Noname manuscript No.
(will be inserted by the editor)

WatchNet++: Efficient and accurate depth-based network for detecting people attacks and intrusion

M. Villamizar, A. Martínez-González, O. Canévet · J-M. Odobez

Received: date / Accepted: date

Abstract We present an efficient and accurate people detection approach based on deep learning to detect people attacks and intrusion in video surveillance scenarios. Unlike other approaches using background segmentation and pre-processing techniques, which are not able to distinguish people from other elements in the scene, we propose WatchNet++ that is a depth-based and sequential network that localizes people in top-view depth images by predicting human body joints and pairwise connections (links) such as head and shoulders. WatchNet++ comprises a set of prediction stages and up-sampling operations that progressively refine the predictions of joints and links, leading to more accurate localization results. In order to train the network with varied and abundant data, we also present a large synthetic dataset of depth images with human models that is used to pre-train the network model. Subsequently, domain adaptation to real data is done via fine-tuning using a real dataset of depth images with people performing attacks and intrusion.

An extensive evaluation of the proposed approach is conducted for the detection of attacks in airlocks and the counting of people in indoors and outdoors, showing high detection scores and efficiency. The network runs at 10 and 28 FPS using CPU and GPU, respectively.

Keywords Video surveillance · people detection · convolutional network · deep learning

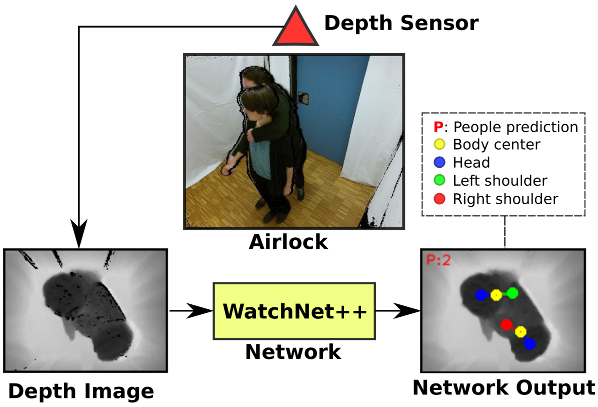

Fig. 1 We propose WatchNet++ which is an efficient and accurate depth-based network for people detection in video surveillance applications. It is able to identify people intrusion by detecting human joints and links with high accuracy.

1 Introduction

In recent years there has been a large deployment of computer vision systems for people detection and counting in video surveillance and analysis applications [4,6,12,25,29,31]. Some systems are, for instance, capable of detecting abnormal behaviors and alerting the teleoperators of potentially dangerous situations. These systems can be of primary necessity for security in public and private places such as banks, airports, shopping centers, and corporate buildings.

In this paper, we study the problem of intruder detection in building entrances from monocular depth cameras. More precisely, we focus on the detection of multiple people in restricted areas where one person is exclusively allowed at a time. This is a difficult problem since the video surveillance system must be able to detect people attacks and trickeries (such as tailgating and piggybacking to fool the system), see Figure 1.

M. Villamizar and O. Canévet, Idiap Research Institute, Switzerland. A. Martínez-González and J-M. Odobez, Idiap Research Institute and École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
E-mail: {michael.villamizar,olivier.canevet,angel.martinez, odobez}@idiap.ch

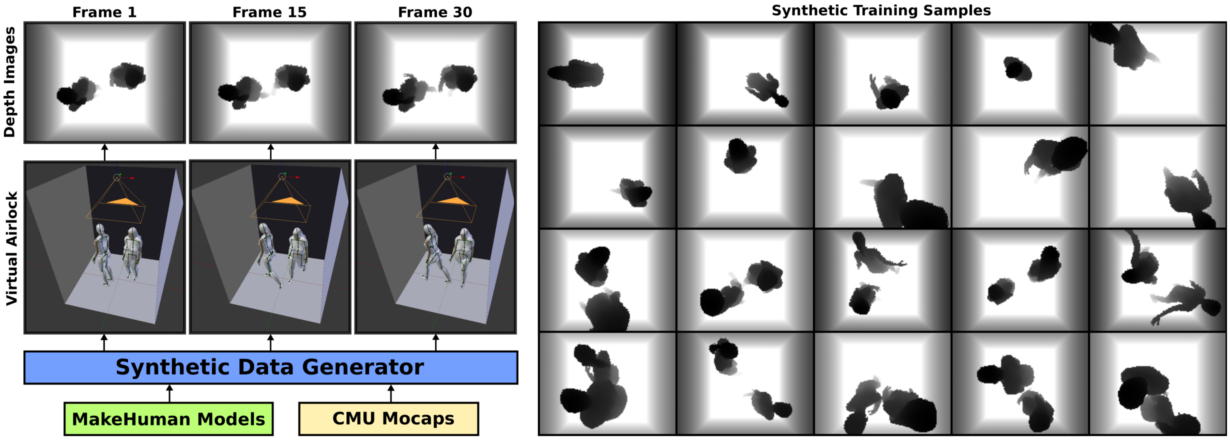

Fig. 2 Synthetic Data Generator (SDG). Left: SDG creates a virtual airlock with one or two people performing different actions to reproduce similar depth images to the Unicity database [10]. Right: Some example images generated by SDG for training WatchNet++.

To deal with this type of attacks, it is necessary to have a good visibility of the scene in order to reduce the degree of body occlusions caused by other people or elements in the scene. To solve this problem, the camera is commonly placed in a zenithal position in such a way that it is much harder to deceive the detection system when people are exactly below the camera [1,2,7,19,25,37]. Additionally, the use of overhead cameras allows to detect upper body parts with greater reliability which leads to better detection accuracy.

Another aspect related to surveillance systems is the privacy and data protection regulations. For example, detection systems based on color cameras have to apply algorithms and controls to maintain people's privacy. This leads to the use of other technologies such as depth cameras that are a great source of information for people detection, but when used alone can avoid this legal inconvenience [3,8,11]. Another advantage of using depth information versus color is that texture is removed allowing to compute lighter and more efficient methods for real-time video surveillance applications.

Aligned to these requirements, we present in this work a deep-learning approach to detect people efficiently and accurately from top-view depth cameras (Fig. 1). Particularly, we focus on the problem of detecting people under attack and intrusion in building airlocks as well as counting people in indoor and outdoor scenarios. The proposed approach relies on a depth-based network, named WatchNet++, that sequentially predicts the location of human body joints and links as well as the prediction the body center in order to estimate the number of people in the scene. We use the head and shoulders as body joints and the links between them and the body center as body links.

Additionally, we present a synthetic dataset which is used to pre-train the network model with abundant data to boost the performance on detection. It has a large number of artificial depth images with 3D human models performing different actions inside a virtual airlock, see Figure 2. To perform domain adaptation between synthetic and real data, we fine-tune the network using a real dataset of depth images including people attacks and intruders [10]. This approach is very convenient for deployment since the pre-trained model can be adapted to new scenarios (e.g corridors, elevators) using a relatively small dataset with real depth images.

This work builds on earlier publications [10,35]. However, this paper presents WatchNet++ which is an extended and improved version of the WatchNet network introduced in [35] for people detection in depth images. Specifically, WatchNet++ has two main novelties over the original version that make it more accurate and improve its detection performance: the use of a new prediction stage that increases the localization accuracy of body joints, and the integration and prediction of body joints and links simultaneously to obtain better detection results. In addition, we include a more extensive experimental validation of the method and provide a comparison with other detection approaches.

The main contributions are:

- An accurate network for people detection that simultaneously predicts body joints and links;
- A series of prediction modules that refines the localization of body joints, links, and centers;
- A synthetic depth image dataset to train the proposed network with abundant data, resulting in better detection rates;
- A detection and tracking approach to count people going through a corridor.

The remainder of the paper is organized as follows: section 2 presents the related work while section 3 introduces the synthetic and real datasets used to train the network. Section 4 describes WatchNet++ and its main constituents. Experimental validation is conducted in section 5. Finally, conclusions are provided in section 6.

2 Related Work

To control the access of people to public or private buildings, video surveillance systems are usually based on counting the number of people in the scene. Techniques for this can be divided into two main categories: feature-based and counting-by-detection methods.

Feature-based counting methods formulate the task as a regression problem, avoiding people detection, where image features are exploited to predict the number of people in the scene. This approach is particularly convenient for crowded scenarios such as public events or demonstrations since the detection of people is very challenging due to the high degree of occlusion [4,20,23,38]. However, this kind of methods provides a rough estimate about the number of people as well as a weak localization, represented commonly through density maps.

The second category relies on visual detectors to localize each person in the image. To cope with occlusions and avoid blind regions, the detectors are mainly focused on localizing the head and shoulders of people [1,3,8,14,29,31,37]. This has shown good results, especially for overhead and depth cameras.

Approaches can also be divided according to whether they are unsupervised or supervised. In both cases, background segmentation or modeling techniques such as Gaussian Mixture Model (GMM) are often exploited to facilitate the extraction of features and ease the detection process. People detection approaches based on unsupervised techniques have shown pretty good efficiency in real world applications [3,7,11,24,29,37]. In [3], for instance, a foreground segmentation module is used to extract head candidates in real time using low-level image processing operations such as edge and blob detection and connected components analysis. Similarly, a top-view detection system was proposed for depth images in [7]. This work uses dynamic background modeling to find objects of interest which are then filtered out via morphological operations to return bounding boxes around people. In [37] an unsupervised water filling algorithm is used to detect head candidates by finding local minimum regions in depth data. This method is accompanied of background modeling via GMM. Despite the good results of these methods, especially in terms of efficiency, they are heavily subject to the quality of background subtraction and the choice of ad hoc thresholds. In addition, these methods are unable to distinguish between humans and other dynamic objects in the scene (e.g baby carriage), leading to false alarms.

Supervised approaches for detecting people have also shown remarkable results. They normally require higher computational costs for both training and testing as well as a representative dataset with annotations for supervised learning. These approaches make use of machine learning algorithms, such as SVM or Boosting, to compute discriminative classifiers [14,25,31,33,34,39] In [25], for example, a top-view people detection system was proposed in challenging scenarios such as tailgating and piggybacking. This method first searches the depth map for local maxima to extract potential head candidates followed by a validation step using a SVM classifier. In general, this kind of methods yield high detection results but they are dependent on the choice of hand-crafted feature descriptors such as Histogram of Oriented Gradients (HOG). Furthermore, some initial steps such as floor detection [33], background modeling [31], and local maxima search in depth maps [25] are crucial to achieve high detection performance.

Recently, the use of deep networks has shown impressive results for people detection using color cameras [5,17,19,26]. Nevertheless, these methods were trained and focused mainly on detecting people from frontal and lateral views, thus showing a low performance for overhead cameras. By contrast, in [35] was presented a convolutional network called WatchNet for detecting people in top-view depth images. Since it only uses depth information its architecture is lighter in comparison to more complex networks [5,26], resulting in an efficient network that can be deployed in video surveillance applications in real time. WatchNet was trained with artificial and real depth data for people detection.

In this work, we propose a counting-by-detection approach for identifying attacks and intrusion in building entrances (airlocks) based on an improved version of WatchNet [35]. As novelties, WatchNet++ predicts body joints and links simultaneously to obtain higher detection rates, and it introduces a new prediction stage that provides features maps of higher spatial resolution to improve the localization accuracy of body joints. The network is also trained with abundant synthetic data in order to compute a pre-trained model that can be easily fine-tuned for different scenarios.

3 Depth Image Datasets

In this section we present the synthetic and real datasets used to train the network. While the first dataset is used to initialize the network model, the second one allows to adapt the network to the real scenario.

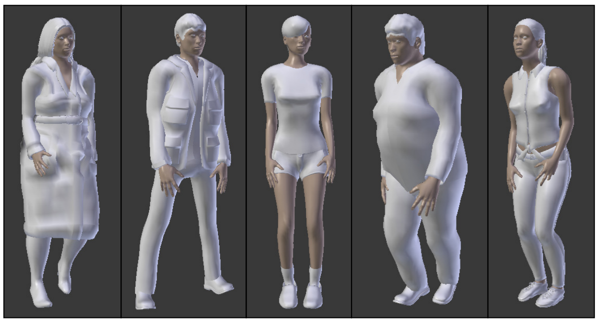

Fig. 3 Examples of 3D human models.

Fig. 4 Motion capture sequence for a 3D human model.

3.1 Synthetic Dataset

The supervised learning of deep network models requires to have at hand a large and diverse enough dataset to boost the network performance and prevent overfitting. Yet, the data is sometimes scarce for scenarios with task-based specifications. In addition, generating the images' annotations for supervised learning presents another inconvenient. This process is usually done manually and requires large amounts of human effort. An attractive alternative is to work with synthetic data [30, 32]. The benefits of this approach are twofold: 1) synthetic data can be generated automatically according to a given scenario for a specific problem, and 2) high quality annotations are generated at no cost.

Thus, to overcome the need for annotated training data, we present a systematic way to generate artificial depth images displaying people inside an airlock and the corresponding annotations (ground truth). We introduce a Synthetic Data Generator (SDG) built on Blender[1] to render people performing multiple behaviors inside a virtual airlock by motion simulation (see Figure 2). The airlock was designed following the specifications mentioned in the Unicity database [10]. Specifically, the airlock has an area of 2 × 2 meters and the camera is placed at the center of the airlock at two different heights: 2.1 and 2.5 meters.

A challenge in generating synthetic data is to introduce enough variability. We achieve this point by considering different body shapes and as many body pose configurations as possible. First, we use 24 3D human characters created with the modeling software Makehuman[2]. The different characters show variations in physical features, such as height and weight, and have been dressed with different clothing outfits to increase shape variation. Some examples are shown in Figure 3.

We add variability in body pose configurations by relying on the publicly available motion capture dataset from CMU labs[3]. We selected motion sequences of people performing diverse actions like walking or jumping. Figure 4 shows some snapshots for a mocap sequence.

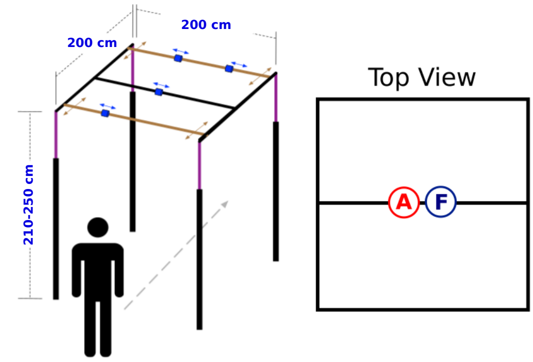

Fig. 5 Left: Schema of the recording structure (taken from [10]). Right: Top view of the recording structure depicting the position of the Argos and Fotonic sensors.

To synthesize depth images along with the required annotations, our SDG works as follows. At each iteration, we randomly select up to two 3D characters along with the corresponding number of mocap sequences, randomly selected. The 3D characters are randomly placed inside the airlock, in such a way that there is no collision between them. Subsequently, SDG samples one every 15 frames from the mocap sequence, performs motion retargeting and generates the corresponding synthetic depth image along with annotations. This is illustrated in Figure 2 (left). As a result, the synthetic database has more than 80k images containing up to two people, observe Figure 2 (right).

3.2 Real Dataset

Data. To fine-tune the network model to the real scenario, as well as evaluating its performance for people detection, we use the Unicity dataset[4] introduced in [10]. This dataset comprises several recorded sequences of people passing through a physical airlock giving access to a restricted area. The recording structure is schematized in Figure 5 (left). It is a 200 cm square airlock with two distinct heights: low and normal height

[1] http://www.blender.org
[2] http://www.makehuman.org/
[3] http://mocap.cs.cmu.edu/
[4] https://www.idiap.ch/dataset/unicity

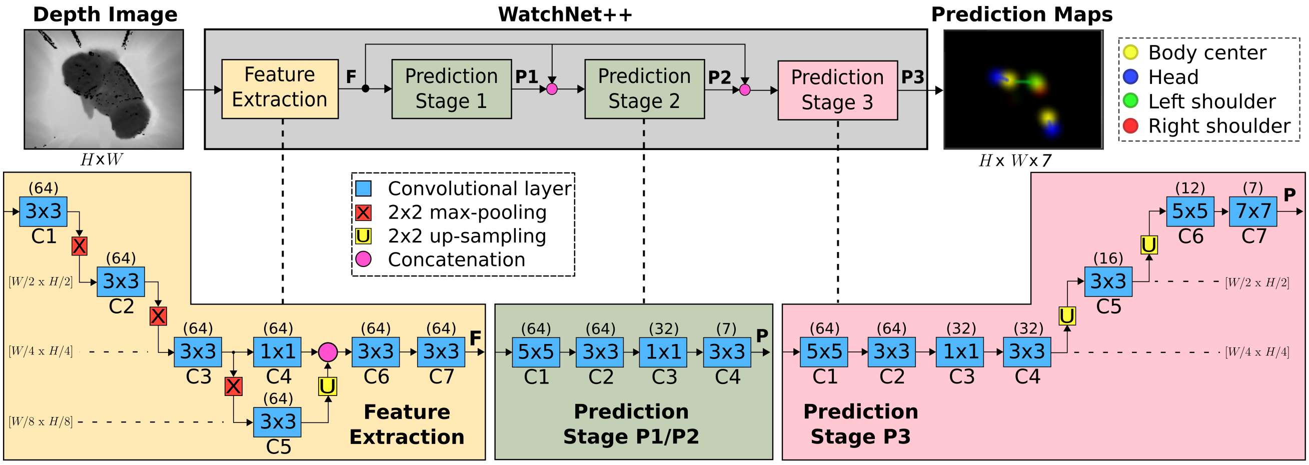

Fig. 6 General scheme of WatchNet++ for people detection. The network comprises a feature extraction module and a series of prediction stages that sequentially refine the prediction maps for human body joints and links as well as body centers.

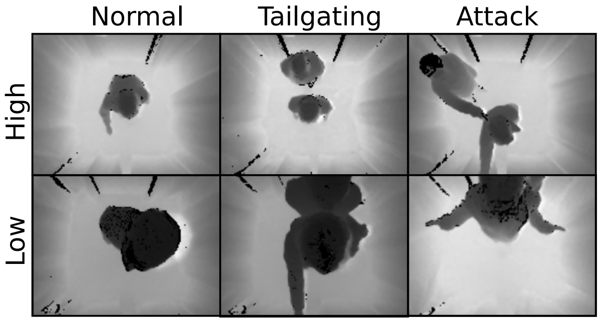

Fig. 7 Depth images of the recorded scenarios using the Argos sensor at two different heights (low and high).

(210 and 250 cm). The sequences were recorded using two industry-oriented depth sensors: Argos3D-P220 and Fotonic G-series. The position of these sensors is shown in Figure 5 (right). They are placed in the middle of the structure, next to each other, facing down.

Specifically, the Unicity dataset consists of 65 video sequences organized according to three different scenarios. The first one is a normal scenario with a single person walking and accessing the restricted area; the second scenario comprises two people trying to fool the surveillance system (e.g tailgating); in the third scenario, two people enter, and one of them attacks and forces the other to get into the restricted area. Figure 7 shows some example images of the recorded scenarios. For training and evaluation, the dataset was split into 33 and 32 sequences respectively, so that a participant does not appear in both sets. In total, the dataset has about 58k depth images for both sensors.

Annotations. Every frame in the dataset was annotated manually with the location of body joints (head and shoulders), the number of people in the airlock, and the degree of visibility of each person in order to evaluate the sensitivity of the detection system in accordance to people visibility. Five levels were defined: *Full:* the person is fully visible (body joints are visible); *Partial:* the person is partially visible and at least one body joint (head or shoulder) is visible; *Truncated:* a large portion of the person is visible but not any joint (e.g lower body); *Difficult:* similar to the truncated label but it only applies for a small portion of the person (e.g a leg or a hand). *Invisible:* the person is not visible in the airlock.

4 WatchNet++

In this section we describe the architecture and components of the proposed network for people detection. WatchNet++ is inspired by the Convolutional Pose Machines (CPM) for people pose estimation in color images [5]. However, our network is a lightweight and efficient version of CPM thanks to the use of depth data –instead of color– which allows reducing the number of convolutional layers and parameters since texture and color are removed. Besides, WatchNet++ includes other network characteristics like skip connections useful for multi-resolution analysis and upsampling operations for increased localization accuracy.

Similar to CPM, WatchNet++ can be thought of as comprising a feature extraction sub-network and a series of prediction stages that progressively refine the localization of human body joints and links in the image. Figure 6 shows a general view of the proposed network architecture. The different parts are described below.

4.1 Feature Extraction Sub-network

This sub-network computes discriminative features (F), from an input depth image of size $W \times H$, for body joints and links prediction that will be shared among the prediction stages. Since we use depth images as input, the complexity of this stage can be reduced compared to [5], and we can therefore deploy a smaller and more efficient feature extractor sub-network, contrary to the original CPM framework that relied on a very deep network to compute features (VGG-19 [28]).

We propose to use a sub-network composed of 7 convolutional layers (C) with filters' size of 3×3, three max-pooling operations (X), and one up-sampling operation (U), see Figure 6. All our convolutional layers use 64 filters to reduce the number of parameters in the network and speed up the forward pass.

A major design choice we follow and which differs from [5] is the use of skip connections [22] to combine features from different resolutions. Specifically, while the layer $C4$ with filters' size of 1×1 computes features at a quarter of the resolution of the input image, the convolutional layer $C5$ computes features at an eighth of the resolution. Then, features from $C5$ are upsampled and combined with $C4$ via concatenation. Finally, layers $C6$ and $C7$ compute the output features F.

The main reason for this particular configuration is to increase the robustness and accuracy of the network to detect people at multiple scales. This is an important aspect in video surveillance systems since the height at which the camera is located varies depending on the room, and depth measures have a different semantic nature than color images. This is the case of the Unicity dataset whose depth images were recorded at two different sensor's heights.

4.2 Prediction Sub-networks

The proposed network has a series of prediction sub-networks where each one provides a set of feature maps (seven maps) encoding the location of the body center, body joints (head and shoulders) and the prediction of body links (coupling the body center and joints). The prediction stages are applied sequentially in combination with the extracted features (F) in order to refine the preceding predictions (see Figure 6). This results in enhanced prediction maps and better detection scores.

Unlike [5,35], whose prediction maps are computed at the same resolution, in this work we propose two types of prediction sub-networks. The first one is an efficient sub-network introduced in [35] that is composed by four convolutional layers with filters of different sizes, keeping low the numbers of filters. This sub-network is computed particularly for the prediction stages $P1$ and $P2$, see Figure 6. The first convolutional layer ($C1$) has filters of size of 5×5 in order to capture larger image spatial context and to encode the spatial relationships among the body joints and links. This spatial/feature co-occurrence has been shown to play an important role to refine the network output predictions. The final layer ($C4$) provides prediction maps of size of $W/4 \times H/4 \times 7$.

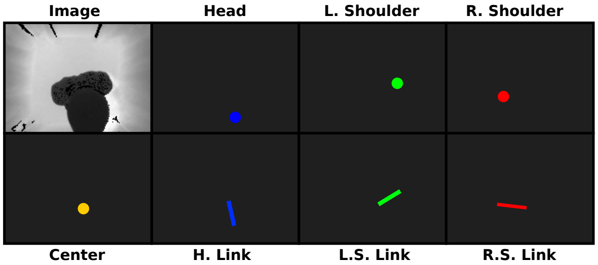

Fig. 8 Ground-truth annotation masks for an image. Body center, joints and links represented in color for better clarity.

The second prediction sub-network is computed for the last prediction stage ($P3$). It has a similar architecture as P1 and P2, but with the novelty that there are three additional convolutional layers ($C5$ - $C7$) and two up-sampling operations with the aim of providing feature maps to higher spatial resolution ($W \times H \times 7$). This allows to compute enhanced feature maps ($C7$) and thus obtaining more accurate results for the localization of body joints and links in the image. The use of up-sampling layers has shown pretty good results in the past for accurate image segmentation [22,27].

Another difference with [35], which only uses body joints, is that we use and predict body joints and links together to encode pairwise relationships between them to obtain better predictions about the body center used for people detection (see Fig. 8). This idea comes from [5], but instead of having two network branches which results in computational overhead, we use a single branch to predict body joints and links, keeping efficiency and reducing the number of network parameters.

4.3 Training Loss and Ground Truth Masks

The training loss for WatchNet is calculated as a linear combination of partial losses across the network. We define the global loss by $L = \frac{1}{N} \sum_{i=1}^{N} L_i$, where N is the number of prediction stages and L_i is the loss for the prediction stage P_i. Specifically, the partial loss for a prediction stage i is defined as the mean squared distance between the prediction maps provided by stage i and the ground-truth annotation masks. For every training image the ground truth is a stack of seven annotations masks encoding the true locations of the body joints, the body center, and the links between them. Figure 8 illustrates the ground truth for a training image. These masks are computed online during training. Gaussian blobs are placed to encode the position of the joints and the body center, while line segments are added to encode the links between joints and the center. This center is computed online as the mean point among joints (head and shoulders).

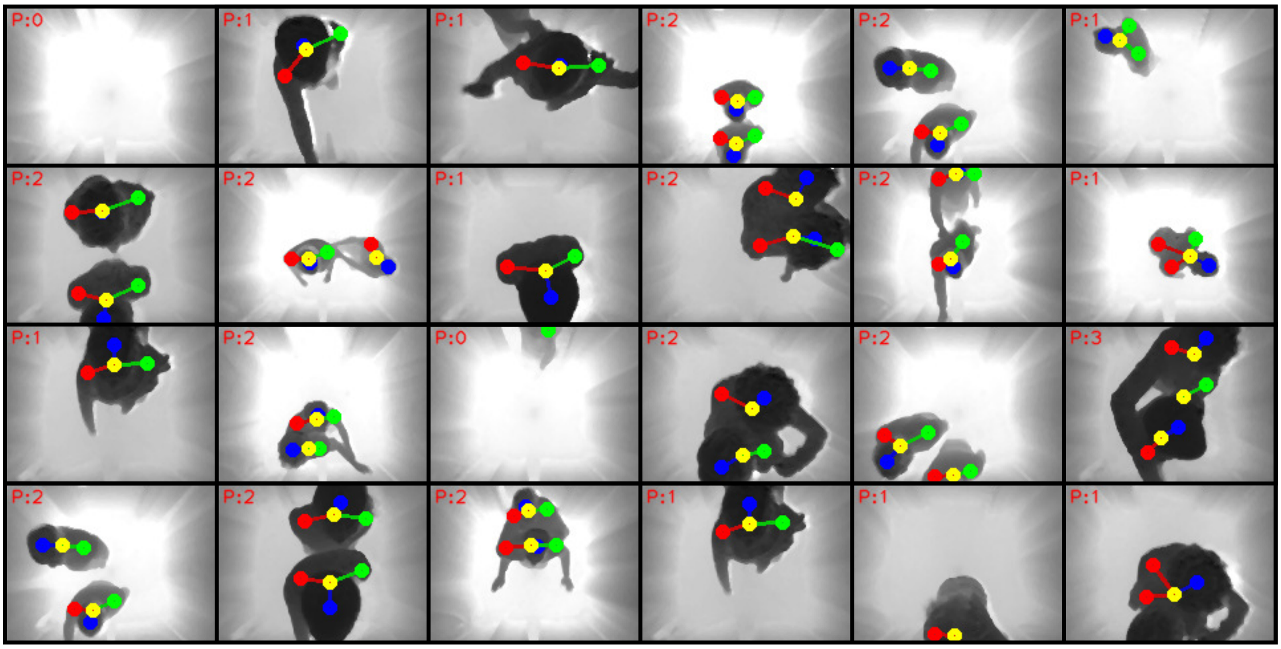

Fig. 9 Some example images with the output of WatchNet++ for people detection in depth images. The network predicts the location of body joints and links as well as the body center, all depicted by blue, green, red and yellow spots respectively. The system also estimates the number of people (P) inside the airlock based on counting the number of body centers.

4.4 General Settings

All our convolutional layers are computed in combination with batch normalization [15] and Rectified Linear Units (ReLU), showing good experimental results and faster training. Note that WatchNet++ is a fully-convolutional network involving only convolution layers. This reduces the number of parameters and enables the network to be independent of the size of the input image [22]. The network weights are initialized using the Xavier's initialization [13]. We use Adam [16] as optimizer with default settings.

4.5 Counting People in Images

At test time, our learned WatchNet++ is applied to each image to compute predictions. We remove predictions whose confidence level is below a threshold β. The choice of β is done accordingly to the user needs (e.g high recall vs high precision).

We use the number of predicted body centers to count the number of people inside the airlock. This choice has shown to be robust in cases when other body joints (e.g head) lie outside the scene [35].

5 Experiments

This section evaluates WatchNet++ for the task of counting people and detecting attacks in building access rooms. Fig. 9 shows some example images.

5.1 Datasets

In this paper we use three datasets to train and evaluate the network for people detection in depth images.

Synthetic Dataset: We use the synthetic dataset described in Sec. 3.1 to pre-train the network with a large number of artificial depth images (80k images) containing up to two people performing different actions.

Unicity Dataset: To adapt the network to real depth images, we fine-tune the network with the Unicity dataset described in Sec. 3.2 and introduced originally in [10]. This dataset is also used to evaluate the network to detect attacks and intruders in airlocks in which only one person is allowed. Images with two people in the room are considered as alarms (positive samples). The dataset has 33 video sequences for training and 32 sequences for testing. Each sequence was recorded by two different depth sensors: Argos and Fotonic.

MIVIA Dataset: This dataset was proposed in [9]. It contains 17 videos sequences of multiple people walking in indoor and outdoor corridors. The sequences were acquired using an overhead depth sensor (Kinect) at fixed height. The goal in this dataset is to count people crossing a virtual line in different directions. This dataset is used to test WatchNet++ in another scenario where the task of detecting people is crucial.

5.2 Evaluation Protocol

We follow the evaluation protocol presented in [10] to measure the performance of the network for detecting attacks and intrusion (alarms). We compute the recall (R), precision (P) and F-measure (F) rates. For evaluation the dataset considers four levels of difficulty defined according to the degree of visibility of people inside the airlock (see Sec. 3.2). Level 1 comprises all images where people's joints are full visible (head and shoulders). In level 2 at least one body joint is visible (e.g a shoulder is visible). Level 1 is thus a subset of level 2. Similarly, level 3 contains level 2 plus all those images where a portion of people is visible, but not their joints. Finally, level 4 is all the images in the test set including difficult cases (e.g a leg is visible only). For all the levels, we include images of the empty airlock to count for negative samples.

To measure the localization accuracy of the body joints in the image, we use the Percentage of Correct Keypoints (PCK) metric [36]. It works as follows: a detected joint is considered correct if the Euclidean distance between the predicted and the true joint is within a certain threshold (radius). Here, we consider radius of 1, 3, 5, and 10 pixels. For every threshold case we compute the recall, the precision and the F-measure rates. Rates at lower thresholds (e.g 1 and 3 pixels) indicate more accurate localization results.

We repeated the training and testing phases five times in order to consider randomness in the network computation. Average detection rates are reported.

5.3 Default Settings

Unless otherwise stated, WatchNet++ is trained with the synthetic dataset for 50k iterations and is fine-tuned with the real training data for 5k iterations. We use three prediction stages in the network and a batch of five samples for training. The network is trained and tested using the Argos depth data (17k training images and 11k testing images). Image resolution is set to 120×160 pixels and the depth image values are normalized in the range between 0 and 1 by dividing the depth values by a distance of 3000 millimeters. To remove noise from depth maps, we resort to inpainting with a filter size of 5. To select β, we run the network in the training set for varying detection thresholds and take the one that achieves the highest F-measure score.

5.4 Network Architecture

In Table 1 we compare the proposed WatchNet++ with its preceding version (WatchNet [35]) according to the two introduced innovations: the use of body links and prediction maps of higher spatial resolution. The comparison is done in terms of the localization accuracy of the body joints (head and shoulders) for varying radius and the rates for detecting alarms (two people in the room) for the different evaluation levels. We see that WatchNet++ obtains better detections rates, particularly for recall which corresponds to detect true alarms. On the other hand, WatchNet attains high precision, but its accuracy for predicting the body joints is low, especially for low distance thresholds (radius of 1 and 3 pixels). This is because the prediction maps provided by the network have a low resolution of 30x40 pixels, while WatchNet++ returns maps with the same resolution of the input image (120x160 pixels). This is achieved by incorporating up-sampling and convolutional operations in the last prediction stage (see Fig. 6). The result is therefore more accurate prediction maps encoding the position of the body joints.

Alarm Detection Rates				**Localization Accuracy**			
Level	R	P	F	Rad.	R	P	F
WatchNet				**WatchNet**			
L-1	96.5	99.6	98.1	R-1	13.7	14.0	14.4
L-2	94.2	**99.8**	96.9	R-3	30.6	31.9	31.2
L-3	81.5	**99.7**	89.7	R-5	64.3	66.9	65.5
L-4	61.3	**99.7**	75.9	R-10	85.3	88.8	87.0
WatchNet++ [w/o links]				**WatchNet++ [w/o links]**			
L-1	96.6	99.9	98.2	R-1	33.3	33.5	33.4
L-2	93.9	99.2	96.5	R-3	53.1	53.3	53.2
L-3	82.6	98.7	89.9	R-5	**75.6**	76.0	75.8
L-4	62.2	98.8	76.4	R-10	**88.5**	88.9	88.7
WatchNet++				**WatchNet++**			
L-1	**97.4**	**100**	**98.7**	R-1	**34.0**	**34.6**	**34.3**
L-2	**95.5**	99.5	**97.5**	R-3	**53.4**	**54.8**	**54.1**
L-3	**82.9**	99.3	**90.3**	R-5	75.3	**77.2**	**76.3**
L-4	**62.3**	99.3	**76.6**	R-10	88.0	**90.3**	**89.1**

Table 1 Left: Alarm detection rates provided by the networks in the Unicity database. The evaluation is done using the recall (R), precision (P) and F-measure (F). Rigth: Localization accuracy of body joints (head and shoulders).

Besides, the table reports the scores for the four evaluation levels mentioned above. Note that the proposed network achieves almost perfect rates for levels 1 and 2 which contain at least one visible body joint. The scores degrade for levels 3 and 4 since people are not fully visible. Figure 9 shows some example images with the output of WatchNet++. The last column has failure examples with wrong body joint predictions (e.g two detected left shoulders for a person) and a case where three body centers are detected.

Table 1 also shows the impact of using body links in WatchNet++. Note that the network without using links obtains lower rates. This proofs that the links are complementary features that contribute to improve the prediction of body joints and its center.

5.5 Training Data

The detection performance evaluation of WatchNet++ according to the size of the training data is shown in Table 2.The network was trained with different training subsets of the Unicity dataset. Synthetic data is discarded in this experiment. We see that the rates increases as the number of training images gets larger.

Alarm Detection Rates				Localization Accuracy			
Level	R	P	F	Rad.	R	P	F
Real Dataset[1k images]				Real Dataset[1k images]			
L-1	83.2	96.9	89.5	R-1	25.3	27.0	26.1
L-2	86.3	95.9	90.8	R-3	40.7	42.8	41.7
L-3	76.0	94.2	84.1	R-5	63.2	66.5	64.8
L-4	59.6	94.6	73.1	R-10	80.3	84.5	82.4
Real Dataset [3k images]				Real Dataset [3k images]			
L-1	88.0	98.3	92.9	R-1	27.0	29.7	28.3
L-2	88.5	**98.2**	93.1	R-3	43.9	47.7	45.7
L-3	76.6	**98.2**	86.0	R-5	65.8	71.5	68.5
L-4	58.2	**98.2**	73.1	R-10	80.6	87.6	83.9
Real Dataset [10k images]				Real Dataset [10k images]			
L-1	89.0	98.2	93.4	R-1	29.0	31.3	31.1
L-2	90.7	97.3	93.9	R-3	45.6	48.3	46.9
L-3	79.6	96.7	87.3	R-5	68.0	72.1	70.0
L-4	61.2	96.9	75.0	R-10	82.8	**87.8**	85.2
Real Dataset [17k images]				Real Dataset [17k images]			
L-1	**92.3**	**98.9**	**95.5**	R-1	**30.2**	**32.6**	**31.4**
L-2	**92.5**	**98.2**	**95.3**	R-3	**46.7**	**49.3**	**48.0**
L-3	**81.1**	97.6	**88.6**	R-5	**68.9**	**72.6**	**70.7**
L-4	**61.8**	97.7	**75.7**	R-10	**83.2**	**87.8**	**85.4**

Table 2 Alarm detection rates and body joint localization accuracy of WatchNet++ according to the size of the real training data (Unicity dataset) and without using synthetic data for pretraining the network.

Alarm Detection Rates				Localization Accuracy			
	R	P	F		R	P	F
Real Dataset				Real Dataset			
L-1	92.3	98.9	95.5	R-1	30.2	32.6	31.4
L-2	92.5	98.2	95.3	R-3	46.7	49.3	48.0
L-3	81.1	97.6	88.6	R-5	68.9	72.6	70.7
L-4	61.8	97.7	75.7	R-10	83.2	87.8	85.4
Synthetic Dataset				Synthetic Dataset			
L-1	79.9	98.8	88.3	R-1	15.0	20.4	17.3
L-2	77.5	96.2	85.8	R-3	25.8	31.8	28.5
L-3	61.3	94.5	74.4	R-5	41.1	50.7	45.4
L-4	47.8	94.8	63.6	R-10	55.5	68.5	61.3
Synthetic + Real Datasets				Synthetic + Real Datasets			
L-1	**97.4**	**100.0**	**98.7**	R-1	**34.0**	**34.6**	**34.3**
L-2	**95.5**	**99.5**	**97.5**	R-3	**53.4**	**54.8**	**54.1**
L-3	**82.9**	**99.3**	**90.3**	R-5	**75.3**	**77.2**	**76.3**
L-4	**62.3**	**99.3**	**76.6**	R-10	**88.0**	**90.3**	**89.1**

Table 3 Detection performance of WatchNet++ in terms of the use of synthetic and real data for training.

However, the best achieved performance is limited because the relative small number of images for training the network (17k images).

To enlarge the training data, we resort to the synthetic dataset which contains depth images with people inside a virtual airlock. Table 3 reports the rates for training the network using only synthetic data (80k images). We see that the detection rates and accuracy are substantially inferior to those obtained by the network using the real dataset. This is because there is a gap between the real and synthetic data domains caused by sensor noise and variations in the simulated scenario.

To overcome this problem WatchNet++ is initially trained with the synthetic data for 50k iterations and then fine-tuned (for 5k iterations) to adapt the network model to the real domain. The table shows the rates after fine-tuning the network. Note that the use of synthetic and real images significantly improves the results, especially the recall rate that corresponds to the detection of attacks and intrusion.

Alarm Detection Rates				Localization Accuracy			
Level	R	P	F	Rad.	R	P	F
1 Prediction Stage				1 Prediction Stage			
L-1	96.1	99.5	97.8	R-1	33.3	33.8	33.6
L-2	92.6	99.6	96.0	R-3	52.2	53.5	52.9
L-3	81.9	99.5	89.8	R-5	74.2	76.0	75.1
L-4	61.5	99.5	76.1	R-10	86.8	88.9	87.8
3 Prediction Stages				3 Prediction Stages			
L-1	97.4	**100.0**	98.7	R-1	34.0	**34.6**	**34.3**
L-2	95.5	99.5	97.5	R-3	53.4	54.8	54.1
L-3	82.9	99.3	90.3	R-5	75.3	77.2	76.3
L-4	62.3	99.3	76.6	R-10	88.0	**90.3**	**89.1**
5 Prediction Stages				5 Prediction Stages			
L-1	**97.7**	**100.0**	**98.8**	R-1	**34.2**	34.2	34.2
L-2	**95.6**	**99.8**	**97.7**	R-3	**54.1**	**54.9**	**54.5**
L-3	**85.3**	**99.8**	**92.0**	R-5	**76.2**	**77.4**	**76.8**
L-4	**64.7**	**99.8**	**78.5**	R-10	**88.1**	89.5	88.8

Table 4 Network performance according to the number of prediction stages.

Alarm Detection Rates				Localization Accuracy			
Level	R	P	F	Rad.	R	P	F
Baseline				Baseline			
L-1	97.0	55.0	70.0	R-3	-	-	-
L-2	**96.0**	74.0	84.0	R-5	-	-	-
L-3	**88.0**	79.0	83.0	R-10	-	-	-
L-4	**72.0**	81.0	76.0	R-15	-	-	-
FCN				FCN			
L-1	92.7	99.0	95.7	R-1	15.4	15.5	15.5
L-2	85.3	98.5	91.4	R-3	32.3	34.9	33.6
L-3	71.2	98.4	82.6	R-5	63.5	68.7	66.0
L-4	53.9	98.4	69.7	R-10	82.8	89.5	86.0
UNet				UNet			
L-1	96.6	99.6	98.1	R-1	32.0	32.3	32.1
L-2	93.7	**99.7**	96.6	R-3	51.5	51.8	51.6
L-3	81.8	**99.5**	89.8	R-5	74.6	75.2	74.9
L-4	61.5	**99.5**	76.0	R-10	**88.7**	89.3	89.0
WatchNet++				WatchNet++			
L-1	**97.4**	**100**	**98.7**	R-1	**34.0**	**34.6**	**34.3**
L-2	95.5	99.5	**97.5**	R-3	**53.4**	**54.8**	**54.1**
L-3	82.9	99.3	**90.3**	R-5	**75.3**	**77.2**	**76.3**
L-4	62.3	99.3	**76.6**	R-10	88.0	**90.3**	**89.1**

Table 5 Comparison of WatchNet++ against a baseline method and two popular network architectures.

5.6 Prediction Stages

The performance of WatchNet++ in terms of the number of prediction stages is shown in Table 4. The larger the number of prediction stages, the higher the detection scores and accuracy. Using five stages the network obtains the best results. However, it is at the expense of a larger number of model parameters. In this work, we use three stages as a compromise between detection performance and the network model size.

5.7 Detection Approaches

The proposed WatchNet++ is compared against other approaches in Table 5. The first approach is the baseline provided with the Unicity dataset [10]. This approach is based on background subtraction which thresholds the estimated volume inside the airlock: when the volume is larger than a predefined threshold, the method classifies

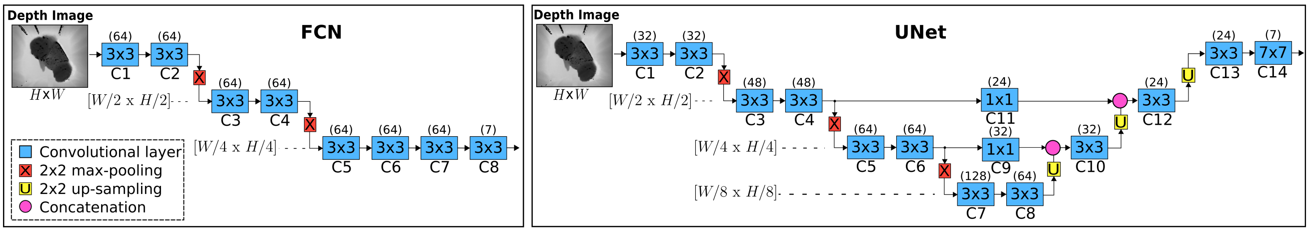

Fig. 10 Implemented network architectures. Left: Fully-Convolutional Network (FCN). Right: U-shape Network (UNet).

the depth image as an alarm. The volume is estimated by simply summing up all the pixels of $B - I$, where B is the depth map of the empty airlock, and I the current depth image. The second approach is a Fully-Convolutional Network (FCN) commonly used for image recognition tasks [18,28]. In this work, the network consists of 7 convolutional layers, two max-pooling operations, and a final convolutional layer for predicting the body joints and links, see Figure 10 (left). Similar to WatchNet++, this network uses 64 filters per layer and a filter size of 3×3 to keep efficiency.

The third approach is an U-shape network (UNet) which is a popular network architecture for semantic segmentation [21,22,27]. UNet comprises down- and up-sampling operations to extract and combine shallow and deep features at multiple resolutions. The output is also a set of prediction maps with the same resolution of the input image. Figure 10 (right) shows the implemented UNet using 14 convolutional layers, three max-pooling and three up-sampling operations, and two skip connections ($C9$ and $C11$) using a filter size of 1×1. To maintain efficiency the number of convolutional filters varies according to the resolution level.

Looking at Table 5, we see that the baseline method achieves very high recall scores for all evaluation levels, but it obtains low precision due to high rates of false positives. Since this method uses background subtraction, it does not provide the localization of people (body joints) in the image. On the other hand, FCN does detect people but obtains lower results than the other networks because FCN does not use the prediction refinement to increase the alarm detection rates, and because FCN yields prediction maps at low resolution (up-sampling is not applied) what results in low localization accuracy. Conversely, UNet returns more accurate prediction maps and thus higher localization rates, showing that up-sampling features is beneficial for people localization. WatchNet++ obtains better scores than UNet and achieves the best alarm detection rates and localization accuracy. This is a consequence of using several predictions stages to enhance the prediction maps.

Alarm Detection Rates				Localization Accuracy			
Level	R	P	F	Rad.	R	P	F
WatchNet				WatchNet			
L-1	97.2	**95.4**	96.3	R-1	16.2	17.1	16.7
L-2	93.1	**94.1**	93.6	R-3	34.2	31.8	32.9
L-3	87.0	92.7	89.8	R-5	66.7	62.2	64.4
L-4	72.1	92.9	81.2	R-10	87.2	81.3	84.1
WatchNet++				WatchNet++			
L-1	**98.3**	94.9	**96.6**	R-1	**33.9**	**36.6**	**35.2**
L-2	**93.8**	93.9	**93.9**	R-3	**44.4**	**41.0**	**42.6**
L-3	**88.3**	**93.1**	**90.6**	R-5	**75.9**	**70.0**	**72.8**
L-4	**73.7**	**93.3**	**82.3**	R-10	**88.8**	**81.9**	**85.2**

Table 6 Performance comparison using the Fotonic sensor.

5.8 Fotonic sensor

WatchNet++ is also compared against WatchNet using the Fotonic depth sensor. The results are shown in Table 6. We see again that the proposed network outperforms the original version both in alarm detection scores and the accuracy for localizing the body joints.

5.9 Counting People

Unlike previous experiments testing the network to detect people attacks and intrusion, WatchNet++ is tested here to measure the flow of people in indoor and outdoor corridors. In particular, the task consists on counting the number of persons crossing a virtual line placed in the scene. This is illustrated in Figure 11 that shows the output of WatchNet++ for detecting people and tracking them in two video sequences of the MIVIA dataset [9]. The second and fourth rows depict the prediction of body joints and links as well as the body centers and the number of estimated people in the scene. The first and third rows show the trajectories of people crossing the virtual line (indicated by a black line). The current number of persons who have gone through the scene is shown at top-left of the frame.

For this task, WatchNet++ is run at every video frame to detect the body joints and its center. In parallel, a simple tracker is executed to track every person in the scene. This tracker is based on matching new detected body centers with the current set of tracklets using the minimum distance. A person passage is detected when the tracklet intercepts the virtual line. For this approach, it is essential that WatchNet++ detects the persons in most of the frames (i.e high recall rate) in

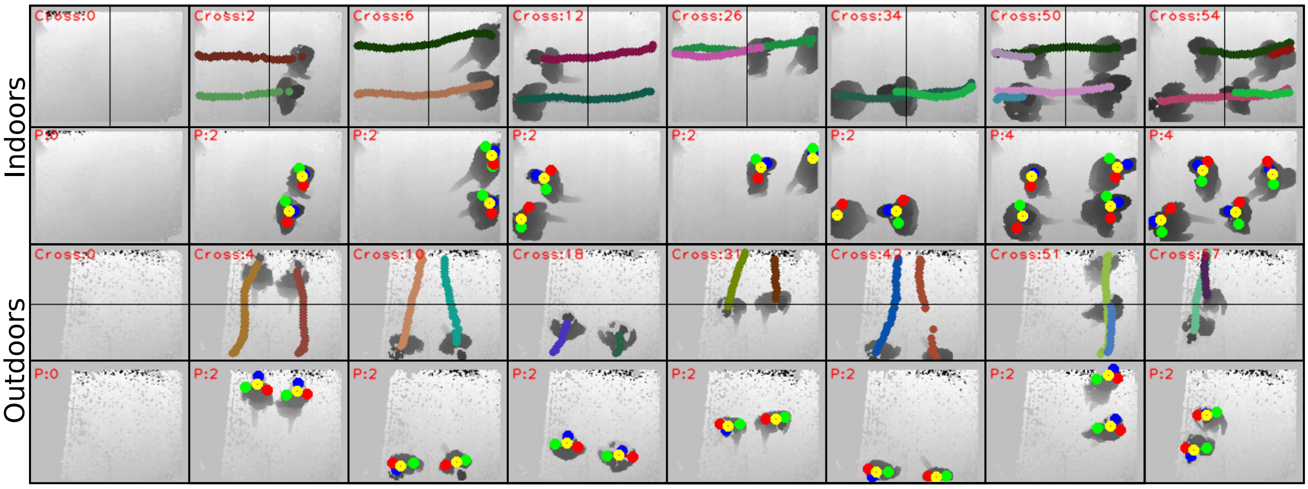

Fig. 11 People counting results using WatchNet++ for indoor and outdoor scenarios in the MIVIA dataset.

People Counting Rates													
		Indoors						Outdoors					
Method	Train data	TP	FN	FP	R	P	F	TP	FN	FP	R	P	F
Method [9]	–	408	7	0	98.3	100	99.1	520	33	0	94.0	100	96.9
WatchNet++	Synthetic	409	6	0	98.6	100	99.3	435	118	0	78.7	100	88.1
WatchNet++	Synth.+Unicity	408	7	0	98.3	100	99.1	493	60	0	89.2	100	94.3
WatchNet++	Synth.+MIVIA	377	0	0	100	100	100	487	5	0	99.0	100	99.5

Table 7 People counting performance in MIVIA dataset for indoor and outdoor scenarios.

order to have continuous tracklets. To consider missed detections, we use an elapsed period of 11 frames for tracklet recovery. Otherwise, such tracklet is deleted.

For evaluation, we use the MIVIA dataset [9] that has 8 indoor videos and 9 outdoor videos with 415 and 553 people passages respectively. The performance is measured using true positives (TP), false negatives (FN), false positives (FP), recall (R), precision (P) and F-measure (F) rates. TP corresponds to the transits of persons that are correctly detected by the method, FP is the number of falsely detected passages of persons, and FN is the number the passages of persons missed by the method.

Table 7 reports the performances for the method presented in [9] and WatchNet++ using different training data. In [9], a simple approach was proposed which does not detect and track people in the video sequences. It consists of two steps: foreground detection via background subtraction and people counting based on a cell grid that interprets the results of the foreground detection step. This method is very efficient and yields high performance rates in both scenarios. However, it depends of the quality of the foreground detection results and some ad-hoc parameters. By contrast, WatchNet++ does detect people everywhere in the depth image and counts the number of people passages using the tracking system. We see that when the network is trained with synthetic data only, the performance is high in indoors (F-measure of 99.3%) but low in outdoors (88.1%). This is caused mainly to the sensor noise that in outdoors degrades largely the quality of depth data. If the network is then fine-tuned with real depth images from the Unicity dataset [10], the performance in outdoors achieves an F-measure of 94.1%, reducing by half the number of false negatives.

Better results are obtained if WatchNet++ is fine-tuned with data from the MIVIA dataset. To this end, the dataset is split into a small training set consisting of two sequences and a testing set having the remaining 15 sequences. For training, we chose the sequences *D_I_S_1* and *D_O_S_1* which correspond to an indoor and an outdoor sequence having a single person per frame. About 2600 depth images were manually annotated with the position of heads and shoulders. The network was fine-tuned for 2k iterations. Observe that the WatchNet++ attains almost perfect recognition rates in both scenarios since the network is well adapted to the environment conditions and sensor noise. Note also that the true positive values are lower than previous results since the dataset was split to fine tune the network, resulting in a smaller number of people crossing the corridor in the test set.

Some example results are shown in Figure 11 in which we see that WatchNet++ is able to localize people in every image and track them during the video sequences. Some failure cases are depicted in Figure 12. In the first and second columns, the network does not detect a person for some frames and the associated track-

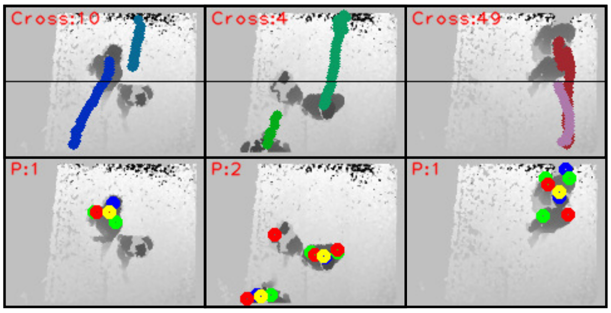

Fig. 12 Some failure cases in the people counting approach.

let is thus deleted. It is due to undefined depth values around shoulders and back (sensor noise). In the third column the problems comes from occlusion of the person. The network predicts then just one body center.

5.10 Efficiency

The proposed network has similar real-time performance as its predecessor [35] for detecting people in depth images since both networks share the same feature extraction module and similar prediction modules. Specifically, WatchNet++ runs in 10 and 28 FPS using CPU and GPU cards respectively.

6 Conclusion

In this work we presented a video surveillance system which is able to detect people attacks and intrusion in access rooms as well as counting people in corridors for motion analysis. The proposed approach consists of a deep network, called WatchNet++, which is an improved version of WatchNet [35]. Our system demonstrated very good results in the Unicity [10] and MIVIA [9] datasets created specifically for the above problems. WatchNet++ also showed superior performance to other network architectures and approaches based on foreground detection. The use of synthetic and real data demonstrated to be beneficial to enlarge the training data and obtain better detection results. This work also proved that using body links, up-sampling convolutional operations, and a cascade of prediction stages the proposed network returns more accurate predictions maps encoding the location of body joints.

Acknowledgments: The work was supported by Innosuisse, the Swiss innovation agency, through the UNICITY (3D scene understanding through machine learning to secure entrance zones) project.

References

1. Ahmad, M., Ahmed, I., Ullah, K., khan, I., Khattak, A., Adnan, A.: Person detection from overhead view: A survey. International Journal of Advanced Computer Science and Applications **10**(4) (2019) 2, 3
2. Ahmed, I., Adnan, A.: A robust algorithm for detecting people in overhead views. Cluster computing **21**(1), 633–654 (2018) 2
3. Bondi, E., Seidenari, L., Bagdanov, A.D., Del Bimbo, A.: Real-time people counting from depth imagery of crowded environments. In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 337–342. IEEE (2014) 2, 3
4. Boominathan, L., Kruthiventi, S.S., Babu, R.V.: Crowdnet: A deep convolutional network for dense crowd counting. In: Proceedings of the 2016 ACM on Multimedia Conference (2016) 1, 3
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017) 3, 5, 6
6. Carincotte, C., Naturel, X., Hick, M., Odobez, J.M., Yao, J., Bastide, A., Corbucci, B.: Understanding metro station usage using closed circuit television cameras analysis. In: ITSC (2008) 1
7. Carletti, V., Del Pizzo, L., Percannella, G., Vento, M.: An efficient and effective method for people detection from top-view depth cameras. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017) 2, 3
8. Chen, S., Bremond, F., Nguyen, H., Thomas, H.: Exploring depth information for head detection with depth images. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 228–234. IEEE (2016) 2, 3
9. Del Pizzo, L., Foggia, P., Greco, A., Percannella, G., Vento, M.: A versatile and effective method for counting people on either rgb or depth overhead cameras. In: 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6. IEEE (2015) 7, 10, 11, 12
10. Dumoulin, J., Canévet, O., Villamizar, M., Nunes, H., Khaled, O.A., Mugellini, E., Moscheni, F., Odobez, J.M.: Unicity: A depth maps database for people detection in security airlocks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2018) 2, 4, 7, 9, 11, 12
11. Galčík, F., Gargalík, R.: Real-time depth map based people counting. In: International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 330–341. Springer (2013) 2, 3
12. Garrell, A., Villamizar, M., Moreno-Noguer, F., Sanfeliu, A.: Teaching robot's proactive behavior using human assistance. International Journal of Social Robotics **9**(2), 231–249 (2017) 1
13. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256 (2010) 7
14. Hu, R., Wang, R., Shan, S., Chen, X.: Robust head-shoulder detection using a two-stage cascade framework. In: 2014 22nd International Conference on Pattern Recognition, pp. 2796–2801. IEEE (2014) 3
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR **abs/1502.03167** (2015) 7
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014) 7
17. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, pp. 11977–11986 (2019) 3

18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012) 10
19. Lejbolle, A.R., Krogh, B., Nasrollahi, K., Moeslund, T.B.: Attention in multimodal neural networks for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 179–187 (2018) 2, 3
20. Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: Advances in neural information processing systems, pp. 1324–1332 (2010) 3
21. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018) 10
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440 (2015) 6, 7, 10
23. Ma, Z., Chan, A.B.: Crossing the line: Crowd counting by integer programming with local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2539–2546 (2013) 3
24. Nalepa, J., Szymanek, J., Kawulok, M.: Real-time people counting from depth images. In: International Conference: Beyond Databases, Architectures and Structures (2015) 3
25. Rauter, M.: Reliable human detection and tracking in top-view depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 529–534 (2013) 1, 2, 3
26. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263–7271 (2017) 3
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer (2015) 6, 10
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 6, 10
29. Song, H., Sun, S., Akhtar, N., Zhang, C., Li, J., Mian, A.: Benchmark data and method for real-time people counting in cluttered scenes using depth sensors. arXiv preprint arXiv:1804.04339 (2018) 1, 3
30. Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 969–977 (2018) 4
31. Tu, J., Zhang, C., Hao, P.: Robust real-time attention-based head-shoulder detection for video surveillance. In: 2013 IEEE International Conference on Image Processing, pp. 3340–3344. IEEE (2013) 1, 3
32. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 109–117 (2017) 4
33. Vera, P., Zenteno, D., Salas, J.: Counting pedestrians in bidirectional scenarios using zenithal depth images. In: Mexican Conference on Pattern Recognition (2013) 3
34. Villamizar, M., Andrade-Cetto, J., Sanfeliu, A., Moreno-Noguer, F.: Boosted random ferns for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 272–288 (2018) 3
35. Villamizar, M., Martínez-González, A., Canévet, O., Odobez, J.M.: Watchnet: Efficient and depth-based network for people detection in video surveillance systems. In: IEEE International Conference on Advanced Video and Signal-based Surveillance (2018) 2, 3, 6, 7, 8, 12
36. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE transactions on pattern analysis and machine intelligence **35**(12), 2878–2890 (2012) 8
37. Zhang, X., Yan, J., Feng, S., Lei, Z., Yi, D., Li, S.Z.: Water filling: Unsupervised people counting via vertical kinect sensor. In: 2012 IEEE ninth international conference on advanced video and signal-based surveillance, pp. 215–220. IEEE (2012) 2, 3
38. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 589–597 (2016) 3
39. Zhu, L., Wong, K.H.: Human tracking and counting using the kinect range sensor based on adaboost and kalman filter. In: International Symposium on Visual Computing (2013) 3