimation model (containing a few fully connected layers) based on the bounding box width and height using the KITTI dataset [65], which has annotated bounding boxes and distances for standing people. To account for the difference in the camera used for the KITTI data and our ZED camera, we used our ZED camera to capture a set of videos with standing people at multiple distances against a variety of backgrounds, with the distance captured using the positional tracking system of the ZED camera. Using this dataset, we were able to learn the mapping from the distance estimated by the model trained on the KITTI data to the distance for video captured by the ZED camera. To apply this model on the people detected in the StreetScene dataset using the YOLO model, which includes both standing and sitting people in the same object category, we looked at the distributions of the height over width ratio among the standing and sitting people in the KITTI data, and found that using a ratio threshold of 2.0 can fairly reliably separate standing people from sitting people. Therefore, we used this ratio threshold to detect standing people in the StreetScene dataset. By applying the distance estimation model to the detected bounding boxes for the standing people followed by the camera mapping, we generated the distance measurements of standing people in the StreetScene dataset.

C. EFFECT OF SPATIAL AND AMPLITUDE RESOLUTION ON OBJECT DETECTION ACCURACY

We first examine the impact of spatial and amplitude resolutions (with corresponding bit rates) on the object detection accuracy on the StreetScene dataset, in which objects appear at varying distances. Results show that the optimal spatial resolution varies with the target bit rate (which is constrained by the network throughput). We applied a pretrained YOLO 5s model [66] on the decompressed videos in the StreetScene dataset to detect the 14 objects of interest. Figure 3 shows the weighted mean average precision (wmAP) over 11 objects (see Table 1) vs. bit rate. The weight for an object type is proportional to its occurrence frequency in the dataset. The figure reveals that there is an optimal spatial resolution at each bit rate that will maximize the wmAP. Specifically, 720P is best for 0.35-6.0 Mbps, 1080P for 6.0-26.2 Mbps, 2.2K for higher bit rates. However, 2.2K provides only marginal improvement over 1080P above 26.2 Mbps. We note that this could be because the YOLO model was trained mainly with low-resolution images. Although WVGA is best at very low rate (below 0.35 Mbps), the achievable AP is too low to be usable.

At 26.2 Mbps and using 1080P resolution, the weighted mean AP is about 54%, which is still far from perfect. However, this relatively low detection accuracy is due to limitations of the YOLO 5s model, which was trained using uncompressed low-resolution images. Better detection models (e.g., models specifically trained for street scenes and/or models that are separately optimized for different resolutions) will likely further improve the detection accuracies.

Note that although the videos in the StreetScene dataset are captured and compressed at 15 Hz, we report the equivalent bit rates for videos at 30 Hz, which is necessary to meet the real-time navigation requirement as further detailed in Sec. II-D. This is accomplished by scaling the actual bit rates corresponding to different spatial resolutions and QPs with different scaling factors determined by a separate experiment where we compressed videos at 30 Hz and 15 Hz separately at multiple resolutions and QPs for several sample videos captured at 30 Hz.
FIGURE 4: Mean detection accuracy (weighted mean AP) for 11 object types vs. bit rate for different spatial resolutions. Different points in the same curve correspond to different QPs.

FIGURE 5: Detection accuracy (AP) for person vs. bit rate for different spatial resolutions. Different points in the same curve correspond to different QPs.

It is tenable that the trend of the detection accuracy vs. rate vs. spatial resolution would be preserved for future, more powerful models.

Figure 5 presents the detection result for the person category. We see a similar trend as in Fig. 4, although the specific rate points where higher resolutions take over the lower resolutions, which shows that the YOLO model is more effective in detecting people than other object categories. This is consistent with the performance reported in [67], likely because there are significantly more instances of person than other objects in the training set.

Sample frames from videos compressed to around 10 Mbps using different settings are shown in Fig. 6(a-b). From the outset, it is not clear which decompressed image will lead to improved object detection. However, from the detection results shown in Fig. 6(c-d), the YOLO model did better for image (b), which was represented with a high spatial resolution but low amplitude resolution.

D. EFFECT OF SPATIAL RESOLUTION ON DETECTION ACCURACY AT DIFFERENT DISTANCES

The performance measures reported so far are aggregated results for objects appearing at varying distances from the camera. Generally, detecting a faraway object is harder than a nearby object. On the other hand, being able to detect an object while it is still far away provides more time for navigation planning. Therefore it is important to understand how the detection accuracy degrades as the distance increases and what is the maximum distance when an object can be detected reliably. How does the distance affects the detection also depends on the physical size of the objects. We show such results for detection of standing pedestrians as a case study, wherein we use the algorithm described in Sec. III-B to detect standing pedestrians and furthermore estimate the distances of the detected person(s) from bounding-box size(s).

We quantized the distances to several bins and determined the AP within each bin. Figure 7 indicates how the detection accuracy drops as the object distance increases, under different spatial resolutions. When the distance is very close, YOLO performs very well even at the WVGA resolution, but the accuracy drops quickly as the distances increases at this low resolution. As expected, higher spatial resolutions enjoy a slower decay rate. The 2.2K resolution leads to significantly better detection than 1080P only when the distance is greater than 21 m. Note that the 2.2K resolution did worse than other lower resolutions in the short distance range in this study. This is likely because the YOLO 5s model was trained on low-resolution videos (close to WVGA). People within a short distance occupy a very large area in the 2.2K image, requiring bounding box sizes that rarely occur in the training data.

Figure 8 shows the detection ranges for different spatial resolutions. Here, we see clearly that going from WVGA to 1080P, we are able to extend the detection range from about 6 m to 12 m. This study shows that we should use at least 1080P video to be able to reliably detect people at a distance important for navigation planning.

E. COMPUTATIONAL COMPLEXITY VS. SPATIAL RESOLUTION

To quantify the benefit of edge computing and the conditions under which local computing is preferable, Table 3 summarizes the computation complexity (measured by the FLOP count), the inference time per video frame on the embedded processor in our VISION backpack (Jetson Xavier NX running at 15 Watts, using a GPU at 1.1 GHz), the inference time per video frame using an edge server equipped with an RTX 8000 GPU, and the detection accuracy by the YOLO 5s model, for videos at different spatial resolutions. The detection accuracy and bit rate for each resolution where the accuracy saturates according to Fig. 4 are also reported.

The accuracy would be similar even when the uncompressed video is processed locally.
IV. WIRELESS EVALUATION

Having analyzed the bit rate and latency requirements for the video processing, we now simulate the wireless network to determine what percentage of time these requirements can be met.

A. USER ROUTE AND RAY TRACING

We simulate a hypothetical end user commuting through the streets of Manhattan – a challenging environment from a wireless perspective, due to the tall-building blockage. In particular, we identified a walking route starting at Lighthouse Guild (a healthcare and research center that assists persons with BVI) and finishing at the NYU Tisch Hospital in Kips Bay, as depicted in Fig. 9. The environment of this...
TABLE 3: Impact of spatial resolution on the detection model complexity, running time on local processor and edge server, respectively, and detection accuracy. The Jetson Xavier NX is used as the local processor, while the server uses an RTX 8000 GPU. The table also reports the achievable detection accuracy for each resolution at the bit rate where the detection accuracy plateaus.

| Resolution | GFLOP | Local Inference Time (ms) | Server Inference Time (ms) | wmAP (11 obj) | AP (person) | Bitrate (Mbps) |
|------------|-------|---------------------------|---------------------------|---------------|------------|----------------|
| 2.2K       | 57.22 | 232.02                    | 23.4                      | 54.62         | 67.27      | 90.93          |
| 1080P      | 83.38 | 178.25                    | 18.7                      | 54.08         | 56.11      | 26.10          |
| 720P       | 79.36 | 95.69                     | 10.3                      | 56.27         | 60.31      | 18.77          |
| WVGA       | 5.36  | 75.02                     | 5.1                       | 36.57         | 44.32      | 9.29           |

FIGURE 9: Main walking route from which we extracted four specific sites represented by the red rectangles. Top to bottom, the sites are Lighthouse Guild, Midtown, Herald Square, and NYU Langone route is similar to where the NYU-NYC StreetScene dataset described in Section III was collected. Along this route, we selected four sites (red rectangles of Fig. 9) to perform a realistic, full stack 5G simulation. This simulation involved two main steps: 1) generation of ray-tracing data of the wireless environment from the realistic 3D layout of the city at these sites, and 2) end-to-end simulation using Network Simulator 3 (ns-3) and the ray-tracing information obtained in the previous step. Here, we describe the generation of ray-tracing data with the figures for the top site in Fig. 9 as an example.

In wireless communications, ray-tracing involves the calculation of the paths that electromagnetic waves, represented as rays, follow while propagating in a known 3D environment according to a set of transmitting and receiving positions. For each combination of transmitting and receiving locations, the output of a ray-tracing simulation consists of a set of propagation information for each ray: path loss, propagation time, phase offset, Angles of Departure (AoD), and Angles of Arrival (AoA). We used Remcom’s Wireless InSite [29] software to generate ray-tracing data in our work. This software has been successfully used in a number of other studies [68]–[70]. Accurate ray-tracing of a 3D scenario at mmWave and sub-6-GHz frequencies requires precise information of buildings materials, since each material affects the propagation of the electromagnetic wave differently at each frequency range. In addition, details of the vegetation in a certain area are essential for accurate ray-tracing at mmWave frequencies. To capture all this information, we imported in Remcom the 3D layout of the City at each of the 4 sites, as provided by GeoPipe [30]. Fig. 10 depicts the 3D layout with building materials and vegetation at the top site of Fig. 9. This data from GeoPipe provides one of the most accurate models for mmWave ray tracing. In particular, the models contain small building features which are known to influence mmWave propagation significantly [71]. For each material (i.e., concrete, wood, glass, etc.) and frequency range, Remcom uses different parameters for the reflection and diffraction coefficients, as well as different propagation properties. We performed the simulation at two frequencies: 1.9 GHz for a typical sub-6 GHz 4G LTE carrier, and 28 GHz for a typical 5G mmWave carrier.

Moreover, we considered a real placement of the base stations in each of the four sites as provided by [72]. This database contains information regarding the actual foreseen placement of 5G mmWave Base Stations (BSs) in New York City. We assumed that the LTE 4G and 5G mmWave base stations are co-located, meaning that, at each cell site, base station equipment is available for both frequencies. This co-location is common since, once the operator has secured a site, it generally utilizes it maximally. Note that in 3GPP terminology, a 4G base station is called a eNB (evolved Node B), and a 5G base station is called a gNB (next generation Node B).

For the ray tracing, each base station site represents a transmitting location, whereas the receiver positions are placed one meter apart along the route taken by the hypothetical user at each specific site (more details in IV-D). The ray tracing then provides the channel from each BS to each position along the route at both frequencies. Note that due to symmetry, the large scale channel parameters are identical in the uplink and downlink. Hence, we can use the estimated channel in both directions.

B. NETWORK SIMULATION

The second step in the evaluation of the wireless offloading involves the end-to-end (full stack) network simulation using...
ns-3. Ns-3 is an open-source, discrete event network simulator which affords end-to-end simulations (i.e., simulations that involve the entire network protocol stack from the physical layer up to the application layer) with support for user mobility and traffic modeling, among other features. To compare the performance of mmWave connectivity with a standard sub-6 GHz system, we run two simulations:

- An LTE system at 1.9 GHz with 40 MHz uplink and 40 MHz downlink total bandwidth; and
- A 5G mmWave system at 28 GHz with 400 MHz total bandwidth that is Time-Division Duplexed (TDD) for the uplink and downlink.

For both the 4G and 5G systems, the deployment would likely be with multiple carriers, as is common today. For example, the LTE system could be two standard carriers of 20 MHz each and the 5G mmWave system with four carriers of 100 GHz each. The parameter values for both systems are shown in Table 4 and are representative of typical 4G and 5G simulation studies, see, e.g., [7], [73], [74]. To account for loading, we assume an individual UE obtains a fraction 0.25 of the total bandwidth, which would represent a moderate loading level relative to standard evaluation methodologies [74]. Hence, we simulate the 4G user as operating in a system with 10 + 10 MHz bandwidth and the 5G user as operating in a system with 100 MHz total bandwidth.

In both the 4G and 5G cases, we model the wearable as a User Equipment (UE) traversing the path described above. The channels from the ray tracing in Section IV-A are imported into the ns-3 simulator.

Since the ray tracing captures only the buildings and foliage, it does not capture blockage from objects such as humans and vehicles in the environment. As discussed in the Introduction, modeling blockage is critical to accurately assess mmWave coverage [23], [24]. To model this additional blockage, we employ the Blockage Model A in [74] which is integrated into the ns-3 simulator [31]. This model adopts a stochastic approach for capturing human and vehicular blocking. In particular, multiple 2D angular blocking regions, in terms of azimuth and elevation angular spreads, are generated around the UE. One blocking region, denoted as self-blocking region, captures the effect of human body blocking, whereas \( K_{NSB} \) non-self-blocking regions with random sizes are used to model other sources of blockage (\( K_{NSB} \) can be changed to increase/decrease the density of blockers). Once the blocking regions are computed, each cluster (or ray, in the case of our ray-tracing data) is attenuated accordingly, based on the angular spreads and position of each blocking component. The parameter \( T_{NSB} \) denotes the time interval at which new blockers are randomly generated.

Given the blocked channels, the ns-3 simulator then models the full stack communication. At the Physical (PHY) and Medium Access Control (MAC) layers, the modeling includes the beam tracking, Channel Quality Information (CQI) reports, rate prediction, Hybrid Automatic Repeat reQuest (HARQ), and scheduling. At the higher layers, the simulator models all the Radio Link Control (RLC) segmentation and buffering, Radio Resource Control (RRC) signaling, and handovers.

An important parameter in the network configuration is the location of the edge server. In cellular systems, data in the uplink traverses a path: UE (wearable) \( \rightarrow \) base station (4G eNB or 5G gNB) \( \rightarrow \) core network \( \rightarrow \) server in the public Internet. The downlink follows the reverse path. In conventional deployments, operators have relatively few gateway points from the core network to servers in the public Internet. The data may thus need to traverse a long path in the core network to the closest gateway resulting in high delay from the base station to the server — see, for example, measurements in commercial networks in [75]. Mobile edge computing reduces the core network delay by placing the edge servers much closer to the base station.

![3D map of the top site, Lighthouse Guild, in the route in Fig. 9 with material information for each building and vegetation provided courtesy of GeoPipe](Image 42x552 to 272x722)
As mentioned above, we performed end-to-end wireless analysis, we instead model the uplink video data as a camera video encoding application in the ns-3 simulator. Ideally, in the uplink, we would model the adaptive multi-camera video encoding application in the ns-3 simulator. This analysis would then be specific to the video adaption algorithm used. To provide a more general and simpler analysis, we instead model the uplink video data as a single TCP stream with full buffer up to a maximum data rate of 120 Mbps. This maximum data rate is sufficient to support four cameras at 30 Mbps each. Since TCP has congestion control, it will automatically adjust the sender rate to the available uplink link capacity. As a simplification, we assume that the video encoding can be adapted to be exactly the same as the TCP rate. Hence, the full buffer TCP rate at any time can be regarded as an approximation of the actual video rate.

For the downlink, data from the edge server to the wearable is used to carry the object detection results. We model this downlink traffic as a constant bit rate application at 30 packets per seconds, corresponding to the expected video frame rate. We assume that the total rate is 1 Mbps, which is ample to specify a large number of detected objects, including their bounding boxes and probabilities belonging to different object classes.

D. SIMULATION RESULTS
As mentioned above, we performed end-to-end wireless simulation of an end user walking through four sites in Manhattan, near Lighthouse Guild, Midtown, Herald Square, and NYU Langone as shown in Fig. 9. As one example, Fig. 11 depicts the route of the user (UE) for the Herald Square site. The UE starts walking from the north-west corner in the figure and moves south-east across 34th Street (yellow line in the picture). In this site, two base stations are present, both with 4G-LTE and 5G. Figure 12(a) shows the SINR over time for this scenario for the two mmWave base stations (gNB 1 and gNB 2), as well as the maximum SINR for the two LTE cells. We see that, in this case, the LTE SINR is continuously high (mostly >40 dB) due to the favorable propagation of the lower frequency (1.9 GHz) carrier. In contrast, there is a period from approximately 80 to 170 seconds where the SINR from both mmWave cells is low. This time period corresponds to a segment of the route where the UE is in Non Line of Sight (NLOS) to both mmWave cells.

The resulting TCP end-to-end throughput is shown in Fig. 12(b). We see that the LTE rate is more continuously available. However, the maximum rate is limited to ~36 Mbps corresponding to the maximum modulation and coding scheme (MCS) with a 10 MHz bandwidth. In contrast, the mmWave system can obtain the full 120 Mbps rate, but the rate falls below the LTE rate during some parts of the NLOS segment.

The simulation results for each of the four scenarios have been aggregated and depicted in Fig. 13. Fig. 13(a) plots what we will call the delay-constrained throughput, which is calculated as follows. We divide the time into intervals of $T$ seconds, where $1/T = 30$ Hz corresponds to the expected video frame rate ($T$ is the frame interval). For each TCP packet transmitted in the interval, we measure its uplink delay from the UE to the edge server application. We also measure the downlink delay for each feedback packet transmitted in that interval as well. Note that these delays contain all the air-link and core network delays. For a given delay constraint, $D_{\text{max}}$, we define the delay-constrained throughput as $b/T$, where $b$ is the number of uplink bits transmitted in the interval $T$ for which the uplink + downlink delay $\leq D_{\text{max}}$.

The delay-constrained rate is computed separately for the LTE and mmWave systems. Since mmWave systems are always deployed with a sub-6 GHz fallback carrier, we also estimate the rate of mmWave+LTE system as the maximum of these two rates under the same delay constraint. Fig. 13(a) plots the delay constrained rates for the LTE and mmWave+LTE systems under delay constraints of $D_{\text{max}} = 30, 40$ and $50$ ms.

We see in Fig. 13(a) that the LTE system achieves a peak uplink rate of approximately 36 Mbps, and attains over 20 Mbps more than 90% of the time. However, this rate is only achievable with a delay constraint of $D_{\text{max}} = 50$ ms. At a tighter delay constraint of $D_{\text{max}} = 40$ ms, the delay-constrained rate of approximately 36 Mbps is supported less than 80% of the time; at $D_{\text{max}} = 30$ ms, there are virtually no LTE data within this delay constraint. In contrast, the
mmWave 5G system coupled with LTE can obtain the peak rate of 120 Mbps at least 40% of the time, even under a delay constraint of $D_{\text{max}} = 30$ ms. However, for approximately 25% of the time, the delay-constrained rate of the mmWave+LTE system is similar to that of the LTE system owing to the fact that the mmWave only coverage is not always available and falls back to the LTE carrier.

To understand the delay differences, Fig. 13(b) plots the CDF of the delays of packets without any delay constraint. We see that for the mmWave system, the minimum delay is $\approx 15$ ms which includes two times the core network delay of $D_{\text{core}} = 5$ ms along with an addition 5 ms for the transmission of the uplink and downlink data. The LTE packet delays are generally higher. Although the core network delays is assumed to be the same, the LTE frame structure as well as the lower throughput results in a higher air-link delay.

V. PERFORMANCE EVALUATION

A. OVERVIEW

We now combine the video processing requirements in Section III with the wireless simulation results in Section IV to assess the potential benefits of wireless offloading. We consider three scenarios for edge connectivity: (1) Local only, where all the processing is performed on the wearable; (2) LTE only, where edge processing can be accessed by the LTE link; and (3) mmWave+LTE where edge processing can be accessed by LTE or mmWave, whichever has the highest rate. For each such scenario, we consider different possible video options in terms of the number of cameras, spatial resolution and bit rate. We can then use the wireless analysis in Section IV to assess the percent of
### Table 5: Comparison summary of example configurations in different connectivity scenarios.

| Item                                      | Local only | LTE only | mmWave+LTE | Adaptive | Remarks                                                                 |
|-------------------------------------------|------------|----------|------------|----------|--------------------------------------------------------------------------|
| **Video configuration and object detection performance** |            |          |            |          |                                                                          |
| Number of monocular cameras               | 1          | 1        | 1          | 1        | 4                                                                       | Variable (1-4)*  |
| Camera resolution                         | WVGA       | 720P     | 1080P      | 1080P    | mostly 1080P                                                            | 1 camera if throughput ≤ 26 Mbps, more cameras if throughput >26 Mbps |
| wmAP (%) for multiple objects             | 36.6       | 50.3     | 54.0       | 54.0     | 54.0 avg†                                                               | See Table 3       |
| AP (%) for person                         | 44.3       | 60.3     | 66.1       | 66.1     | 65.8 avg†                                                               | See Table 3       |
| Detection range for person (meter)        | 6          | 9        | 12         | 12       | mostly 12                                                               | See Fig. 8        |
| **Bandwidth, Delay, and Availability**    |            |          |            |          |                                                                          |
| Total uplink data rate (Mbps)             | 0          | 0        | 26         | 26       | 104                                                                      | variable          |
| Video frame delay (ms)                    | 33         | 33       | 33         | 33       | 33                                                                       | We assume video is captured at 30 Hz. |
| Video encoding delay (ms)                 | 0          | 0        | 17         | 17       | 17                                                                       | We assume that video encoding takes at most 1/60 (s) per frame, because today’s smart phones can capture video at 60 Hz. Local processing does not need to compress video. |
| Inference time (ms)                       | 75         | 96       | 19         | 19       | 19                                                                       | See Table 3       |
| Median round-trip time (RTT) (ms)         | -          | -        | 37         | 15       | 15                                                                       | See Fig. 13(b).   |
| Median total delay (ms)                   | 108        | 129      | 106        | 84       | 84                                                                       | Frame delay+encoding delay+inference time+RTT |
| Availability with total delay ≤ 100 ms    | 0%         | 0%       | 0%         | 75%      | 65%                                                                      | Computed from CDF of delay constrained throughput. See text.    |
| Availability with total delay ≤ 150 ms    | 100%       | 100%     | 95%        | 97%      | 67%                                                                      |                                                                  |

†The average wmAP and AP presented are for a total delay of ≤ 100 ms. These numbers are 53.9% and 66.0%, respectively for a total delay ≤ 150 ms.

There are many other feasible configurations, including:

1. Processing 1080P video locally for increased detection accuracy, at a total delay of 211 ms.
2. Sending both views of each stereo camera or one view plus depth map, with increased up link rate, and reduced availability.

### B. VIDEO CONFIGURATIONS

In Table 5 for edge computing with mmWave or LTE connectivity, we have considered the case where the video from each monocular camera would be delivered at 1080P spatial resolution, 30 Hz temporal resolution, at 26 Mbps. Based on the video analysis in Section III, this configuration provides a high object detection accuracy and good detection range – see the wmAP and detection range rows in Table 5. Going beyond 1080P resolution and 26 Mbps brings only very slight gains, and yet processing 2.2K video will consume substantially more computation time.

For the local processing scenario, we have considered only WVGA and 720P. With local processing, the inference time for higher spatial resolution 1080P (see Table 5) would substantially exceed the delay budget. As shown in Table 5, this lower resolution results in both a lower object detection...
accuracy and reduced object detection range.

C. DELAY ANALYSIS
The delay computations in Table 5 consider four components:

- **Video frame delay** which is the interval of one video frame (i.e., the inverse of the frame rate). For edge computing, the video frame interval needs to be considered because an object may appear any time between two adjacent frames.
- **Video encoding delay** which is the time to encode the video for edge computing.
- **Round-trip time (RTT)** which is the total time to transmit the packets from the wearable to the edge server and back.
- **Inference time** which is the time for object detection network (either local or edge) to compute the detection results.

In our analysis, the video frame delay, encoding delay, and inference time are fixed. The only variable component is the RTT. Table 5 shows the median RTT and the corresponding median total time. We see that mmWave+LTE offer dramatically lower median RTT of 15 ms relative to the median RTT in LTE only of 37 ms. Recall that the RTT includes 2D_{core} = 10 ms of assumed delay from the base station (gNB or eNB) through the core network to the edge server and back. Note that with direct local processing, only the frame delay and inference time are required since there is no video encoding or communication.

As suggested in [21], to avoid detection mismatch, we will assume that the local processor runs a simple object tracking algorithm that predicts the locations of objects detected for the last frame for which the edge server detection results were fed back. For example, if the total delay for edge processing is twice of the frame interval, at the time when frame \( t \) is captured, detection results for frame \( t − 2 \) will be used as the reference, the motion between frame \( t \) and frame \( t − 2 \) will be used to predict the locations of these detected objects in frame \( t \). In the mean time, frame \( t \) will be delivered to the edge server with its processing results to be fed back at time when frame \( t + 2 \) will be captured. The motion vectors between frames generated for video compression can be leveraged for the local object tracking. Such an approach should be able to track small movements of previously detected objects within a few frames, while also report any newly appearing objects within the 100 ms delay.

D. SUPPORTABLE VIDEO STREAMS UNDER DIFFERENT DELAY CONSTRAINTS
From the wireless simulation results in Figure 13, we derive the probability of supporting one or more camera streams under different round-trip delay constraints. These are plotted as heat maps in Fig. 14(a). Fig. 14(a) shows that, with multi-connectivity using both mmWave and LTE links, we can support one video stream over 75% of the time, and support two streams over 73% of the time, under the roundtrip delay constraint of 30 ms. All four cameras can be supported 65% of the time. Furthermore, we can support one and four cameras with high availability (91% and 67%, respectively) if the delay constraint is relaxed to 40 ms. On the other hand, with the LTE link only (Figures 14(b)) and a round-trip delay of 40 ms, the availability for supporting one and four cameras drops to 79% and 0%, respectively, since the peak LTE rate is approximately 36 Mbps. The LTE link can sustain one camera with high probability (93%) only if the delay constraint is relaxed to 50 ms.

For any configuration, we can also compute the probability that the total delay will meet a certain delay target. For example, from Section II-D, the estimated total delay requirement is 100 ms. This is not met by local processing. When using offloading, the video frame delay, encoding delay, and inference take a total of 33 + 17 + 19 = 69 ms, so there would be \( D_{max} = 100 − 69 = 31 \) ms for the RTT. Similarly, if we relax the total delay requirement to 150 ms, the communication delay constraint would be \( D_{max} = 150 − 69 = 81 \) ms. The availability numbers listed in the final two rows of Table 5 are the percentage of time the delay-constrained throughput meets the minimum rate requirement at the RTT of 31 and 81 ms, respectively. The availability percentage for these two delay constraints is not shown in Fig. 14 but we have extracted the numbers from the wireless simulation results in a similar manner. Note that the availability for RTT of 31 ms is the practically same as for 30 ms. Using Fig. 14 one can also find the corresponding availability for total delays of 110 ms (RTT of 40 ms) or 120 ms (RTT of 50 ms).

In practice, given the limited total throughput, it may be better to only upload the front facing camera stream at a high rate (for highest detection accuracy), and use lower rate for other cameras (side and back facing) to enhance situational awareness. As an example, Figure. 14(c) shows the probability of supporting one camera at the full rate of 26 Mbps and additional cameras at 10 Mbps each. In this case, with mmWave+LTE connectivity and a total delay 100 ms, we could increase the probability to support 4 cameras from 65% to 72%.

E. ADAPTIVE OFFLOADING
Given the variability in availability, particularly in mmWave, it is natural to consider a strategy that adaptively selects the rate, number of cameras, and whether to use local or remote processing based on the available uplink bandwidth. For example, as a simple strategy, when the available throughput is ≥26 Mbps, we can transmit one or more cameras. When the throughput is lower than 26 Mbps, we can still offload the video for edge computing, but at lower rates, enabling adaptive switching between camera number and video quality based on bandwidth constraint and functional need (refer to Fig. 14 as an example). From Fig. 4 when the throughput is between 6 Mbps and 26 Mbps, the system...
should deliver the video at 1080P but at a lower rate, leading to proportionally lower detection accuracy. When the rate is between 1 Mbps and 6 Mbps, the system should upload the video at 720P resolution. In the “adaptive” column of Table 5, the detection accuracy is derived by assuming an average detection accuracy of 51.5 and 63.2 for multi-object and person, respectively, when the bit rate is between 6 and 26 Mbps (which occurs 2% of the time with a delay constraint of 30 ms, from Fig. 13(a)); and an accuracy of 41.1 and 49.7, respectively, when the bit rate is between 1 Mbps and 6 Mbps (which occurs 1% of the time). The overall availability for adaptive offloading under total 100 ms delay is the probability that the throughput is ≥ 1 Mbps at a RTT constraint of 30 ms.

Under the relaxed total delay constraint of 150 ms, when the throughput is between 10 and 26 Mbps (with probability of 0.8%), the wearable could locally process the uncompressed video at 720P video resolution for better detection performance (see Fig. 4). Therefore the availability is 100%. The average accuracy would be 53.92% and 66.00%, for multi-object and person, respectively.

VI. CONCLUSIONS AND FUTURE WORK
Mobile edge computing coupled with the high data rate capabilities of mmWave holds significant promise for accessing powerful video analytics by wearable devices. We have assessed the feasibility of such capabilities for an advanced smart wearable with multiple high-resolution cameras where the wireless and video requirements are particularly demanding. Several new elements were required in the analysis including developing a large labeled video data set, evaluation of object detection algorithms at variable resolutions and bit rates, and detailed and high accuracy wireless simulations with ray tracing. Overall, the wireless...
simulations provides a high level of realism and can identify the key limitations in high data rate edge computing. These tools can be applied in other applications and may prove valuable as video processing and spatial intelligence becomes more widely-used in mobile scenarios.

For the VISION application, our simulation results suggest that at bandwidths and loading similar to current deployments, systems in traditional sub-6 GHz bands combined with low delay mobile edge computing can provide gains by offloading camera data a large fraction of time, improving the accuracy and detection range. However, meeting the end-to-end delay requirements of 100 ms is challenging. The mmWave bands can reduce the delay to meet these requirements and provide additional capabilities including multiple cameras at high resolution. However, due to blockage and the limited range of mmWave signals, the peak performance is not uniformly available at typical cell-site densities. Thus, fall back to lower frequency carriers and local processing combined with adaptation in the video resolution and number of camera streams will be required.

In the current work, we have abstracted this adaptation by assuming that the number of cameras and their bit rate can be adjusted to the available throughput. An obvious line of future work is to actually simulate a particular adaptive video application over wireless links and assess its performance. Additionally, we have simply relied on the pre-trained YOLO network that was trained on uncompressed low-resolution images. A second line of work is to train multiple detection networks for different resolutions and bit rates or a single network that can perform well across resolutions and bit rates. More ambitiously, one can also consider new compression schemes that are trained end-to-end with object detection accuracy as the goal, as opposed to the standard compression algorithms, which are optimized for image reconstruction.

Acknowledgements

This work was supported in part by the NSF grant 1952180 under the Smart and Connected Community program as well as the industrial affiliates of NYU WIRELESS. In addition, Azzino, Mezzavilla and Rangan were supported by NSF grants 1925079, 1564142, and 1547332 and the Semiconductor Research Corporation (SRC). Yuan and Wang were also supported by NSF grant 2003182. Yu Hao was also partially supported by NYUAD Institute (Research Enhancement Fund - RE132). Special thanks are also given to Remcom that provided the Wireless Insite software and GeoPipe that provided the 3D models.

REFERENCES

[1] N. Niknejad, W. B. Ismail, A. Mardani, H. Liao, and I. Ghani, “A comprehensive overview of smart wearables: The state of the art literature, recent advances, and future challenges,” Engineering Applications of Artificial Intelligence, vol. 90, p. 103529, 2020.
[2] W. Sun, J. Liu, and H. Zhang, “When smart wearables meet intelligent vehicles: Challenges and future directions,” IEEE wireless communications, vol. 24, no. 3, pp. 58–65, 2017.
[3] C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazeltown, E. Isaac, Y. Jia, B. Jia et al., “Machine learning at Facebook: Understanding inference at the edge,” in 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), IEEE, 2019, pp. 331–344.
[4] V. C. Leung, X. Wang, A. Jamalipour, X. Chen, and S. Bouzefrane, “IEEE Access Special Section Editorial: Edge Computing and Networking for Ubiquitous AI,” IEEE Access, vol. 9, pp. 90933–90936, 2021.
[5] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, “In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning,” IEEE Network, vol. 33, no. 5, pp. 156–165, 2019.
[6] S. Rangan, T. S. Rappaport, and E. Erkip, “Millimeter-wave cellular wireless networks: Potentials and challenges,” Proceedings of the IEEE, vol. 102, no. 3, pp. 366–385, 2014.
[7] M. Shafie, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, “5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice,” IEEE Journal on Selected Areas in Communications, vol. 35, no. 6, pp. 1201–1221, 2017.
[8] A. N. Uwaechia and N. M. Mahyuddin, “A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges,” IEEE Access, vol. 8, pp. 62367–62414, 2020.
[9] J.-R. Rizzo, C. Feng, W. Riewpaiboon, and P. Mongkoltip, “A low-vision navigation platform for economies in transition countries,” 10 2020, pp. 1–3.
[10] L. Niu, C. Qian, J.-R. Rizzo, T. Hudson, Z. Li, S. Enright, E. Sperling, K. Conti, E. Wong, and Y. Fang, “A wearable assistive technology for the visually impaired with door knob detection and real-time feedback for hand-to-handle manipulation,” 10 2017, pp. 1500–1508.
[11] W. Gui, B. Li, S. Yuan, J.-R. Rizzo, L. Sharma, C. Feng, A. Tzes, and Y. Fang, “An assistive low-vision platform that augments spatial cognition through proprioceptive guidance: Point-to-tell-and-touch,” 11 2019, pp. 381–3822.
[12] A. Boldini, A. L. Garcia, M. Sorrentino, M. Beheisthi, O. Ogedegbe, Y. Fang, M. Portfiri, and J.-R. Rizzo, “An inconspicuous, integrated electronic travel aid for visual impairment,” ASME Letters in Dynamic Systems and Control, vol. 1, no. 4, p. 041004, 2021.
[13] X. Li, H. Cui, J.-R. Rizzo, E. K. Wong, and Y. Fang, “Cross-safe: A computer vision-based approach to make all intersection-related pedestrian accessible for the visually impaired,” 2019.
[14] J.-R. Rizzo, Y. Pan, T. Hudson, E. K. Wong, and Y. Fang, “Sensor fusion for ecologically valid obstacle identification: Building a comprehensive assistive technology platform for the visually impaired,” in 2017 7th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO), 2017, pp. 1–5.
[15] R. A. Shoureshi, J. R. Rizzo, and T. E. Hudson, “Smart wearable systems for enhanced monitoring and mobility,” in Advances in Science and Technology, vol. 100, Trans Tech Publ, 2017, pp. 172–178.
[16] P. Phamduy, J.-R. Rizzo, T. Hudson, M. Torre, K. Levin, and M. Portfiri, “Communicating through touch: Macro fiber composites for tactile stimulation on the abdomen,” IEEE Transactions on Haptics, vol. PP, pp. 1–1, 12 2017.
[17] A. Boldini, J.-R. Rizzo, and M. Portfiri, “A piezoelectric-based advanced wearable: obstacle avoidance for the visually impaired built into a backpack,” in Nano-, Bio-, Info-Tech Sensors, and 3D Systems IV, vol. 11378, International Society for Optics and Photonics, 2020, p. 1137806.
[18] B. Jedari, G. Premnaskar, G. Illahi, D. Francescosco, A. Mehrabi, and A. Ylä-Jääski, “Video Caching, Analytics, and Delivery at the Wireless Edge: A Survey and Future Directions,” IEEE Communications Surveys & Tutorials, vol. 23, no. 1, pp. 431–471, 2020.
[19] E. Li, L. Zeng, Z. Zhou, and X. Chen, “Edge AI: On-demand accelerating deep neural network inference via edge computing,” IEEE Transactions on Wireless Communications, vol. 19, no. 1, pp. 447–453, 2019.
[20] H. Sun, Y. Yu, K. Sha, and B. Lou, “mvideO: edge computing based mobile video processing systems,” IEEE Access, vol. 8, pp. 11 615–11 623, 2019.
[21] L. Liu, H. Li, and M. Gruteser, “Edge assisted real-time object detection for mobile augmented reality,” in The 25th Annual International Conference on Mobile Computing and Networking, 2019, pp. 1–6.
[22] X. Ran, H. Chen, X. Zhu, Z. Liu, and J. Chen, “Deepdecision: A mobile deep learning framework for edge video analytics,” in IEEE INFOCOM 2018-IEEE Conference on Computer Communications. IEEE, 2018, pp. 1421–1429.
[23] G. R. MacCartney, T. S. Rappaport, and S. Rangan, “Rapid fading due to human blockage in pedestrian crowds at 5G millimeter-wave frequencies,”
in GLOBECOM 2017-2017 IEEE Global Communications Conference. IEEE, 2017, pp. 1–7.

[24] C. Slezk, V. Semkin, S. Andrevev, Y. Koucheryavy, and S. Rangan, "Exemplary effects of dynamic human-body blockage in 60 GHz commu- nications," IEEE Communications Magazine, vol. 56, no. 12, pp. 60–66, 2018.

[25] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave mimo systems," IEEE journal of selected topics in signal processing, vol. 10, no. 3, pp. 436–453, 2016.

[26] M. Girodani, M. Freeze, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 173–196, 2018.

[27] J. Rizzo, "Lab website," https://rizzolab.org.

[28] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Koratikattara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7310–7311.

[29] Remcom, "Wireless InSite," https://www.rencom.com/wireless-insite-propagation-software.

[30] Geoipipe, "3D Data Models," https://geopi.pe.

[31] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-end simulation of 5g mmwave networks," IEEE Communications Surveys Tutorials, vol. 20, no. 3, pp. 2237–2263, 2018.

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

[33] VideoLAN Organization, "x265," https://www.x265.org.

[34] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré, and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in 2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017, pp. 6533–6540.

[35] R. M. Sidek, and A. A. Mutalib, "Smart cane: Assistive cane for visually-impaired people," arXiv preprint arXiv:1110.5156, 2011.

[36] W. Elmannani and K. Ellerbyth, "Sensor-based assistive devices for visually-impaired people: current status, challenges, and future directions," Sensors, vol. 17, no. 3, p. 565, 2017.

[37] W. Elmannani and K. Ellerbyth, "Sensor-based assistive devices for visually-impaired people: current status, challenges, and future directions," Sensors, vol. 17, no. 3, p. 565, 2017.

[38] B. Dougherty, K. Kehler, R. Jamara, N. Patterson, D. Valenti, and F. Vera- Diaz, "Abandonment of low-vision devices in an outpatient population," 01 2016.

[39] R. Farcy, R. Leroux, A. Jucha, R. Damascini, C. Grégoire, and A. Zogaghi, "Electronic travel aids and electronic orientation aids for blind people: technical, rehabilitation and everyday life points of view," in Proceedings of the Conference and Workshop on Assistive Technology for Vision and Hearing Impairment. Citeseer, 2006.

[40] J. M. Bochsler, G. E. Legge, R. Gage, and C. S. Kallie, "Recognition of signs and steps by people with low vision," Investigative Ophthalmology & Visual Science, vol. 54, no. 1, p. 288–294, January 2013.

[41] G. E. Legge, C. Downey, N. Giudice, and B. Tjan, "Indoor airport wayfinding for blind and visually impaired travelers," 01 2016.

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

[43] J. Hu, Z. Ma, and Y. Wang, "Optimization of spatial, temporal and amplitude resolution for rate-constrained video coding and scalable video adaptation," in 2012 19th IEEE International Conference on Image Pro- cessing. IEEE, 2012, pp. 717–720.

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

[45] S. Labs, "ZED 2 Camera product page," https://www.stereolabs.com/zed-2.

[46] FFmpeg project, “FFmpeg,” https://ffmpeg.org.

[47] J. Redmon, A. Farhadi, and Y. Wang, “YOLOv3: Better, Faster, Stronger,” 2016.

[48] W. Xia, S. Rangan, M. Mezzavilla, A. Lozano, G. Geraci, V. Semkin, and G. Loianio, “Millimeter wave channel modeling via generative neural networks,” in Proc. IEEE Globecom Workshops., 2020, pp. 1–6.

[49] A. Alkhathe, “DeepMIMO: A generic deep learning dataset for millime- ter wave and massive MIMO applications,” in Proc. of Information Theory and Applications Workshop (ITA), San Diego, CA, Feb 2019, pp. 1–8.

[50] W. Xia, S. Rangan, M. Mezzavilla, A. Lozano, G. Geraci, V. Semkin, and G. Loianio, “Millimeter wave channel modeling via generative neural networks,” in Proc. IEEE Globecom Workshops., 2020, pp. 1–6.
ZHONGZHEN (JACKY) YUAN received the B.S. and M.S. in electrical engineering from the New York University Tandon School of Engineering, Brooklyn, NY, USA, in 2019 and 2020, respectively. He had held internships positions in Media Lab at Tencent America, and Incubation Lab at Dolby Laboratories.

He is currently a Ph.D. candidate in electrical engineering at New York University, under the supervision of Prof. Yao Wang. His research interests include image/video compression, computer vision, and deep learning.

TOMMY AZZINO received the B.Sc. and M.Sc. in information engineering and telecommunications engineering, respectively, from the University of Padova, Italy, in 2016 and 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering with New York University Tandon School of Engineering, Brooklyn, NY, USA, under the supervision of Prof. S. Rangan. He held visiting research positions with the National Institute of Standards and Technology (NIST), Department of Commerce, Gaithersburg, MD, USA, in 2018, Nokia Bell Labs France, Paris, France, in 2019.

He took part to the 2018 edition of the Seeds for the Future Project promoted by Huawei Technologies in Shenzhen, PRC, in 2018. His research interests include 5G mobile cloud/edge computing, 5G network simulation and prototyping, MAC and network-layer resource allocation, wireless communications, and machine learning for wireless communications.

YU HAO received his B.Eng degree in Computer Science and Technology from Beijing University of Technology, Beijing, China in 2019. He received an M.S degree in Computer Science from New York University, New York, USA in 2021. He is currently pursuing the Ph.D. degree in Computer Science at New York University. His research interests include Computer Vision, Computer Graphics, and Deep Learning.

YIXUAN LU was born in Beijing, China in 1996. He received a B.E. degree in automation from Nanjing University of Science and Technology in 2019 and now pursuing a master’s degree in New York University majoring in Computer Engineering. His research interests include video/image compression methods, computer vision, and deep neural networks.

HAOYANG PEI was born in Yuncheng, Shanxi, China in 1995. He received a B.E. degree in electrical engineering and its automation from North China Electric Power University in 2017 and now pursuing a master’s degree in New York University majoring in Computer Engineering. His research interests include video/image compression, computer vision, and deep learning.

ALAIN BOLDINI received B.Sc. degree in Aerospace Engineering and M.Sc. degree in Aeronautical Engineering from Politecnico di Milano, Milan, Italy, in 2015 and 2017, respectively, and M.Sc. in Aerospace Engineering from Politecnico di Torino, Turin, Italy, in 2018. He is currently a PhD candidate in Mechanical Engineering at the New York University Tandon School of Engineering. He is involved in research in smart advanced materials, assistive technologies, and dynamical systems.

MARCO MEZZAVILLA received the B.A.Sc., the M.Sc., and the Ph.D. at the University of Padua, Italy, all in Electrical Engineering. He held visiting research positions at the NEC Network Laboratories in Heidelberg (Germany, 2009), at the Telematics Department at Politecnico di Catalafia (UPC) in Barcelona (Spain, 2010), and at Qualcomm Research in San Diego (USA, 2012). He joined New York University (NYU) Tandon School of Engineering in 2014, where he is currently a Research Faculty and leads multiple mmWave-related research projects, mainly focusing on 5G and beyond PHY/MAC design. He is serving as reviewer for many IEEE and ACM conferences, journals, and magazines. His research interests include design and validation of communication protocols and applications for 5G and beyond broadband wireless technologies, millimeter-wave communications, multimedia traffic optimization, radio resource management, spectrum sharing, convex optimization, cognitive networks, and novel network paradigms. He is a Senior Member of the IEEE.
MAHYA BEHESHTI received her M.D. degree from Gulf Medical University, School of Medicine, Ajman, UAE. She is a Physician-scientist at NYU Langone Medical Center’s Rusk Rehabilitation. She is also pursuing the Ph.D. degree in Mechanical and Aerospace Engineering Department, NYU Tandon School of Engineering. She is the author of 3 book chapters, more than 17 peer-reviewed articles and 13 poster presentations. She has also given multiple talks. Her research interests lie within the realm of neurorehabilitation, human-machine interface, specifically EEG-based pattern recognition and wearable technologies.

Dr. Beheshti was a recipient of Distinguished Alumni Award of Health-care Academia from Gulf Medical University and was chosen as Intern of the Year during her medical training. She is a member of American Society of Neuroradiology (ASNR), American Medical Association (AMA), National Multiple Sclerosis Society, and Association of Academic Physiatry (AAP). She is IEEE student member (IEEE RAS, IEEE EMBS, and IEEE WIE).

MAURIZIO PORFIRI (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in engineering mechanics from Virginia Tech, Blacksburg, VA, USA, in 2000 and 2006, respectively, and the “Laurea” (Hons.) degree in electrical engineering and the Ph.D. degree in theoretical and applied mechanics from the Sapienza University of Rome, Rome, Italy, and the University of Toulon, Toulon, France (under a dual-degree program between Italy and France), in 2001 and 2005, respectively.

He is currently an Institute Professor with the New York University Tandon School of Engineering, Brooklyn, NY, USA, having appointments with the Center for Urban Science and Progress, Department of Mechanical and Aerospace Engineering, Department of Biomedical Engineering, and Department of Civil and Urban Engineering. He is the Director of the Dynamical Systems Laboratory. He is involved in conducting and supervising research on dynamical systems theory and applications. He has authored more than 350 journal publications. He has been included in the “Brilliant 10” list of Popular Science in 2010. His research featured in major media outlets, including CNN, NPR, Scientific American, and Discovery Channel.

Dr. Porfiri is the recipient of the National Science Foundation CAREER Award, the Outstanding Young Alumnum Award by the College of Engineering, Virginia Tech, the American Society of Mechanical Engineers (ASME) Gary Anderson Early Achievement Award, the ASME DSCD Young Investigator Award, and the ASME C.D. Mote, Jr. Early Career Award. His other significant recognitions include invitations to the Frontiers of Engineering Symposium and the Japan–America Frontiers of Engineering Symposium organized by the National Academy of Engineering He has served on the Editorial Board of the ASME Journal of Dynamics Systems, Measurements and Control, the ASME Journal of Vibrations and Acoustics, Flow, the IEEE CONTROL SYSTEMS LETTERS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: REGULAR PAPERS, and Mechatronics. He is a Fellow of the ASME.

TODD E. HUDSON is an Assistant Professor of Rehabilitation Medicine at New York University’s Grossman School of Medicine, holding cross-appointments in Neurology, and also in the Department of Biomedical Engineering at the New York University Tandon School of Engineering.

Prof. Hudson is a Computational Neuroscientist whose research focuses on modeling sensory and motor systems, particularly with regard to eye and arm movements in healthy and disease states, movement planning, and spatial orientation. He received his PhD from Columbia University, and has authored over 50 peer reviewed articles, as well as the textbook ‘Bayesian Data Analysis for the Behavioral and Neural Sciences’ from Cambridge University Press.

WILLIAM SEIPLE received the B.S. and M.S. degrees in psychology from the Albright College, Reading, PA and the University of North Carolina, Greensboro, NC and the Ph.D. degree in zoology from the University of Illinois, Urbana, IL.

He is the Chief Research Officer at Lighthouse Guild, Research Professor of Ophthalmology at New York University School of Medicine, and Adjunct Faculty at the Institut de la Vision, Paris. He is the author of two books chapters, and more than 150 peer-reviewed articles.

His research interests include development and assessment of functional interventions for people with vision loss, rehabilitation training, and visual electrophysiology and psychophysics of vision.

YI FANG received the BS and MS degrees in biomedical engineering from Xi’an Jiaotong University, Xi’an, China, in 2003 and 2006, respectively, and the PhD degree in mechanical engineering from Purdue University, West Lafayette, Indiana, in 2011. He is currently an assistant professor with the Department of Electrical and Computer Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. His research interests include three-dimensional computer vision and pattern recognition, large-scale visual computing, deep visual computing, deep cross-domain and cross modality multimedia analysis, and computational structural biology.
SUNDEEP RANGAN received the B.A.Sc. at the University of Waterloo, Canada and the M.Sc. and Ph.D. at the University of California, Berkeley, all in Electrical Engineering. He has held postdoctoral appointments at the University of Michigan, Ann Arbor and Bell Labs. In 2000, he co-founded (with four others) Flarion Technologies, a spin-off of Bell Labs, that developed Flash OFDM, the first cellular OFDM data system and pre-cursor to 4G cellular systems including LTE and WiMAX. In 2006, Flarion was acquired by Qualcomm Technologies. Dr. Rangan was a Director of Engineering at Qualcomm involved in OFDM infrastructure products. He joined NYU Tandon (formerly NYU Polytechnic) in 2010 where he is currently a Professor of Electrical and Computer Engineering. He is a Fellow of the IEEE and the Associate Director of NYU WIRELESS, an industry-academic research center on next-generation wireless systems.

JOHN-ROSS (JR) RIZZO was born in Wayne, NJ, USA. He received his medical degree from New York Medical College, Alpha Omega Alpha Honors, on academic scholarship. He is currently a physician-scientist at NYU Langone Medical Center’s Rusk Rehabilitation, where he serves as vice chair of Innovation and Equity for Physical Medicine and Rehabilitation with cross-appointments in the Department of Neurology and the Departments of Biomedical & Mechanical and Aerospace Engineering at NYU-Tandon School of Engineering. He is also the Associate Director of Healthcare for the NYU Wireless Laboratory in the Department of Electrical and Computer Engineering at NYU-Tandon. He leads the Visuomotor Integration Laboratory (VMIL) and the REACTIV Laboratory (Rehabilitation Engineering Alliance and Center Transforming Low Vision), where his team focuses on assistive technology for the visually impaired and benefits from his own personal experiences with vision loss. He is the author of 10 book chapters, more than 80 peer-reviewed articles and many poster presentations. His research interests lie within the realm of neurorehabilitation, assistive technologies, and health equity.

Dr. Rizzo was awarded the prestigious Crain’s 40 under 40 award in New York Business for his medical devices, including his wearable technology. He has also been featured in several lay articles and also featured in videos and press releases. In 2018, he was a highlighted speaker in NYU’s TEDx “Re-Vision” Series. He is a member of American Medical Association, American College of Physicians, American Academy of PM&R (AAPM&R), Association of Academic Physiatry (AAP), and American Heart Association.

YAO WANG is a Professor at New York University Tandon School of Engineering (formerly Polytechnic University, Brooklyn, NY), with joint appointment in Departments of Electrical and Computer Engineering and Biomedical Engineering. She is also Associate Dean for Faculty Affairs for NYU Tandon since June 2019. Her research areas include video coding and streaming, multimedia signal processing, computer vision, and medical imaging.

She is the leading author of a textbook titled Video Processing and Communications, and has published over 250 papers in journals and conference proceedings. She received New York City Mayor’s Award for Excellence in Science and Technology in the Young Investigator Category in year 2000. She was elected Fellow of the IEEE in 2004 for contributions to video processing and communications. She received the IEEE Communications Society Leonard G. Abraham Prize Paper Award in the Field of Communications Systems in 2004, and the IEEE Communications Society Multimedia Communication Technical Committee Best Paper Award in 2011. She was a keynote speaker at the 2010 International Packet Video Workshop, INFOCOM Workshop on Contemporary Video in 2014, the 2018 Picture Coding Symposium, and the 2020 ACM Multimedia Systems Conference (MMSys’20). She received the NYU Tandon Distinguished Teacher Award in 2016.