Learning to Detect Fallen People in Virtual Worlds

Fabio Carrara
fabio.carrara@isti.cnr.it
ISTI CNR
Italy

Lorenzo Pasco
l.pasco@studenti.unipi.it
University of Pisa
Italy

Claudio Gennaro
claudio.gennaro@isti.cnr.it
ISTI CNR
Italy

Fabrizio Falchi
fabrizio.falchi@isti.cnr.it
ISTI CNR
Italy

ABSTRACT

Falling is one of the most common causes of injury in all ages, especially in the elderly, where it is more frequent and severe. For this reason, a tool that can detect a fall in real time can be helpful in ensuring appropriate intervention and avoiding more serious damage. Some approaches available in the literature use sensors, wearable devices, or cameras with special features such as thermal or depth sensors. In this paper, we propose a Computer Vision deep-learning based approach for human fall detection based on largely available standard RGB cameras. A typical limitation of this kind of approaches is the lack of generalization to unseen environments. This is due to the error generated during human detection and, more generally, due to the unavailability of large-scale datasets that specialize in fall detection problems with different environments and fall types. In this work, we mitigate these limitations with a general-purpose object detector trained using a virtual world dataset in addition to real-world images. Through extensive experimental evaluation, we verified that by training our models on synthetic images as well, we were able to improve their ability to generalize. Code to reproduce results is available at https://github.com/lorepas/fallen-people-detection.

CCS CONCEPTS

• **Software and its engineering** → *Virtual worlds training simulations*; • **Computing methodologies** → **Object detection**; *Activity recognition and understanding*.

KEYWORDS

visual fallen people detection, virtual worlds for synthetic data, object detection, scarce data

ACM Reference Format:
Fabio Carrara, Lorenzo Pasco, Claudio Gennaro, and Fabrizio Falchi. 2022. Learning to Detect Fallen People in Virtual Worlds. In *International Conference on Content-based Multimedia Indexing (CBMI 2022), September 14–16, 2022, Graz, Austria.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3549555.3549573

1 INTRODUCTION

As reported by World Health Organization [15], falls are the second leading cause of unintentional injury deaths worldwide. The people who suffer the most from falls are adults over 60; if not timely assisted, falls may lead to severe injuries and even death. For these reasons, a monitoring system that can identify a fall can be beneficial to giving first aid as soon as possible.

Among available techniques, vision-based ones offer a cheap and minimally invasive solution, as they do not require specialized hardware other than a standard video camera and do not involve wearing battery-powered sensors. Computer vision algorithms based on deep learning have already demonstrated excellent performance in detecting people and objects from video streams and can be adapted to detect fallen people. However, training a robust deep-learning-based detection model implies having a large labeled dataset covering a wide variety of scenarios. This is usually not available, as large sets are usually too expensive to obtain, limiting the generalization capabilities of the trained models.

This paper proposes to use synthetic data from virtual worlds to build robust vision-based fall detection models. We exploit characters and scenarios of a highly photo-realistic video game to create automatically-labeled synthetic images of fallen and non-fallen people. Controlling the game engine allows us to collect labeled data under different settings, varying scenes, characters, lighting, number of people in the scene, and camera pose. We create and release *Virtual World Fallen People* (VWFP) — a collection of 6071 synthetic labeled images depicting fallen and non-fallen people.

We set up a frame-based visual fall detection pipeline based on widely used object detector deep neural networks. We explore different configurations of synthetic data usage to train more robust fallen people detectors. Experiments on existing non-synthetic benchmarks show that synthetic data helps improve the generalization capabilities of detection models compared to the same models trained only on small training collections of real images.

2 RELATED WORK

Due to its potential societal impact, fallen people detection is an actively studied field of research with many proposed approaches and methodologies.

Sensor-based Fall Detection. A significant body of work proposes to detect falls employing sensors, alone or combined with video data. For example, Dovgan et al. [7] adopt an approach that performs different tests in sensor data applied to different body parts. Martínez-Villaseñor et al. [13] uses a multimodal approach, collecting both data from videos and sensors, in which the subjects perform 11 different activities (six normal daily activities and five different types of falls). Falls are recognized using a shallow classifier on handcrafted features. Another multimodal approach can be seen in Kwolek and Kepski [8], in which both video images and acceleration data are collected through two Microsoft Kinect cameras.

CBMI 2022, September 14–16, 2022, Graz, Austria

ACM ISBN 978-1-4503-9720-9/22/09…$15.00
https://doi.org/10.1145/3549555.3549573

An SVM on defined features detects falls. Antonello et al. [1] propose an open-source solution for an autonomous robotic platform for home care. They build a dataset that contains both images and point clouds for this scope. Geometric consistency checks together with SVM classifiers are used to discern fallen people. In general, those solutions require specific hardware and setups to be deployed without the guarantee of being robust to different scenarios.

Visual Fallen People Detection. Detecting fallen people from video streams makes systems easier to deploy, as video cameras are accessible and often already available, e.g., existing surveillance cameras. However, extracting information from RGB data poses additional challenges. Seminal work [4] tries to detect a fall by using handcrafted spatiotemporal descriptors and shallow classifiers trained on a limited number of scenarios. More recent solutions adopt AI-based computer vision; Maldonado-Bascón et al. [12] use the YOLO object detector to find people in images taken by a mobile-patrol robot. They adopt an SVM to decide whether detected people are fallen given geometric features on bounding box position and aspect ratio. The main roadblock to a robust solution in these works is the limited generalization capabilities given by ad-hoc training sets or camera configurations.

Synthetic data for Fall Detection. The most relevant work to our proposal of using synthetic data for fall detection is Asif et al. [2]. The authors propose a multimodal segmentation and human skeleton pose estimation model to detect people and their pose from RGB data. The segmentation map and skeleton position are fed to FallNet, a deep classifier that discerns falls from non-falls and is trained on synthetic segmentation and skeleton data generated with the MakeHuman tool. Unlike their work, our approach aims to train a single detection model able to discern fallen and non-fallen people from RGB data directly. Our solution is more straightforward and probably more efficient, but it is more challenging to generate the synthetic RGB data needed to train the model. For the generation of RGB data, we follow a collection procedure of Di Benedetto et al. [5] with the difference that we collect also fall/non-fall information instead of object bounding boxes.

3 VISUAL FALLEN PEOPLE DETECTION

Our goal is to compare the performance of detectors trained with and without virtual data on available real-data benchmarks. To this end, we set up a common pipeline to detect fallen people from video streams described below.

We formulate the task of visual fallen people detection as a frame-based analysis of the video stream. Each analyzed frame is processed by a deep object detection neural network to find people in the image and, if any, classify them as either fallen or non-fallen. Although the pipeline could be improved by including temporal information and additional logic (e.g., adding alarm hysteresis), here we keep a simple detection pipeline for the sake of easy comparison.

We assume to have a synthetic training set and a real training set. We explore four configurations of training data for our detector models, which are

Real-only Data (R), a baseline configuration in which the training phase uses only the real data,

Table 1: Datasets Statistics. Datasets were adapted as described in Section 4.

	# imgs	# fallen	# non-fallen	annot/img
VWFP (ours)	6071	7456	26125	5.53 ± 2.68
FPDS [12]	6832	5019	2247	1.06 ± 0.26
train set	4699	3863	1004	1.03 ± 0.20
validation set	1174	765	413	1.00 ± 0.06
test set	959	391	830	1.27 ± 0.48
Elderly [11]	412	357	65	1.02 ± 0.15
URFD [8]	421	182	239	1.00 ± 0.02

Virtual-only Data (V) in which the training phase uses only the synthetic game data,

Virtual then Real Data (V → R) in which the model is initially trained on synthetic data and then fine-tuned on real data, and

Virtual and Real Data (V + R) in which the model is trained on a mixture of synthetic and training data.

Once detector models are trained, we test them by measuring their performance on real-world benchmark data.

4 DATASETS

Table 1 collects statistics of the datasets used in this work and described in this section. First, we introduce our novel synthetic dataset for visual fallen people detection and its collection procedure. Then, we review the existing real-data benchmarks we will use to validate the trained models.

4.1 Virtual World Fallen People (VWFP)

We collect a novel synthetic dataset for fallen people detection called *Virtual World Fallen People* (VWFP). VWFP comprises images extracted from the highly photo-realistic video game *Grand Theft Auto V* developed by *Rockstar North*[1]. Each image is automatically labeled by extracting from the game engine the information about people present in the scene, i.e., their bounding boxes and their status (fallen or non-fallen). Specifically,

(1) we set up a scenario picking a time of the day, a weather condition, and a position in the game map at random,
(2) we instantiate at most 30 pedestrians around the center of the scene, let them wander in the area, and kill approximately half of them, having the consequence that they fall on the ground,
(3) we place the camera in a random position pointing it to the scene,
(4) we take a snapshot of the scene and collect the 2D bounding boxes and status (fallen/non-fallen) of visible pedestrians, and
(5) we repeat steps (3) and (4) five times to obtain more viewpoints of the same scene.

The above actions are implemented via game engine API calls available via modding hooks. Technical details are available in

[1] https://www.rockstarnorth.com/

Di Benedetto et al. [5, 6]. We filter out images containing no labeled objects that mostly correspond to bad game locations (e.g., oceans) and highly occluded camera angles. We also filter out annotations for objects that are more than 80% occluded (e.g., by walls, trees, or other pedestrians) in the capturing viewpoint. After cleaning, we obtain 6,071 images depicting 7,456 fallen and 26,125 non-fallen people. Figure 1 shows samples from VWFP. The dataset is publicly available [3].

4.2 Benchmark Datasets

In this section, we discuss three frame-based real-world benchmark datasets for visual fallen people detection that we adopt in our experimental evaluation, i.e., FPDS [12], Elderly [11], and URFD [8] datasets. All the datasets comprise a collection of images with bounding box annotations localizing fallen and non-fallen people in the scene. Figure 2 shows some samples from each collection. Details of each benchmark are given in the following paragraphs.

FPDS. comprises 6,832 images depicting fallen or non-fallen people. This is one of the largest benchmarks for visual fall detection with frame-level annotations. Since video data were meant to be captured and analyzed by a domestic robot, images are captured using a single camera from 76 cm above the floor. Moreover, most images depict indoor scenarios with precisely a single person in the scene, even if outdoor scenarios and multi-instances images do occur. We adopt the original train, validation, and test splits provided by the authors. However, through manual inspection of the dataset, we filter out images having no or bad annotations (e.g., incorrect values for the coordinates of bounding boxes). Dataset statistics after cleaning are available in Table 1.

Elderly. is a smaller collection of images collected by the same authors of FPDS with the same intent. This collection comprises 413 images of volunteer subjects over 65 years old depicted in standing, sitting, and lying postures in indoor scenarios. As in FPDS, people bounding boxes and fallen/non-fallen status are provided for each image, and we manually fixed missing or wrong annotations on dataset inspection. With respect to FPDS, object distribution is even more skewed towards fallen people.

URFD. is a collection of 70 image sequences comprising 30 fall sequences and 40 sequences of daily living activities. Fall events are recorded indoors with 2 Microsoft Kinect cameras and corresponding accelerometer data at 30 fps. We only kept RGB data with bounding box annotations of fall sequences for our purposes. Moreover, we pick every fifth frame from each sequence to reduce the redundancy of consequent video frames. In the end, we retain 421 images for evaluation consisting of 239 non-fallen and 182 fallen instances.

5 EXPERIMENTAL EVALUATION

5.1 Model and Training Details

Here we describe procedures for defining and training detectors that are common for all performed experiments.

As object detection model, we use the widely adopted Faster-RCNN [14] with a ResNet-50 with Feature Pyramid Network [9] as backbone and a two-class detection head. We initialize all the model parameters from the model snapshot pre-trained on the COCO train2017 set [10] except for the detection head in which parameters are randomly initialized. We freeze the parameters of the backbone network except for the last three residual blocks that are trainable.

The model is trained with SGD for 10 epochs with a learning rate of $5 \cdot 10^{-3}$ divided by 10 every 3 epochs, momentum of 0.9, and weight decay of $5 \cdot 10^{-4}$. We applied data augmentation using random cropping, random horizontal flipping, and random color augmentations (contrast/brightness, color jitter, or grayscale transformations). We cope with class unbalance by weighting images based on the distribution of object classes in it; we assign weights to each annotation such that fallen and non-fallen objects are balanced in the entire training set, and then we set the weight of each image as the mean weight of its objects.

At each epoch, we measure the loss and the COCO mean average precision (mAP) on a validation set, and we select the snapshot that gave the maximum mAP as the final model.

5.2 Configuration Comparison

We train four models using the configurations of training data described in Section 3 and the training procedure detailed above. In (V), we use the proposed VWFP dataset with a random 80/20 train/validation split. In (R), we use the train and validation splits of FPDS. In (V $\rightarrow$ R), we fine-tune the model obtained in (V) using the train and validation splits of FPDS, or equivalently, we apply procedure (V) and (R) sequentially on the same model. In (V + R), we use VWFP, but we replace 30% of synthetic samples with real samples randomly picked from FPDS for both the training and validation splits.

We then test these four models on the test set of FPDS, on the URFD, and on the Elderly benchmarks, and we report results in Table 2. For each test set and configuration, we report the mAP as a threshold-independent metric and Recall, Precision, and F_1-score as threshold-dependent metrics. To make our results comparable to the work of Maldonado-Bascón et al. [12], for threshold-dependent metrics, we discard the localization information provided by the model (bounding boxes) and consider only the presence or absence of the fallen/non-fallen classes as in a binary classification problem. In this context, we report results using the threshold that maximizes the F_1-score.

We can observe that the (V + R) configuration consistently achieves good performance compared to other configurations, especially on Elderly, the benchmark on which FPDS data transfer worst, where it reaches 0.82 mAP compared to 0.68 of (R). Mixing synthetic and real data also provides a basic domain adaptation that helps transfer knowledge from the synthetic to the real domain. The (R) baseline configuration performs best when applied on the same domain used in training (FPDS) but can achieve significant results when transferred to other datasets. Still, results suggest that including virtual data can improve the generalization capability of the model free of labeling cost. The (V $\rightarrow$ R) configuration instead is often suboptimal; we deem that separating training phases on synthetic and real data tends to increase overfitting in each domain, preventing the model from learning robust features not specific to a particular domain. Note that without any real data, the (V) configuration still achieves

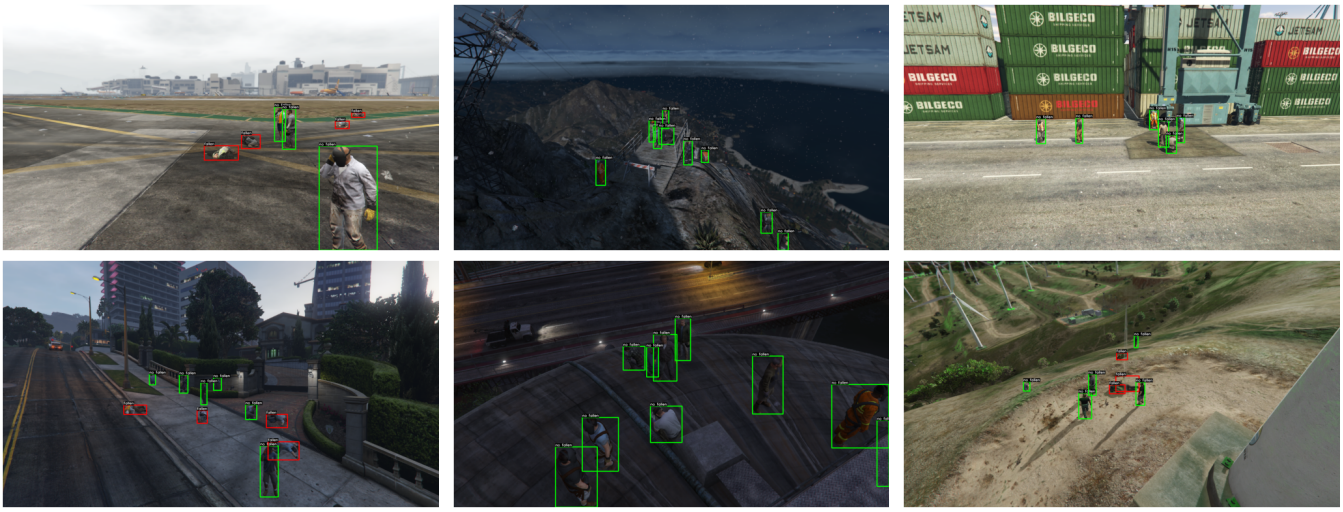


Figure 1: Samples from the Virtual World Fallen People (VWFP) Dataset. Green and red bounding boxes represent non-fallen and fallen people, respectively. The game engine and contents enable us to capture variability in background scenes, pedestrian looks and behaviors, and lightning and occlusion conditions. Best viewed in electronic format.

reasonable performance in some benchmarks (0.75 mAP on FPDS and 0.74 on URFD), indicating that useful knowledge is present in synthetic data and can be harnessed, especially in scenarios where real data is hard to collect.

6 CONCLUSIONS

The main challenge for robust detection of fallen people from visual data is the lack of large and varied training dataset. In this work, we tackled this problem by proposing the use of synthetic training data generated from virtual worlds to improve the generalization capabilities of fallen people detector models.

We generated and publicly released a varied and highly photo-realistic synthetic dataset of fallen and non-fallen people exploiting modern video games. We performed several experiments training widely used detector models with different configurations of real and synthetic images and testing them on publicly available visual fallen people detection datasets. Results showed that synthetic data improved the generalization capability of models when tested on unseen real scenarios. Specifically, mixing synthetic data with a small portion of real data in the training set gives the best detection performance in terms of mAP on unseen benchmarks.

Future directions include improving the synthetic dataset by exploring more contents offered by virtual worlds (more pedestrian models, poses, and behaviors, richer indoor scenarios, and more challenging lightning and occlusions). Moreover, future work may apply our proposal to more complex detection pipelines (e.g., including temporal information) and thus extend its evaluation by exploiting larger and different (e.g., non-frame-based) benchmarks.

ACKNOWLEDGMENTS

This work was partially funded by: AI4Media - A European Excellence Centre for Media, Society and Democracy (EC, H2020 n. 951911).

Table 2: Fallen people detection performance. For threshold-dependent metrics, we report values obtained with the threshold that maximizes the F_1-score. The best and second best entries are respectively indicated in bold and underlined.

(a) On FPDS test subset.

	mAP	thr	Recall	Precision	F_1-score
YOLO+SVM [12]	-	-	0.95	0.92	0.93
(R)	**0.97**	0.50	0.99	0.99	**0.99**
(V)	0.75	0.40	0.80	0.76	0.78
(V → R)	0.94	0.90	0.93	0.99	<u>0.96</u>
(V + R)	<u>0.96</u>	0.90	0.93	0.99	<u>0.96</u>

(b) On URFD.

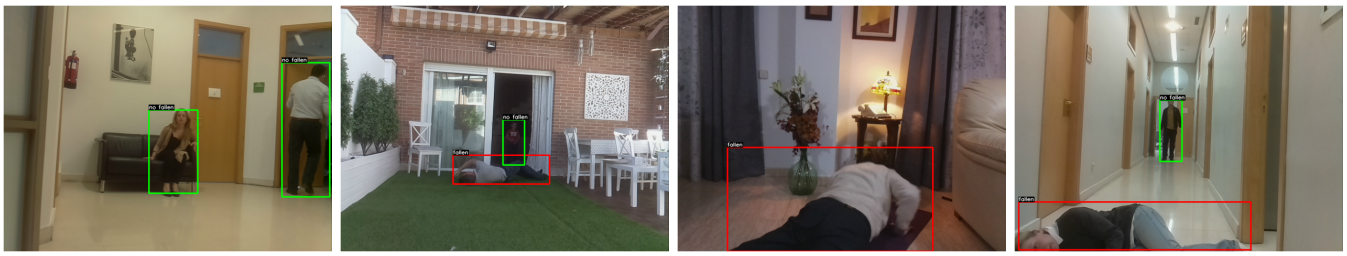
	mAP	thr	Recall	Precision	F_1-score
(R)	0.93	0.99	0.98	0.64	0.78
(V)	0.74	0.20	0.66	0.73	0.69
(V → R)	<u>0.98</u>	0.99	0.91	1.00	**0.95**
(V + R)	**0.99**	0.90	0.91	0.97	<u>0.94</u>

(c) On Elderly.

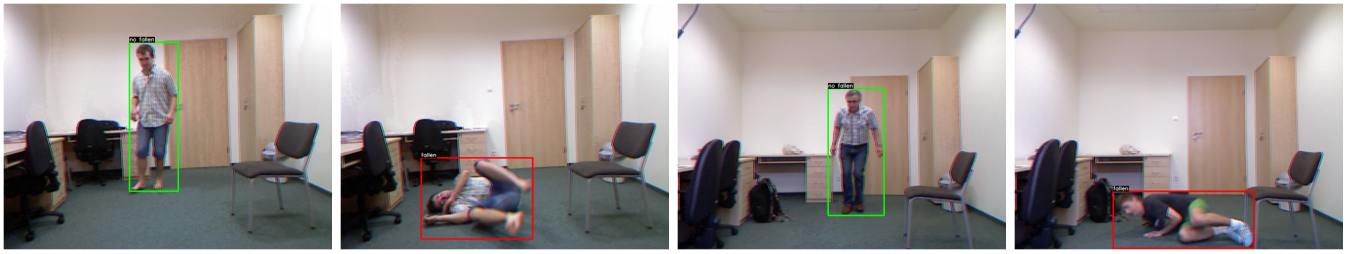
	mAP	thr	Recall	Precision	F_1-score
(R)	<u>0.68</u>	0.80	0.94	0.90	**0.92**
(V)	0.46	0.60	0.70	0.72	0.71
(V → R)	0.57	0.50	0.97	0.83	<u>0.90</u>
(V + R)	**0.82**	0.80	0.87	0.90	0.88

REFERENCES

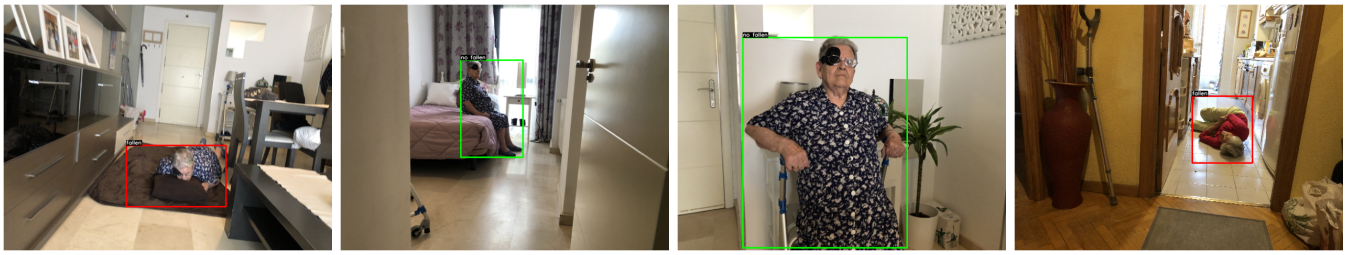
[1] Morris Antonello, Marco Carraro, Marco Pierobon, and Emanuele Menegatti. 2017. Fast and Robust detection of fallen people from a mobile robot. In *Intelligent*



(a) FPDS [12]



(b) URFD [8]



(c) Elderly [11]

Figure 2: Sample images from non-synthetic benchmarks. Green and red bounding boxes represent non-fallen and fallen people instances, respectively.

Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on. IEEE.

[2] Umar Asif, Benjamin Mashford, Stefan Von Cavallar, Shivanthan Yohanandan, Subhrajit Roy, Jianbin Tang, and Stefan Harrer. 2020. Privacy Preserving Human Fall Detection using Video Data. In *Proceedings of the Machine Learning for Health NeurIPS Workshop (Proceedings of Machine Learning Research, Vol. 116)*, Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones (Eds.). PMLR, 39–51. https://proceedings.mlr.press/v116/asif20a.html

[3] Fabio Carrara, Lorenzo Pasco, Claudio Gennaro, and Fabrizio Falchi. 2022. *VWFP: Virtual World Fallen People Dataset for Visual Fallen People Detection.* https://doi.org/10.5281/zenodo.6394684

[4] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. 2013. Optimized spatio-temporal descriptors for real-time fall detection: Comparison of support vector machine and Adaboost-based classification. *Journal of Electronic Imaging* 22 (10 2013), 041106–041106. https://doi.org/10.1117/1.JEI.22.4.041106

[5] Marco Di Benedetto, Fabio Carrara, Enrico Meloni, Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. 2021. Learning accurate personal protective equipment detection from virtual worlds. *Multimedia Tools and Applications* 80, 15 (2021), 23241–23253.

[6] Marco Di Benedetto, Enrico Meloni, Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. 2019. Learning safety equipment detection using virtual worlds. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–6.

[7] Erik Dovgan, Mitja Lustrek, Bogdan Pogorelc, Anton Gradisek, Helena Burger, and Matjaz Gams. 2011. Intelligent elderly-care prototype for fall and disease detection. *Zdravniski Vestnik* 80 (11 2011), 824–831.

[8] Bogdan Kwolek and Michal Kepski. 2014. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine* 117 (10 2014), 489–501. https://doi.org/10.1016/j.cmpb.2014.09.005

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[11] Saturnino Maldonado-Bascón, Cristian Iglesias-Iglesias, Pilar Martín-Martín, and Sergio Lafuente-Arroyo. 2019. Elderly Dataset. https://gram.web.uah.es/data/datasets/fpds/index.html. Accessed: 2019-03-28.

[12] Saturnino Maldonado-Bascón, Cristian Iglesias-Iglesias, Pilar Martín-Martín, and Sergio Lafuente-Arroyo. 2019. Fallen People Detection Capabilities Using Assistive Robot. *Electronics* 8, 9 (2019). https://doi.org/10.3390/electronics8090915

[13] Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafort-Asturiano. 2019. UP-Fall Detection Dataset: A Multimodal Approach. *Sensors* 19, 9 (2019). https://doi.org/10.3390/s19091988

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 [cs.CV]

[15] World Health Organization. [n.d.]. *Falls.* https://www.who.int/news-room/fact-sheets/detail/falls