IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 22, 2020, accepted April 2, 2020, date of publication April 6, 2020, date of current version April 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985985

Open-World Person Re-Identification With RGBD Camera in Top-View Configuration for Retail Applications

MASSIMO MARTINI, MARINA PAOLANTI, (Member, IEEE), AND EMANUELE FRONTONI
Dipartimento di Ingegneria dell'Informazione, 60131 Ancona, Italy
Corresponding author: Marina Paolanti (m.paolanti@univpm.it)

This work was supported by the Grottini Laboratory.

ABSTRACT Person re-identification (re-ID) is currently a notably topic in the computer vision and pattern recognition communities. However, most of the existing works on re-ID have been designed for closed world scenarios, rather than more realistic open world scenarios, limiting the practical application of these re-ID techniques. In a common real-world application, a watch-list of known people is given as the gallery/target set for searching through a large volume of videos where the people on the watch-list are likely to return. This aspect is fundamental in retail for understanding how customers schedule their shopping. The identification of regular and occasional customers allows to define temporal purchasing profiles, which can put in correlation the customers' temporal habits with other information such as the amount of expenditure and number of purchased items. This paper presents the first attempt to solve a more realistic re-ID setting, designed to face these important issues called Top-View Open-World (TVOW) person re-id. The approach is based on a pretrained Deep Convolutional neural Network (DCNN), finetuned on a dataset acquired by using a top-view configuration. A special loss function called triplet loss was used to train the network. The triplet loss optimizes the embedding space such that data points with the same identity are closer to each other than those with different identities. The TVOW is evaluated on the TVPR2 dataset for people re-ID that is publicly available. The experimental results show that the proposed methods significantly outperform all competitive state-of-the-art methods, bringing to different and significative insights for implicit and extensive shopper behaviour analysis for marketing applications.

INDEX TERMS Open-world re-identification, triplet loss, deep neural network, top-view configuration.

I. INTRODUCTION

Person re-identification (re-ID) is the task of recognising individuals at different locations and times, which involves different camera views, poses and lighting [1]. This topic has gained increasing interest in the computer vision community due to its challenging nature, and its important practical role underpinning many visual surveillance functionalities, including person searching and tracking across disjointed cameras [2]. Person re-ID has been adopted in several domains ranging from video surveillance to retail [3].

In a common real-world application, a watch list of known people is given as the gallery/target set for searching through a large volume of video recording locations where said people are likely to return. This aspect is fundamental in retail to understand how customers schedule their shopping. The identification of regular and occasional customers allows temporal purchasing profiles to be defined, which can correlate customer temporal habits with other information, such as expenditure amounts and numbers of purchased items. This knowledge enables novel marketing strategies tailored to the temporal and systematic behavior of each customer, as well as new innovative services and increased customer awareness based on shopping schedule recommendations [4], [5].

Video captured by store cameras usually contain people who are not part of a watch list. Moreover, a target person can appear similar to a non-target person whilst dissimilar to target gallery videos due to significant changes lighting and view angle conditions across camera views. To further aggravate the problem, there may only be a single gallery image (a one-shot) available for each target person, preventing the effective learning of a target's appearance variations. Facing

The associate editor coordinating the review of this manuscript and approving it for publication was Tomasz Trzcinski.

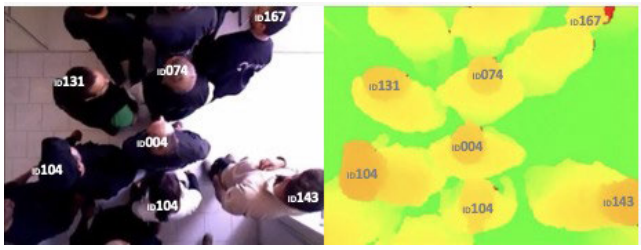


FIGURE 1. Example of RGB-D videos acquired in a persistent crowded environment with person identification. The figure depicts both RGB (left) and Depth (right) streams, showing that the top-view approach allows to avoid occlusions between people, a situation where the frontal approach often fails.

re-ID issues becomes difficult in a crowded retail environment with many occlusions [6], especially where probe sets contain mostly irrelevant (non-target) people. This problem is called open-world re-ID [7]–[9]. For such a challenging problem, depending on a fully automated system to provide exhaustive accurate verification against each targeted individual is neither scalable nor tractable. Nonetheless, it is adequate to expect an automated system to produce some screening by dealing with an easier problem: checking if a targeted person is in a given set (group-based person verification), whilst leaving the more challenging task of individual identification within the set for a human operator. Since watch lists are typically small, human verification can be carried out quickly and more robustly. Many approaches investigate either the best feature representation [10]–[14] or the best matching metrics [15], [16] when using person re-ID under difficult appearance changes across camera views.

They are not suitable to reidentify people in the retail environment as they assume a closed-world setting with probe sets containing exactly the same people in the gallery set. For probe sets consisting of mostly non-target people (many more than those in the gallery set), the re-ID problem becomes more arduous. They also do not consider retail environments where analytic interactions and re-ID are developed with the aim of learning shopper skills based on occlusion-free RGB-D cameras in a top-view configuration [17]–[19].

Furthermore, reidentifying a person in more crowded situations is a problem that remains largely unresolved due to many serious issues, such as the exhibition of persistent occlusion, appearance changes and dynamic or complex backgrounds. All of these issues cause extreme problems when encountered with a crowded environment, since conventional surveillance technologies have difficulty understanding video (Figure 1).

This paper presents the first attempt to solve a more realistic re-ID setting, facing these important issues using top-view open-world (TVOW) person re-ID. Among TVOW's important characteristics is its basis on a pretrained deep convolutional neural network (DCNN) that has been fine-tuned on a dataset acquired via a top-view configuration. A special loss function called triplet loss was used to train the network, optimising the embedding space such that data points with matching identities are closer to each other than those with different identities. Similar to Herman's work [20], triplet loss allows end-to-end learning between input images and a desired embedding space. We can also compare people by computing the Euclidean distance of their embeddings. In addition to the normal metrics used in a closed-set environment, particular metrics were defined, employed and evaluated for an open-set environment. The TVOW perspective was evaluated on a new publicly available dataset: TVPR2 for person re-ID.[1] The experimental results showed that the proposed methods were suitable for this task, bringing different and significative insights for implicit and extensive shopper behaviour analysis in marketing applications.

The main contributions of this paper, compared to the state of art, are: i) a solution, for real retail environments with a great variability in data acquired, derived from a large experience over 10.4 million shoppers observed in two years in different types of stores and in different countries; ii) a framework for shopper re-ID in crowded environments; iii) a new dataset with an RGB-D camera in top-view configuration with data acquired in a real retail environment that is publicly available to the scientific community for testing and comparing different approaches; iv) a new deep learning approach based on triplet loss and able of working both in closet-set and open-set environments; v) a comparison of TVOW approach with state-of-the-art methods.

The paper is organised as follows. Section II provides a description of the approaches used for people re-ID. Section III describes our approach and offers details on the TVPR2 dataset. In Section IV, we offer an extensive comparative evaluation of our approach with respect to state-of-the-art methods, as well as a detailed analysis of each component of our approach. Finally, in Section V, we draw conclusions and discuss future directions for this field of research.

II. RELATED WORKS

Open-set re-ID is much closer to practical video surveillance applications but its low recognition rates under low false accepted rates of existing results show that this setting is very challenging [2]. Historically, the scientific community has been devoted mainly to closed-set re-ID, a mature technology [21] that is convenient and fair for conducting research given its various baselines, datasets and evaluations.

However, open-set re-ID is a realistic approach that considers irrelevant people (those not part of a gallery) during recognition [22]. It can be defined as a person verification task instead of person identification, allowing verification of those who are part of a gallery and images in which those subjects appear [23]. The evaluation metrics are different. In fact, two metrics were defined by Zheng *et al.* [7], namely high true target recognition (TTR) and low false target recognition (FTR), which focus on calculating the likelihood of target and non-target numbers of images being verified as target identities.

[1] http://vrai.dii.univpm.it/content/tvpr2-dataset

The first work for open-set re-ID was proposed by Zheng *et al.* [24]. The authors showed a transfer ranking framework for set-based verification. Another approach, Cancela *et al.* [25], was based on online conditional random field inference.

In Liao *et al.* [8], open-set re-ID was decomposed into detection and identification while also being presented as two generic evaluation metrics (i.e., identification rate and false acceptance rate).

In Wang *et al.* [26], the authors tested a regularised kernel subspace learning model for one-shot verification by learning crossview identity-specific information from just unlabeled data.

Zheng *et al.* [7] presented clearer descriptions of open-set challenges and standard evaluation metrics, describing a group-based setting and a transfer local relative distance comparison model for addressing label scarcity. For performance evaluation, they used TTR and FTR.

Zhu *et al.* [9] proposed a hashing approach (cross-view identity correlation) and introduced a large-scale setting characterised by huge size probe images and an open person population.

Common re-ID approaches are usually based on frontal image datasets, but sensors installed in top-view configuration have been revealed as especially effective in crowded environments [6]. The latter configuration has several advantages because it prevents occlusion due to objects and other people while ensuring personal privacy, as faces are not recorded. Liciotti *et al.* [27] proposed a method to extract anthropometric features through image processing techniques, then training machine-learning algorithms for re-ID tasks. Their tests were carried out on a dataset of 100 people acquired using a top-view RGB-D camera.

Haque *et al.* [28] developed an attention-based model that deduces human body shape and motion dynamics by using depth information. Their approach was a combination of convolutional and recurrent neural networks leveraging unique 4-D spatio-temporal signatures to identify small discriminative regions indicative of human beings. Their tests were assessed on a DPI-T dataset, which consisted of 12 persons appearing in 25 videos while wearing different sets of clothing and holding different objects.

In [29], the authors started with a two-flow Convolutional Neural Network (CNN) (one for RGB and one for depth) and a final fusion layer. They improved on this approach with a multimodal attention network [30], adding an attention module to extract local and discriminative features that were fused with globally extracted features. In another work, Lejbolle *et al.* [31] presented a SLATT network with two types of attention modules (one spatial and one layer-wise). The authors collected also the OPR dataset from a university canteen, which was composed of 64 persons captured twice (entering and leaving a room). However, these datasets are not publicly available.

Recently, the person re-ID task is often solved using a triplet loss function, with excellent performance. In the work of Hermans *et al.* [20], the authors propose a batch hard function expecially designed for person re-ID problem: they show that, for both models trained from scratch or pretrained ones, using a well designed triplet loss can outperform most state-of-the-art methods. Yuan *et al.* [32] present a triplet loss that achieves good performance with large-scale re-ID datasets and has direct transferability with unseen datasets. Our framework uses a triplet loss function based on the work of Hermans *et al.* [20].

Due to Liciotti *et al.*'s success, and considering the advantages of the triplet loss function in improving networks performances, our study combine these approaches by the design of TVOW which considers open-set scenario is closer to retail applications than closed-set settings.

III. MATERIALS AND METHODS

In this section, we introduce the TVOW framework as well as the dataset used for evaluation. The framework is depicted in Figure 2. We use a novel modified DCNN for re-ID that is composed of the following phases:

- *Data Acquisition*: The dataset is acquired through the use of an RGB-D camera.
- *Person Detection*: Using the depth channel, people can be detected.
- *Preprocessing*: By combining depth information with RGB information, the background is removed from the image and only the important information (the person) remains.
- *Triplet Loss DCNNs*: Data augmentation techniques are used to fine-tune the networks, which are pretrained on the ImageNet dataset [33]. The triplet loss function is used for network training.
- *Evaluation*: Defining and evaluating specific metrics for this work.

Further details are given in the following subsections. The framework is comprehensively evaluated using the publicly available TVPR2 dataset.

A. TVPR2 DATASET

In this work, we collected a new dataset for person re-ID called TVPR2 (Top-View Person Re-identification 2). It was acquired following the procedure outlined in [27], which is closer to realistic settings. The dataset comprises 235 videos containing RGB and depth information. Each person during a recording session walked with an average gait within the area under the camera in one direction, then it turned back and repeated the same route in the opposite direction. The number of people present in the videos also varies from one to eleven with the entire dataset comprising 1027 unique individuals.

B. PREPROCESSING

The first problem to solve in a crowd environment is how to isolate individual people in each frame. Once isolated, we can proceed to extracting personal features and performing re-ID. Before using the dataset, individual frames were subjected to a preprocessing phase, shown in Figure 3. Firstly, RGB

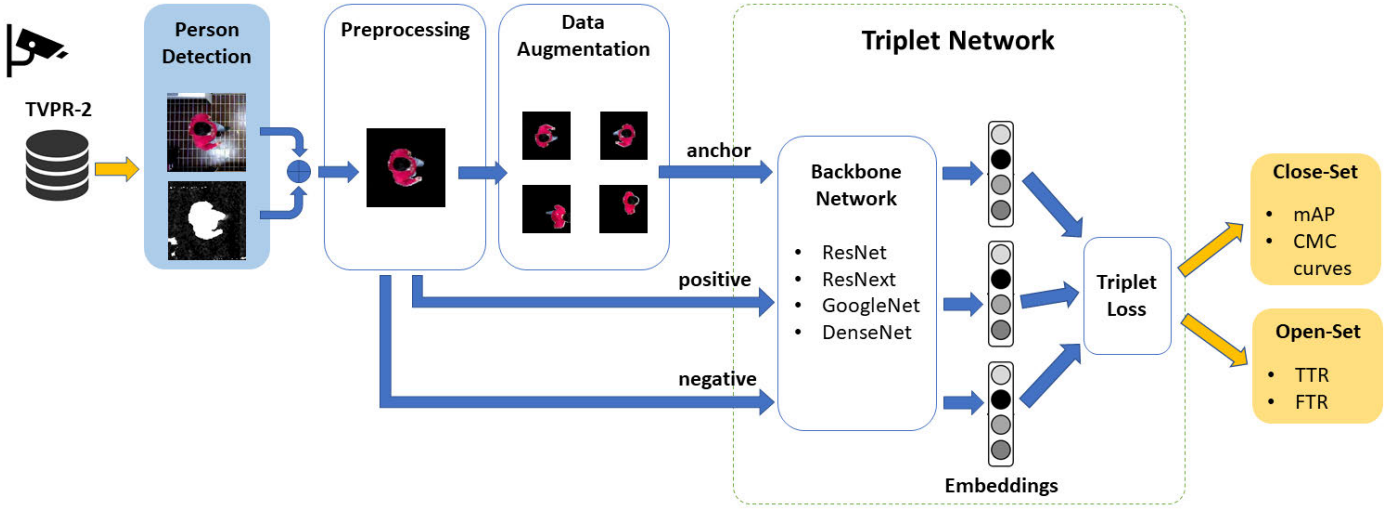


FIGURE 2. TVOW framework. Four phases are followed: Data acquisition, Person Detection, Data processing, Training of the Triplet Loss DCNN and performance evaluation.

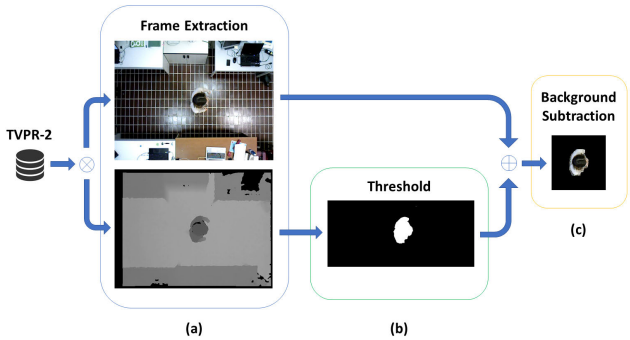


FIGURE 3. Preprocessing phase for the people detection task on an example frame of TVPR2 Dataset. (a) Frame Extraction for both streams. (b) Threshold on the Depth channel based on person's height. (c) Background subtraction by using the contour with the biggest area.

and Depth frames are extracted from their related streams, which was temporally and spatially synchronized. These have dimensions 320 × 240 pixels. A person detection algorithm made a crop of each person using a 150 × 150 pixel bounding box. This was made possible by using the depth channel and a threshold for a person's minimum height. In this way, noise produced by the frame background was removed to allow focusing on more important details (i.e., the person). The 150 × 150 pixel size was chosen experimentally, given the average dimensions of people in the dataset of between 80 × 80 and 125 × 125 pixels. As a further improvement, it was possible to use the depth information to remove the background inside the cropped image. This step was implemented using the previous mask to determine the outline with the largest area and then remove everything outside of that area. These cropped images were then used as input for our deep learning method.

C. TRIPLET LOSS DCNNs

Before training the networks, data augmentation techniques were applied to increase the dataset and improve network performance. Subsequently, images were given as input to a DCNNs. In this phase, various state-of-the-art networks were tested, pretrained on the public ImageNet dataset and then retrained on the TVPR2 dataset using the fine-tuning technique. The network has been trained using a triplet hard loss. With this technique, the input image a (*anchor*) is transformed into a feature embedding space. An image p of the same class (defined as *hard positive*) is taken as an image n of a different class (defined as *hard negative*). The network is subsequently trained to bring the anchor a closer to the hard positive p while simultaneously moving it away from the hard negative n. For the triplet loss function, we used the batch hard function proposed in [20], designed for person re-ID tasks: they show that, for both models trained from scratch or pretrained ones, using a well designed triplet loss can outperform most state-of-the-art methods. Batches of PK frames are created by randomly sampling P person IDs and K frames of each person. The triplet used to calculate the loss function is determined by selecting the hardest positive and the hardest negative samples within the batch for each sample a of the batch itself. Our triplet loss is defined as follows:

$$L_{Triplet} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\ anchors} + [m + \overbrace{\max_{p=1..K} D(a^i, p^i)}^{hardest\ positive} - \overbrace{\min_{\substack{j=1..P \\ n=1..K \\ j\neq 1}} D(a^i, n^j)}^{hardest\ negative}] \quad (1)$$

where the hard positive samples refer to poses of the same person in different frames and hard negative samples refer to similar-looking people.

D. EVALUATION METRICS

In literature, there are few evaluation metrics for the open-set environment. The only existing ones were studied in [34] and are described below. Liao *et al.* [8] proposed calculating the cumulative matching curve (CMC) rate (usually used for close-set re-ID) at a fixed false accept rate to indicate the likelihood of misidentifying a person. In the work of Wang *et al.* [26], the false accept rate evaluation was used on two standard datasets for frontal person re-ID. According to [34], neither of these two studies worked well because the CMC metric is dependant on similar identity correspondence in a closed-set scenario. A completely different approach was used by Zheng *et al.* [7] and Zhu *et al.* [9] which adopted the True Target Rate (TTR) and False Target Rate (FTR) for evaluation in an open-set environment. We will use these metrics in our approach. Should several non-target persons be placed in the probe population, the aim is not only to measure performance based on how well target probe persons are matched, but also how badly non-target persons pass through the verification process. To evaluate the performance of different open-set environment methods, we will compare their measured TTR values.

To evaluate various approaches under a different verification standard, we can compare their TTR values against a series of given FTR values. We can describe TTR and FTR values in the following equations:

$$TTR = \frac{N_{t2t}}{N_t}; \tag{2}$$

$$FTR = \frac{N_{nt2t}}{N_{nt}}; \tag{3}$$

where TTR is the number of accurate verifications N_{t2t} (target probe images are matched in the gallery) divided by the number of probe images from target persons N_t and FTR is the number of false verifications N_{nt2t} (non-target probe images treated as target persons) divided by the number of probe images from non-target persons N_{nt}.

Although TTR differs from the CMC rate, it also indicates the probability of the correct target, which means they can considered comparable to some extent.

For evaluation, according to [34], TTR values (with certain FTR values) are preferred over traditional CMC rates. TTR values can measure performance by verifying target and non-target persons, and are independent of one-to-one identity correspondence (a closed-set hypothesis).

TTR and FTR values must be calculated using the same test, and their purpose is to show how the network behaves in the presence of people unknown to it. The optimal result would be a high value of TTR with a low FTR value, showing that the network succeeded in correctly separating targets from non-targets.

A high FTR value, regardless of the corresponding TTR value, would mean the network had failed to exclude non-targets and subsequently assigned them identities of targets. To calculate pairs of these parameters, we use the following matching algorithm:

TABLE 1. Mean average precision for close-set configuration.

mAP(%)	100 persons	300 persons	1000 persons
ResNet-50	92,10	81,09	76,86
ResNeXt-50	93,30	84,23	78,74
DenseNet-161	91,56	79,20	69,08
GoogleNet	75,54	54,58	41,00

- Calculate the Euclidean distance $d(\widetilde{x^p}, \widetilde{x}_i^g)$, between the probe $\widetilde{x^p}$ and all gallery vectors $\widetilde{x}_i^g$ defined as $i^* = arg\min_i d(\widetilde{x^p}, \widetilde{x}_i^g)$ considering the i^*th element of the gallery as the identity to assign to the probe.
- Given a threshold ϕ_m, we identify person $\widetilde{x^p}$ as target if $d(\widetilde{x^p}, \widetilde{x}_{i^*}^g) < \phi_m$. Otherwise, the person is a non-target.
- We consider a person a target when $d(\widetilde{x^p}, \widetilde{x}_{i^*}^g) < \phi_m$ and simultaneously $\widetilde{x}_{i^*}^g$ and $\widetilde{x^p}$ belongs to that person. Otherwise, $\widetilde{x^p}$ is treated as a non-target.
- The steps are repeated for each vector of probe.
- TTR and FTR values are calculated according to Equations 3.

IV. RESULTS AND DISCUSSIONS

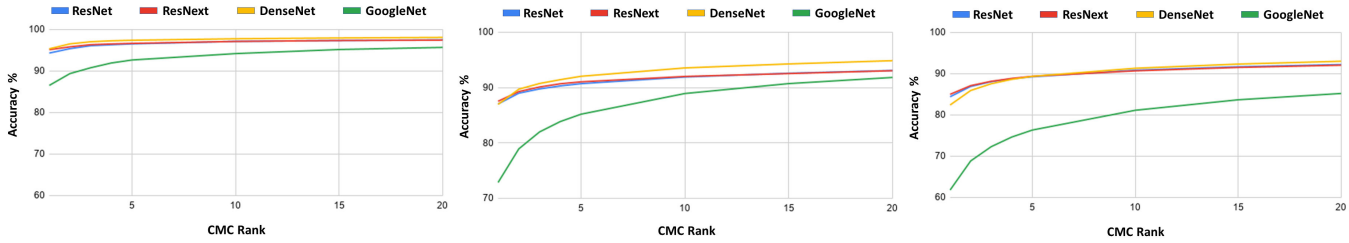
In this section, the results of the experiments conducted using the TVPR2 dataset are presented. Our experiments were separated in two phases. First, we used the new TVPR2 dataset to find the best combination of hyperparameters and backbone network. Second, we tested the best configuration on a state-of-the-art dataset, namely the TVPR dataset [27].

Our approach was tested using the backbones of several state-of-the-art networks, pretrained using the ImageNet dataset, then fine-tuned on our TVPR2 dataset. The chosen networks were:

- ResNet-50 [35], fine-tuned on Layer4, with 68,9 Millions of parameters;
- ResNext-50 [36], fine-tuned on Layer4, with 25 Millions of parameters;
- DenseNet-161 [37], fine-tuned on DenseBlock4 Layer, with 28.7 Millions of parameters;
- GoogleNet [38], fine-tuned on Inception5b Layer, with 6.7 Millions of parameters.

The dataset used for testing was TVPR2. During network training, data augmentation techniques, such as flipping, rotation, random crop and padding, were used.

The tests were carried out initially using a close-set configuration through the mean average precision (mAP) metric and CMC curves. In the second phase, an open-set environment was tested using various combinations of TTR and FTR values. The tests were repeated with a variable number of id. Our networks were trained sequentially with 100, 300 and 1000 people from the dataset. The metrics used in this particular configuration were the mAP and the CMC. Table 1 shows the mAP for each tested network. By contrast, Figure 4 compares the CMC curves of every backbone for three different ranges of a person's IDS.



(a) 100 people. (b) 300 people. (c) 1000 people.

FIGURE 4. CMC curves for close-set configuration. The tests were repeated with a variable number of ids: (a) 100 people, (b) 300 people and (c) 1000 people.

These results indicate increasing the number of people in the phase of training results in performance deterioration. In particular, the GoogleNet exhibited more difficulty when learning and the largest deterioration with more people. The most stable networks were ResNet and ResNext.

As shown by the CMC curves (Figure 4), the performances by ResNet, ResNeXt and DenseNet were very similar. The rank1 exceeded 90%, indicating no particular increase with higher rank. GoogleNet remains worst performing network.

As the the best results (on average) were obtained using ResNext-50 as the backbone network (with retraining via fine-tuning on Layer4), this configuration was used for the next comparison. The following tests were then performed to evaluate performance in an open-set environment. In this situation, there are an unknown number of people that the network has not used for the training. The objective of these tests is to judge the ability of the networks to correctly identify already known targets while discerning unknown non-targets. To do this, new state-of-the-art metrics have been chosen, TTR and FTR values, already introduced in previous sections. The experiments were performed by varying the number of targets, using 100, 300 and 500 people per trial. For each of these cases, we then increased the number of non-targets by a percentage value of 10%, 50% and 100% compared to the number of targets, as shown in Figure 5.

The first graph (Figure 5(a)) compares TTR and FTR values by using 100 targets and 10 non-targets for testing. This graph indicates how obtaining a low FTR value generally leads to a decreased TTR value. For an FTR value of 30%, the TTR is closer to 100%, a situation caused by the positioning of non-targets in the feature space. For a threshold of 30%, most of the non-targets' features are far from those of the targets, while going below 30% non-targets caused the targets to start moving closer and creating confusion. The TTR value decreased because the matching algorithm applied a threshold on the distance, which considered all features beyond it as belonging to non-targets. Regarding network performance, the results are clear: ResNeXt was more robust than the others.

The graphs on the left of Figure 5 show the metrics variation when the number of targets is increased while maintaining the percentage of non-targets. Figure 5(b) compares TTR and FTR values by using 300 targets and 30 non-targets and Figure 5(c) compares TTR and FTR values given input of 500 targets and 50 non-targets. The shape of the curves is similar to the case with 100 targets, but the values decreased generally as the number of targets increased. This result was predictable as other metrics deteriorated for the same reasons. In the 500 target case, however, there was an anomalous worsening of ResNext.

The graphs on the right of Figure 5 show the effect of increasing the number of non-targets from 10% (Figure 5(a)) to 50% (Figure 5(d)) and 100% (Figure 5(e)). The difference between the 10% and 50% cases is a slight deterioration in the TTR value given an FTR value range between 10% and 20%. This difference is caused by additional non-targets being positioned around the threshold values and being interpreted as targets. Increasing non-targets from 50% to 100% produced practically identical graphs, indicating the new non-targets were distributed in the same positions as the old ones. This behaviour is due to the operating principle of the triplet networks; in fact, they learn to cluster classes and this effect extends beyond learned classes to newly encountered classes. This important result demonstrates the strength of triplet networks in an open-set environment.

Table 2 provides a comparison of our approach, Triplet Loss DCNN (TL-DCNN), with other state-of-the-art methods concerning person re-ID from a top-view perspective. TVDH is the method of Liciotti *et al.* [27] to extract anthropometric features through image processing techniques, then training machine-learning algorithms for re-ID tasks. In RGB-D-CNN, Lejbolle *et al.* [29] started with a two-flow Convolutional Neural Network (CNN) (one for RGB and one for depth) and a final fusion layer. Then they improved this approach with a multimodal attention network called MAT [30], adding an attention module to extract local and discriminative features that were fused with globally extracted features. In another work, Lejbolle *et al.* [31] presented a SLATT network with two types of attention modules (one spatial and one layer-wise). Table 2 shows the results in terms of CMC curves (rank-1, rank-5, rank-10 and rank-20).

As a final step, we tested and compared the TL-DCNN approach with the most recent state-of-the-art approach SLATT [31], using our TVPR2 dataset. The test was made

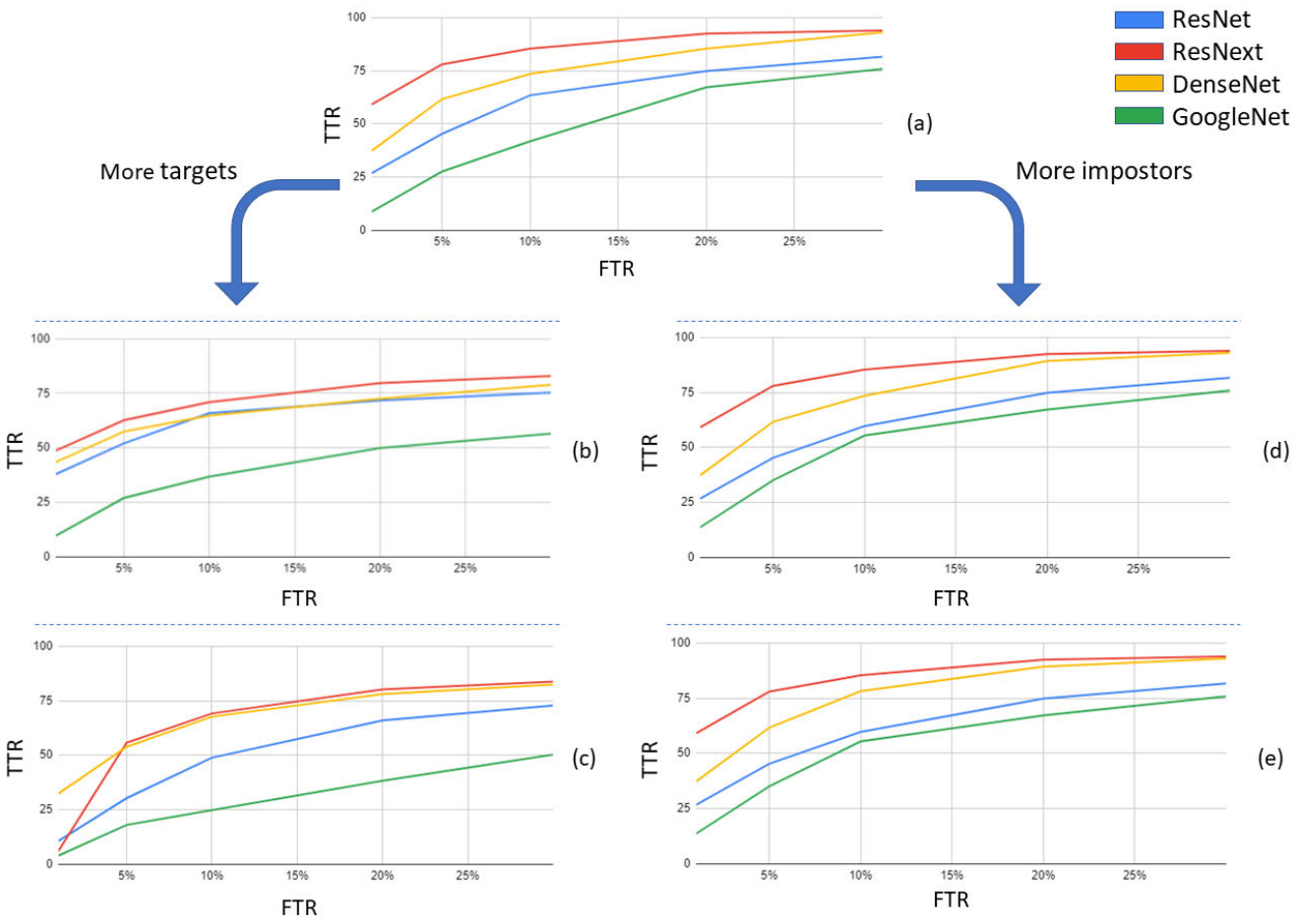


FIGURE 5. TTR and FTR value results for an open-set environment. (a) Test on 100 targets and 10 non-targets. (b) Test on 300 targets and 30 non-targets. (c) Test on 500 targets and 50 non-targets. (d) Test on 100 targets and 50 non-targets. (e) Test on 100 targets and 100 non-targets.

TABLE 2. Test using the TVPR dataset to compare TL-DCNN with other state-of-the-art methods.

Method/Rank	r = 1	r = 5	r = 10	r = 20
TVDH [27]	75,50	87,50	89,20	91,20
RGB-D-CNN [29]	92,55	97,87	97,87	100,00
MAT [30]	94,68	97,87	97,87	100,00
SLATT [31]	93,62	96,81	97,87	100,00
TL-DCNN	95,13	98,03	98,75	100,00

TABLE 3. Testing using the TVPR2 dataset comparing TL-DCNN with SLATT. Results are based on mAP and CMC curves.

Method	mAP	r = 1	r = 5	r = 10	r = 20
SLATT [31]	90,18	91,30	93,77	94,08	95,30
TL-DCNN	93,30	94,02	96,68	97,12	98,55

in a closed-set environment with 100 targets. Table 3 shows results of this comparison in terms of mAP and CMC curves (rank-1, rank-5 rank-10 and rank-20).

Our method, based on depth information, allows a person detection with the minimum possible error, because the RGB and Depth frames are automatically synchronized by the camera, both spatially and temporally. In this way, we can also remove the noise due to the background around the person (through the background removal phase) and then learn only the main features of the person itself. Moreover, the use of triplet loss with hard batch allows to train the network in an efficient way, because it increases the distance between the features of the sample frame (anchor) and the frames of different people (negative), while it decreases the distance with the frames of the same person but in different poses (positive).

From these results, it is possible to evaluate which people were not recognised more frequently and with whom they were confused. Figure 6 highlights four examples of mismatched targets.

Finally, we have added a further test to compare our depth-based method of person detection with a state-of-the-art one: the Region of Interest (ROI) was extracted from the original frame by using a You Only Look Once (YOLO) detector, trained only on the RGB frames. In our approach, we have improved the phase of ROI detection by using a threshold on the depth channel. From Figure 7, it is possible to infer the improvements respect the two used methods. Using

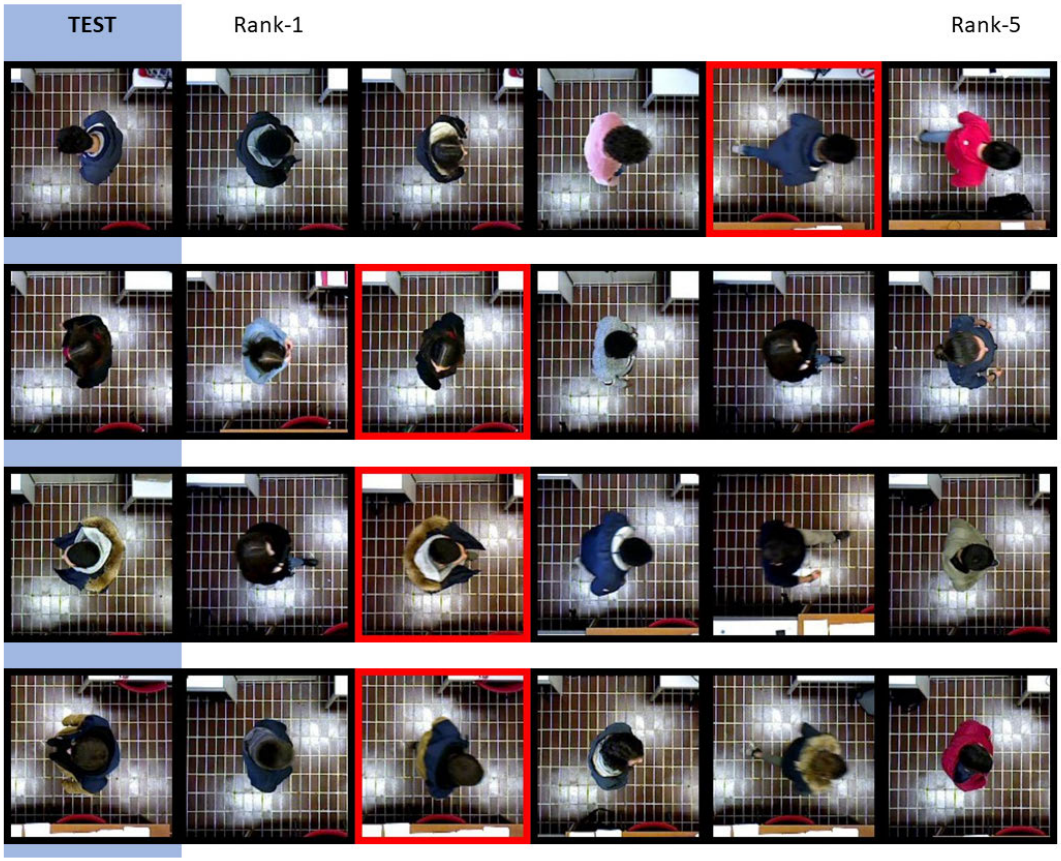


FIGURE 6. Examples of missmatched IDs for a visual analysis of the results. The first column shows the RGB frame of the person in the test set (obviously the relative depth frame is also given as input). The others show representative images for the first 5 predicted IDs. The red box figures the ground truth.

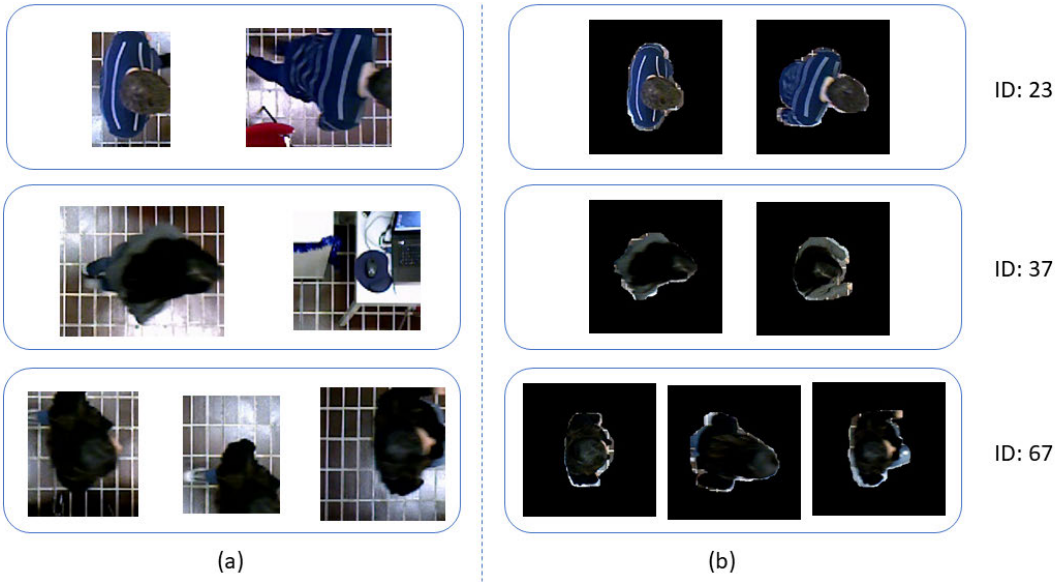


FIGURE 7. Preprocessing comparison for person detection. (a) Some incorrect detections using YOLO. (b) Correct detections of identical targets using our approach based on depth information.

the YOLO-based method, several errors are committed in person detection: the person could be partially picked up, or shot differently on continuous frames; or even confused with some objects (Figure 7.a). Our approach ensures that a height threshold is set on the depth channel, so as to remove all the lower objects than people. Furthermore, the detection will be done only on objects that have passed a certain limit area, to be sure that they are really people. Finally, using the depth information, the background will be removed and only the person's information will remain(Figure 7.b).

V. CONCLUSION AND FUTURE WORKS

We have presented a novel deep learning framework for person re-ID, the TVOW approach, able to work in both closed-set and open-set environments. The approach is based on a pretrained DCNN, fine-tuned on a dataset acquired with a top-view configuration. A special loss function, triplet loss, was also used to train the network. In addition to the normal metrics used in a closed-set environment, particular metrics were also used to work in an open-set environment. The paper describes one of the more extensive test based on real data from real retail scenarios in the literature.

The results showed that the proposed methodology is suitable and accurate, overcoming and advancing the state of the art. For this purpose, TVPR2, a new public dataset was collected and shared with the framework source codes to ensure comparisons with the proposed method and future improvements and collaborations over these challenging scenarios. The results yield high accuracy and demonstrate the effectiveness and the suitability of the proposed approach, especially for crowded scenario where accurate people counting and re-identification is needed (i.e. intelligent retail environments).

Future works should improve and better integrate the triplet loss DCNNs with more complex architectures able to improve performances. Incremental learning methods will be investigated to improve the on-line performances of the re-identification algorithm. Further investigation on DCNNs generalisations are needed to prove the effectiveness of the approach in very different retail categories (from grocery to fashion) and in cross-country human behaviours and attitudes.

ACKNOWLEDGMENT

The authors would like to thank R. Pietrini, M Contigiani, and L Di Bello for their support.

REFERENCES

[1] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Person Re-Identification*. Springer, 2014, pp. 1–20.
[2] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: http://arxiv.org/abs/1610.02984
[3] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-IDentification via multi-loss dynamic training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8514–8522.
[4] M. Paolanti, L. Romeo, M. Martini, A. Mancini, E. Frontoni, and P. Zingaretti, "Robotic retail surveying by deep learning visual and textual data," *Robot. Auto. Syst.*, vol. 118, pp. 179–188, Aug. 2019.
[5] N. Ferracuti, C. Norscini, E. Frontoni, P. Gabellini, M. Paolanti, and V. Placidi, "A business application of RTLS technology in intelligent retail environment: Defining the shopper's preferred path and its segmentation," *J. Retailing Consum. Services*, vol. 47, pp. 184–194, Mar. 2019.
[6] D. Liciotti, M. Paolanti, R. Pietrini, E. Frontoni, and P. Zingaretti, "Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1384–1389.
[7] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 591–606, Mar. 2016.
[8] S. Liao, Z. Mo, J. Zhu, Y. Hu, and S. Z. Li, "Open-set person re-identification," 2014, *arXiv:1408.0872*. [Online]. Available: http://arxiv.org/abs/1408.0872
[9] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng, "Fast open-world person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2286–2300, May 2018.
[10] P. Dollar, Z. Tu, H. Tao, and S. Belongie, "Feature mining for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.
[12] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2008, pp. 262–275.
[13] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 413–422.
[14] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.
[15] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 780–793.
[16] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.
[17] D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, and V. Placidi, "Shopper analytics: A customer activity recognition system using a distributed RGB-D camera network," in *Proc. Int. Workshop Video Analytics for Audience Meas. Retail Digit.* Singapore: Springer, 2014, pp. 146–157.
[18] D. Liciotti, M. Paolanti, E. Frontoni, and P. Zingaretti, "People detection and tracking from an RGB-D camera in top-view configuration: Review of challenges and applications," in *Proc. Int. Conf. Image Anal. Process.* Springer, 2017, pp. 207–218.
[19] M. Paolanti, L. Romeo, D. Liciotti, A. Cenci, E. Frontoni, and P. Zingaretti, "Person re-identification with RGB-D camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection," *Sensors*, vol. 18, no. 10, p. 3471, 2018.
[20] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737
[21] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Comput. Surveys*, vol. 46, no. 2, pp. 1–37, Nov. 2013.
[22] X. Li, A. Wu, and W.-S. Zheng, "Adversarial open-world person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 280–296.
[23] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image Vis. Comput.*, vol. 32, no. 4, pp. 270–286, Apr. 2014.
[24] W.-S. Zheng, S. Gong, and T. Xiang, "Transfer re-identification: From person to set-based verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2650–2657.
[25] B. Cancela, T. Hospedales, and S. Gong, "Open-world person re-identification by multi-label assignment inference," in *Proc. Brit. Mach. Vis. Conf.*, 2014.

[26] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 769–773.
[27] D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti, "Person re-identification dataset with RGB-D camera in a top-view configuration," in *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*. Springer, 2016, pp. 1–11.
[28] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1229–1238.
[29] A. R. Lejbolle, K. Nasrollahi, B. Krogh, and T. B. Moeslund, "Multimodal neural network for overhead person re-identification," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2017, pp. 1–5.
[30] A. R. Lejbolle, B. Krogh, K. Nasrollahi, and T. B. Moeslund, "Attention in multimodal neural networks for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 179–187.
[31] A. R. Lejbolle, K. Nasrollahi, B. Krogh, and T. B. Moeslund, "Person re-identification using spatial and layer-wise attention," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1216–1231, 2020.
[32] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," 2019, *arXiv:1912.07863*. [Online]. Available: http://arxiv.org/abs/1912.07863
[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[34] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1092–1108, Apr. 2020.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 630–645.
[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

MASSIMO MARTINI is currently pursuing the Ph.D. degree with the Department of Information Engineering (DII), Universitá Politecnica delle Marche. His research interests include point cloud semantic segmentation, re-identification, and social media intelligence.

MARINA PAOLANTI (Member, IEEE) is currently a Postdoctoral Research Fellow and a Contract Professor with the Department of Information Engineering (DII), Università Politecnica delle Marche. During her Ph.D., she has worked at GfK Verein, Nuremberg, Germany, for Visual and Textual sentiment analysis of brand related social media pictures using deep convolutional neural networks. Her research interests include deep learning, machine learning, image processing, and computer vision. Her research focuses on deep learning, and computer vision techniques applied to retail and cultural heritage. She is a member of CVPL and AI*IA.

EMANUELE FRONTONI received the Ph.D. degree in intelligent artificial systems from the Department of Information Engineering (DII), Universitá Politecnica delle Marche, in 2006. He is currently a Professor of computer science with the Universitá Politecnica delle Marche, Italy. His Ph.D. thesis was on vision-based robotics. His research interests include artificial intelligence and computer vision techniques applied to robotics, the Internet of Things, e-health, and ambient assisted living. He is a member of ASME MESA TC, CVPL, and AI*IA.