sensors

MDPI

Article

Bodyprint—A Meta-Feature Based LSTM Hashing Model for Person Re-Identification

Danilo Avola [1,2,*], Luigi Cinque [1], Alessio Fagioli [1], Gian Luca Foresti [3] and Daniele Pannone [1] and Claudio Piciarelli [3,*]

[1] Department of Computer Science, Sapienza University, 00198 Rome, Italy; cinque@di.uniroma1.it (L.C.); fagioli@di.uniroma1.it (A.F.); pannone@di.uniroma1.it (D.P.)

[2] Department of Communication and Social Research, Sapienza University, 00198 Rome, Italy

[3] Department of Mathematics, Computer Science and Physics, University of Udine, 33100 Udine, Italy; gianluca.foresti@uniud.it

* Correspondence: avola@di.uniroma1.it (D.A.); claudio.piciarelli@uniud.it (C.P.)

Received: 30 June 2020; Accepted: 14 September 2020; Published: 18 September 2020

check for updates

Abstract: Person re-identification is concerned with matching people across disjointed camera views at different places and different time instants. This task results of great interest in computer vision, especially in video surveillance applications where the re-identification and tracking of persons are required on uncontrolled crowded spaces and after long time periods. The latter aspects are responsible for most of the current unsolved problems of person re-identification, in fact, the presence of many people in a location as well as the passing of hours or days give arise to important visual appearance changes of people, for example, clothes, lighting, and occlusions; thus making person re-identification a very hard task. In this paper, for the first time in the state-of-the-art, a meta-feature based Long Short-Term Memory (LSTM) hashing model for person re-identification is presented. Starting from 2D skeletons extracted from RGB video streams, the proposed method computes a set of novel meta-features based on movement, gait, and bone proportions. These features are analysed by a network composed of a single LSTM layer and two dense layers. The first layer is used to create a pattern of the person's identity, then, the seconds are used to generate a bodyprint hash through binary coding. The effectiveness of the proposed method is tested on three challenging datasets, that is, iLIDS-VID, PRID 2011, and MARS. In particular, the reported results show that the proposed method, which is not based on visual appearance of people, is fully competitive with respect to other methods based on visual features. In addition, thanks to its skeleton model abstraction, the method results to be a concrete contribute to address open problems, such as long-term re-identification and severe illumination changes, which tend to heavily influence the visual appearance of persons.

Keywords: person re-identification; long short-term memory networks; 2D skeletons; RGB video sequences; joints based meta-features; binary coding; hashing

1. Introduction

Last years have seen the design of increasingly advanced computer vision algorithms to support a wide range of critical tasks in a plethora of application areas. These algorithms often have the responsibility of taking determining decisions in issues where failure would lead to serious consequences. In References [1–3], for example, vision-based systems are used for inspection of pipeline infrastructures. In particular, in the first work, the authors presented a method for subsea pipeline corrosion estimation by using colour information of corroded pipes. To manage the degraded underwater images, the authors developed ad-hoc pre-processing algorithms for image restoration and enhancement. Differently, in the second and third work, the authors focused on large infrastructures, such as sewers and waterworks. Both works proposed an anomaly detector based on unsupervised machine learning techniques, which, unlike other competitors in this specific field, do not require annotated samples for training the detection models. Correct evaluations by the algorithms reported above can provide countless advantages in terms of maintenance and hazard determination. Even in hand and body rehabilitation, as shown in References [4–6], last few years have seen the proliferation of vision-based systems able to provide measurements and predictions about the recovery degree of lost skills by patients affected with strokes and degenerative diseases. The first work focused on a hand skeleton model together with pose estimation and tracking algorithms both to determine palm and fingers pose estimation and to track their movements during a rehabilitation exercise. Similarly, but using a body skeleton model, the second and third work used a customized long short term memory (LSTM) model to analyze movements and limitations of patients' body, treating a full body rehabilitation exercise like a long action over time. Hand and body skeleton models, in addition to allowing the transposition of patients' virtual avatars into immersive environments, also allow therapists to better detect joints that require of more exercises, thus optimizing and customizing the recovery task.

Without doubt, one of the application areas in which computer vision has had more impact, in the last ten years, is the active video surveillance, that is, those automatic systems able to replace human operators in complex tasks, such as intrusion detection [7,8], event recognition [9,10], target identification [11,12], and many others. In References [13], for example, the authors proposed a reinforcement learning approach to train a deep neural network to find optimal patrolling strategies for Unmanned Aerial Vehicles (UAVs) visual coverage tasks. Unlike other works in this application area, their method explicitly considers different coverage requirements expressed as relevance maps. References [14–16], instead, even if use pixel-based computer vision techniques, that is, low level processing, are able to achieve remarkable results in terms of novelty recognition and change detection, thus providing a step forward in the field of aerial monitoring. Moving to stationary and Pan–Tilt–Zoom (PTZ) cameras, very recent works, as that reported in Reference [17], shown that, even in challenging application fields, robust systems can be implemented and applied, for security contexts, in everyday life. In particular, the authors proposed a continuous-learning framework for context-aware activity recognition from unlabelled video. In their innovative approach, the authors formulated an information-theoretic active learning technique that utilizes contextual information among activities and objects. Other interesting current examples are presented in References [18,19]. In the first, a novel end-to-end partially supervised deep learning approach for video anomaly detection and localization using only normal samples is presented. Instead, in the second, the authors propose an abnormal event detection hybrid modulation method via feature expectation subgraph calibrating classification in video surveillance scenes. Both works proven to provide remarkable results in terms of robustness and accuracy.

Active surveillance systems, in real contexts, are composed of different algorithms, each of which collaborates with the others for the achievement of a specific target. For example, algorithms for separating background from foreground, for example, References [20–22], are often used, as pre-processing stage, both to detect the objects of interest in the scene and to have a reference model of the background and its variations over time. Another example are the tracking algorithms, for example, Reference [10,23], which are used to analyze moving objects. Among these collaborative algorithms, person re-identification ones, for example, References [24,25], play a key role, especially in security, protection, and prevention areas. In fact, being able to identify a person's identity and being able to verify the presence of that person in other locations hours or even weeks before is a fundamental step for the areas reported above. Anyway, despite the efforts of many computer vision researchers in this application area, person re-identification task still presents several problems largely unsolved. Most of the person re-identification methods, in fact, are based on visual features extracted from images to model a person's appearance [26,27]. This leads to a first class of problems since, as known, visual features have many weaknesses, including illumination changes, shadows, direction of light, and many others. Another class of problems regards the background clutter [28,29] and occlusions [30,31], which tend, in uncontrolled environments, to lower system performances in term of accuracy. A final class of problems, very important from a practical point of view, is referred to the long-term re-identification and camouflage [32,33]. Many systems, based totally or in part on visual features, are not able to re-identify persons under the two issues reported above, thus making the use of these systems limited in real contexts.

In this paper, a meta-feature based LSTM hashing model for person re-identification is presented. The proposed method takes inspiration from some of our recent experiences in using 2D/3D skeleton based features and LSTM models to recognize hand gestures [34], body actions [35], and body affects [36] in long video sequences. Unlike these, the 2D skeleton model, in this paper, is used to generate biometric features referred to movement, gait, and bone proportions of the body. Compared to the current literature, beyond the originality of the overall pipeline and features, the proposed method presents several novelties suitably studied to address, at least in part, the classes of problems reported above, especially in uncontrolled and crowded environments. First, the method is fully based on features extracted from 2D skeleton models present in a scene. The algorithm used to extract the models, reported in Reference [37], is proven to be highly reliable in terms of accuracy, even for multi-person estimation. The algorithm uses each input image only to produce the 2D locations of anatomical keypoints for each person in the image. This means that all the aspects, for example, illumination changes, direction of light, and many others, that can influence the re-identification by systems based on visual appearance of people, can, in the proposed method, be overlooked since the images are only used to generate 2D skeletons. Second, the proposed pipeline does not use visual features, but only features derived by analysing the 2D skeleton joints. These features, through the use of the LSTM model, are designed to catch both dynamic correlations of different body parts during movements and gaits as well as bone proportions of relevant body parts. The term meta-features was born from the fact that our features are hence conceived to generalize the body in terms of movements, gaits, and proportions. Thanks to this abstraction, long-term re-identification and some kinds of camouflages can be better managed. Third, thanks to the two final layers of the proposed network, which implement the bodyprint hash through binary coding, the representative space of our method can be considered uncountable, thus providing a tool to potentially label each and every human being in the world. The proposed method was tested on three challenging datasets designed for person-re-identification in video sequences,that is, iLIDS-VID [38], Person Re-ID (PRID) 2011 [39], and Motion Analysis and Re-identification Set (MARS) [40], showing remarkable results, compared with key works of the current literature, in terms of re-identification rank. Summarizing, the contributions of the paper with respect to both the present open issues and the current state-of-the-art in terms of pipelines and models can be outlined as follows:

- Concerning the present person re-identification field open issues, the proposed method, which is not based on visual appearance, can be considered a concrete support in uncontrolled crowded spaces, especially in long-term re-identification where people, usually, change clothes and some aspects of their visual appearance over time. In addition, the use of RGB cameras allows the method to be used in both indoor and outdoor environments; overcoming, thanks to the use of 2D skeletons, well-known problems related to scene analysis, including illumination changes, shadows, direction of light, background clutter, and occlusions.
- Regarding the overall pipeline and model, the approach proposed in this paper presents different novelties. First, even if some features are inspired by selected works of the current literature in recognizing hand gestures, body actions, and body affects; other features are completely new as also is their joint utilize. Second, for the first time in the literature, an LSTM hashing model for person re-identification is used. This model was conceived not only to exploit the dynamic patterns of the body movements, but also to provide, via the last two layers, a mechanism by which to provide a labelling approach for millions of people thanks to binary coding properties.

The rest of the paper is structured as follows. In Section 2, a concise literature review focused on person re-identification methods based on video sequences processing is presented. In Section 3, the entire methodology is described, including the proposed LSTM hashing model and meta-features. In Section 4, three benchmark datasets and comparative results with other literature works, are reported. Finally, Section 5 concludes the paper.

2. Related Work

In this section, selected works that treat person re-identification task on the basis of video sequences by deep learning techniques are presented and discussed. The same works are then reported in the experimental tests, that is, Section 4, for a full comparison.

A first remarkable work is reported in Reference [40], where the authors, first and foremost, introduced the Motion Analysis and Re-identification Set (MARS) dataset, a very large collection of challenging video sequences that, moreover, were also used for part of the experimental tests in the present paper. Beyond the dataset, the authors also reported an extensive evaluation of the state-of-the-art methods, including a customized one based on Convolutional Neural Networks (CNNs). The latter, supported by the Cross-view Quadratic Discriminant Analysis (XQDA) metric learning scheme [41] on iLIDS-VID and PRID 2011 datasets, and by Multiple Query (MQ), only on the MARS dataset, shown to outperform several competitive approaches, thus demonstrating a good generalization ability. The work proposed in Reference [42], instead, adopts a Recurrent Neural Network (RNN) architecture in which features are extracted from each frame by using a CNN. The latter incorporates a recurrent final layer that allows information to flow between time-steps. To provide an overall appearance feature for the complete sequence, the features, from all time-steps, are then combined by using temporal pooling. The authors conduced experiments on iLIDS-VID and PRID-2011 datasets, obtaining, on both, very significant results. Other two competitors, of the work proposed in this paper, are described in References [43,44]. In the first, the authors presented an end-to-end deep neural network architecture, which integrates a temporal attention model to selectively focus on the discriminative frames and a spatial recurrent model to exploit the contextual information when measuring the similarity. In the second, the authors described a deep architecture with jointly attentive spatial-temporal pooling, enabling a joint learning of the representations of the inputs as well as their similarity measurement. Their method extends the standard RNN-CNNs by decomposing pooling into two steps—a spatial-pooling on feature map from CNN and an attentive temporal-pooling on the output of RNN. Both works were tested on iLIDS-VID, PRID 2011, and MARS, obtaining outstanding results compared with different works of the current literature. The authors of Reference [45] proposed a method for extracting a global representation of subjects through the several frames composing a video. In particular, the method attends human body part appearance and motion simultaneously and aggregates the extracted features via the vector of locally aggregated descriptors (VLAD) [46] aggregator. By considering the adversarial learning approach, in Reference [47] the authors presented a deep few-shot adversarial learning to produce effective video representations for video-based person re-identification, using few labelled training paired videos. In detail, the method is based on Variational Recurrent Neural Networks (VRNNs) [48], which can capture temporal dynamics by mapping video sequences into latent variables. Another approach is to consider the walking cycle of the subjects, such as the one presented in Reference [49], where the authors proposed a super-pixel tracking method for extracting motion information, used to select the best walking cycle through an unsupervised method. A last competitor is reported in Reference [50], where a deep attention based Siamese model to jointly learn spatio-temporal expressive video representations and similarity metrics is presented. In their approach, the authors embed visual attention into convolutional activations from local regions to dynamically encode spatial context priors and capture the relevant patterns for the propagation through the network. Also this work was compared on the three mentioned benchmark datasets, showing remarkable results.

3. Methodology: LSTM Hashing Model and Meta-Features

In this section, the proposed architecture for person re-identification, based on skeleton meta-features derived from RGB videos and bodyprint LSTM hashing, is presented. A scheme summarizing the architecture is shown in Figure 1. In detail, starting from an RGB video, the OpenPose [37] library is first used to obtain skeleton joints positions. This representation is then fed to the feature extraction module, where meta-features analysing movements both locally (i.e., in a frame-by-frame fashion) and globally (e.g., averaging over the whole motion) are generated, fully characterizing a person. Subsequently, meta-features are given as input to the bodyprint LSTM hashing module. In this last component, local meta-features are first analysed via a single LSTM layer. The LSTM representation of local meta-features is then concatenated to global meta-features and finally fed to two dense layers, so that a bodyprint hash is generated through binary coding.

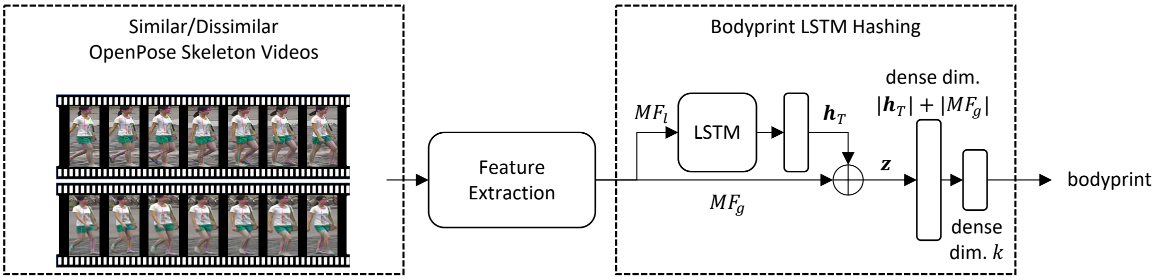


Figure 1. Long short term memory (LSTM) bodyprint hashing architecture scheme.

3.1. Skeleton Joint Generation

The first and foremost step, for a video-based person re-identification system using skeleton information, is a skeleton joint generation phase. At this stage, the well-known OpenPose library is exploited so that the main joints of a human skeleton can be extracted from RGB video sequences. In detail, this library leverages a multi-stage CNN so that limbs Part Affinity Fields (PAFs), a set of 2D vector fields encoding location and orientation of limbs over the image domain, are generated. By exploiting these fields, OpenPose is able to produce accurate skeletons from RGB frames and track them inside RGB videos. The extensive list of skeleton joints, representing body-foot position estimations, is depicted in Figure 2. As can be seen, up to 25 joints are described for a skeleton, where a joint is defined by its (x, y) coordinates inside the RGB frame. The identified joints correspond to—nose (0), neck (1), right/left shoulder, elbow, wrist (2–7), hips middle point (8), right/left hip, knee, ankle, eye, ear, big toe, small toes, and heel (9–24). While joint positions alone do not provide useful information, due to their strict correlation to the video they are extracted from, they can still be used to generate a detailed description of body movements, via the feature extraction module.

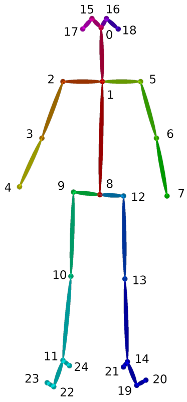


Figure 2. OpenPose body-foot keypoint schematic.

3.2. Feature Extraction

In this module, skeleton joint positions identified via OpenPose are used to generate a detailed description of full-body movements. Skeleton joint positions can be exploited to produce meaningful features able to describe, for example, gait, using the lower half of the body as shown by several gait re-identification works [51,52], or body emotions, leveraging features extracted from both lower and upper body halves [36]. Motivated by the encouraging results already obtained via hand-crafted skeleton features in References [34–36], two meta-feature groups built from skeleton joint positions are proposed in this work to analyze the whole body: local and global meta-features defined MF_l and MF_g, respectively.

Concerning the MF_l set, a frame-by-frame analysis of body movements is provided via the upper and lower body openness, frontal and sideways head tilt, lower, upper, head, and composite body wobble, left and right limbs relative positions, cross-limb distance and ratio, head, arms, and legs location of activity, as well as lower, upper, and full-body convex triangulation meta-features, for a total of 21 local descriptors per frame. Through these features, gait cues and relevant changes associated to the entire body structure can be captured as they happen.

Regarding the MF_g group, a condensed description of body characteristics and movements is specified via the head, arms, and legs relative movement, limp, as well as feet, legs, chest, arms, and head bones extension meta-features, for a total of 19 global descriptors. Through these features, bias towards specific body parts, which might become apparent during a walk, and detailed body conformations, describing possible limb length discrepancy, for example, can be depicted.

3.2.1. Body Openness

Body Openness (BO) is employed as devised in a previous study [36]. This meta-feature can describe both lower and upper body openness, defined BO_l and BO_u, respectively. The former, is computed using the ratio between the ankles-hips distance, and left/right knees distance; indicating whether a person has an open lower body (i.e., open legs) or has bent legs. Similarly, BO_u is calculated utilising the ratio between left/right elbows distance, and the neck-hips distance, thus capturing a broaden out chest. Intuitively, low BO_l values depict bent legs, corresponding to a crouched position; mid-range BO_l values correspond to straight yet open legs; while high BO_l values denote straight but closed legs (i.e., standing position). Comparably, low BO_u values indicate a broaden out chest; mid-range BO_u values correspond to straight arms and torso; while high BO_u values depict open elbows and bent torso. Formally, given a video sequence S, these quantities are computed for each frame $f \in S$ as follows:

$$BO_l = \frac{d(hip_{middle},\, ankles_{avg})}{d(knee_{left},\, knee_{right})}, \tag{1}$$

$$BO_u = \frac{d(neck,\ hip_{middle})}{d(elbow_{left},\ elbow_{right})}, \tag{2}$$

where $d(\cdot,\cdot)$ is the Euclidean distance; hip_{middle}, $knee_{left}$, $knee_{right}$, $neck$, $elbow_{left}$, and $elbow_{right}$, correspond to OpenPose joints 8, 13, 10, 1, 6, and 3, respectively; while $ankles_{avg}$ indicates the average (x, y) ankles position, that is, OpenPose joints 11 and 14. Summarizing, BO is a local meta-feature (i.e., $BO \in MF_l$) describing 2 quantities, that is, lower and upper body half openness.

3.2.2. Head Tilt

Head Tilt (HT) reports the degree of frontal and sideways head tilt, defined HT_{fr} and HT_s, respectively. The former, is calculated exploiting the angle between the neck-hips and nose-neck axes. The latter, is computed using the angle between the nose-neck and left-right eye axes. Intuitively, for positive HT_{fr} values, the head is facing downward, while for negative values the head is facing upward. Similarly, for positive HT_s values the head is tilted to the right side, while for negative values there is a tilt to the left side. Formally, given a video sequence S, these measures are calculated for each frame $f \in S$ as follows:

$$HT_{fr} = \frac{m_{nose-neck} - m_{neck-hip_{middle}}}{1 + m_{neck-hip_{middle}} * m_{nose-neck}}, \tag{3}$$

$$HT_s = \frac{m_{eye_{left}-eye_{right}} - m_{nose-neck}}{1 + m_{nose-neck} * m_{eye_{left}-eye_{right}}}, \tag{4}$$

where $nose$, $neck$, hip_{middle}, eye_{left}, and eye_{right}, represent the OpenPose joints 0, 1, 8, 16, 15, respectively; while the slope m of a given axis is computed using:

$$m = \tan\theta = \frac{y_2 - y_1}{x_2 - x_1}, \tag{5}$$

where x and y indicate the joint coordinates used to compute the various axes. Summarizing, HT is a local meta-feature (i.e., $HT \in MF_l$) indicating 2 measures, that is, frontal and sideways head tilt.

3.2.3. Body Wobble

Body Wobble (BW) describes whether a person has an unsteady head, upper or lower body part, defined BW_h, BW_u and BW_l measures, respectively. Moreover, this meta-feature is employed to also indicate the composite wobbling degree of a person, designated as BW_c, by accounting for the head, upper and lower halves of the body when computing this quantity. Similarly to HT, the angles between the neck-hips axis and either the left-right eye, shoulder, or hip axes, are exploited to compute these values. Moreover, the composite body wobble is computed by averaging the absolute values of head, upper and lower body wobble, to depict the general degree of body wobble. Intuitively, BW_h, BW_u and BW_l describe toward which direction the corresponding body part is wobbling during a motion, while BW_c characterizes the movement by capturing possible peculiar walks (e.g., a drunk person usually wobbles more than a sober one). Formally, given a video sequence S, BW_h, BW_u and BW_l are computed for each frame $f \in S$ as follows:

$$BW_h = \frac{m_{eye_{left}-eye_{right}} - m_{neck-hip_{middle}}}{1 + m_{neck-hip_{middle}} * m_{eye_{left}-eye_{right}}}, \tag{6}$$

$$BW_u = \frac{m_{shoulder_{left}-shoulder_{right}} - m_{neck-hip_{middle}}}{1 + m_{neck-hip_{middle}} * m_{shoulder_{left}-shoulder_{right}}}, \tag{7}$$

$$BW_l = \frac{m_{hip_{left}-hip_{right}} - m_{neck-hip_{middle}}}{1 + m_{neck-hip_{middle}} * m_{hip_{left}-hip_{right}}}, \tag{8}$$

where eye_{left}, eye_{right}, $neck$, hip_{middle}, $shoulder_{left}$, $shoulder_{right}$, hip_{left}, and hip_{right}, indicate the OpenPose joints 16, 15, 1, 8, 5, 2, 12, and 9, respectively. Finally, the composite body wobble BW_c is derived by the other body wobble meta-features as follows:

$$BW_c = \frac{|BW_h| + |BW_u| + |BW_l|}{3}. \tag{9}$$

Summarizing, BW is a local meta-feature (i.e., $BW \in MF_l$) denoting 4 values, that is, lower, upper, head, and composite body wobble.

3.2.4. Limbs Relative Position

Limbs Relative Position (LRP) indicates the opposing arm and leg relative position with respect to the neck-hips axis (i.e., a vertical axis). This meta-feature is defined for left arm/right leg and right arm/left leg pairs, named LRP_{lr} and LRP_{rl}, respectively. Intuitively, an arm and the opposing leg tend to be synchronised when walking, and oscillate together. Thus, by computing the difference between opposing limbs and the vertical neck-hips axis (i.e., their relative position), it is possible to describe whether the synchronous oscillation is happening or not. Formally, given a video sequence S, relative positions are computed for each frame $f \in S$ as follows:

$$\begin{aligned}
LRP_{lr} &= \Delta relative\ distance(neck, hips, arm_{left_{avg}}, leg_{right_{avg}}) \\
&= distance(neck, hips, arm_{left_{avg}}) - distance(neck, hips, leg_{right_{avg}}) \\
&= \frac{|(neck_y - hips_y)arm_{left_{avgx}} - (neck_x - hips_x)arm_{left_{avgy}} + neck_x hips_y - neck_y hips_x|}{\sqrt{(neck_y - hips_y)^2 + (neck_x - hips_x)^2}} \\
&- \frac{|(neck_y - hips_y)leg_{right_{avgx}} - (neck_x - hips_x)leg_{right_{avgy}} + neck_x hips_y - neck_y hips_x|}{\sqrt{(neck_y - hips_y)^2 + (neck_x - hips_x)^2}},
\end{aligned} \tag{10}$$

$$\begin{aligned}
LRP_{rl} &= \Delta relative\ distance(neck, hips, arm_{right_{avg}}, leg_{left_{avg}}) \\
&= distance(neck, hips, arm_{right_{avg}}) - distance(neck, hips, leg_{left_{avg}}) \\
&= \frac{|(neck_y - hips_y)arm_{right_{avgx}} - (neck_x - hips_x)arm_{right_{avgy}} + neck_x hips_y - neck_y hips_x|}{\sqrt{(neck_y - hips_y)^2 + (neck_x - hips_x)^2}} \\
&- \frac{|(neck_y - hips_y)leg_{left_{avgx}} - (neck_x - hips_x)leg_{left_{avgy}} + neck_x hips_y - neck_y hips_x|}{\sqrt{(neck_y - hips_y)^2 + (neck_x - hips_x)^2}},
\end{aligned} \tag{11}$$

where $neck$ and $hips$ correspond to OpenPose joints 1, and 8, respectively; while $arm_{left_{avg}}$, $arm_{right_{avg}}$, $leg_{left_{avg}}$, and $leg_{right_{avg}}$ are the average (x, y) positions of left arm, right arm, left leg, and right leg, computed using joints (5, 6, 7), (2, 3, 4), (12, 13, 14), and (9, 10, 11), respectively. Summarizing, LRP is a local meta-feature (i.e., $LRP \in MF_l$) depicting 2 measures, that is, left/right and right/left arm-leg relative position.

3.2.5. Cross-Limb Distance and Ratio

Cross-Limb Distance Ratio ($CLDR$) denotes the cross distance between left arm/right leg and right arm/left leg pairs, as well as their ratio, defined $CLDR_{lr}$, $CLDR_{rl}$, and $CLDR_r$, respectively. Similarly to the LRP meta-feature, $CLDR$ represents the synchronised oscillation of opposite limbs, although using only the average limbs position and their ratio instead of a reference axis. Intuitively, low $CLDR_{rl}$, and $CLDR_{rl}$ distances indicate a synchronous cross-limb oscillation; while high values denote an irregular movement. Concerning $CLDR_r$, the closer its value is to 1, the more synchronised is the oscillation. Indeed, through these meta-features, possible peculiar movements can be grasped during a motion. For example, arms held behind the back while walking would result in low synchronous oscillation, and could be captured via low $CLDR_{lr}$, $CLDR_{rl}$, and $CLDR_r$ values. Formally, given a video sequence S, cross-limb distances are computed for each frame $f \in S$ as follows:

$$CLDR_{lr} = d(arm_{left_{avg}},\ leg_{right_{avg}}), \tag{12}$$

$$CLDR_{rl} = d(arm_{right_{avg}},\ leg_{left_{avg}}), \tag{13}$$

where $d(\cdot,\cdot)$ is the Euclidean distance; while $arm_{left_{avg}}$, $arm_{left_{avg}}$, $leg_{left_{avg}}$, and $leg_{left_{avg}}$ are the average (x, y) positions of left arm, right arm, left leg, and right leg, computed using joints (5, 6, 7), (2, 3, 4), (12, 13, 14), and (9, 10, 11), respectively. Finally, $CLDR_r$, that is, the cross-limb distance ratio, is calculated using the following equation:

$$CLDR_r = \frac{1}{1 + CLDR_{lr} - CLDR_{rl}}. \tag{14}$$

Summarizing, $CLDR$ is a local meta-feature (i.e., $CLDR \in MF_l$) describing 3 values, that is, left/right, right/left arm-leg cross limb relative position as well as their ratio.

3.2.6. Location of Activity

Location of Activity (LOA) quantifies movement of each body component, and is computed for the head, left/right arms and legs, defined LOA_h, LOA_{al}, LOA_{ar}, LOA_{ll}, and LOA_{lr}, respectively. Intuitively, these meta-features analyse two consecutive frames to capture position changes, and directly measure the movement of a given component, so that they can determine which part is being moved the most during the motion. For example, a person with arms flailing about, would result in high LOA_{al} and LOA_{ar} values. Formally, given a video sequence S, the set of components to be analysed $C = \{h,\ al,\ ar,\ ll,\ lr\}$, and the set containing all joints of a given body part $J_c, c \in C$, the various LOC_c are computed for each pair of consecutive frames $f, f' \in S$ as follows:

$$LOA_{c,f,f'} = \Delta c_{f,f'} = \sum_{j \in J_c} \frac{d(j_f,\ j_{f'})}{|J_c|}, \tag{15}$$

where $c \in C = \{h, al, ar, ll, lr\}$ and $h,\ al,\ ar,\ ll,\ lr$ correspond to $head$, arm_{left}, arm_{right}, leg_{left}, and leg_{right} components, composed by joints (0, 1, 15, 16, 17, 18), (5, 6, 7), (2, 3, 4), (12, 13, 14), and (9, 10, 11), respectively; while $d(\cdot,\cdot)$ is the Euclidean distance. Summarizing, LOA is a local meta-feature (i.e., $LOA \in MF_l$) computing 5 values, that is, head, arms, and legs location of activity.

3.2.7. Body Convex Triangulation

Body Convex Triangulation (BCT) depicts the center of mass distribution by analysing triangle structures built between neck and wrists joints; middle hips and ankles joints; as well as head and feet joints. These meta-features, defined BCT_u, BCT_l, and BCT_f, represent the upper, lower, and full-body mass distribution, respectively. Intuitively, by analysing the triangle inner angles ratio difference, it is possible to describe whether a person is leaning left, right, or has a balanced mass distribution (i.e., positive, negative, and 0 values, respectively) in relation to the upper, lower, and full-body triangles observed during a motion. Formally, given a triangle with vertices A, B, and C, the corresponding angles $\theta_\alpha, \theta_\beta, \theta_\gamma$ are first computed as follows:

$$\theta_\alpha = \cos^{-1}\left(\frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}\right), \tag{16}$$

$$\theta_\beta = \cos^{-1}\left(\frac{\mathbf{c}\cdot\mathbf{d}}{\|\mathbf{c}\|\|\mathbf{d}\|}\right), \tag{17}$$

$$\theta_\gamma = 180 - \theta_\alpha - \theta_\beta, \tag{18}$$

where $\cdot$ is the dot product; $\|\cdot\|$ represents the vector magnitude; while $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$, indicate vectors $\overrightarrow{AB}$, $\overrightarrow{AC}$, $\overrightarrow{BA}$, and $\overrightarrow{BC}$, respectively. Then, given a video sequence S, the difference between the ratios of angles adjacent to the triangle base, and the non adjacent one, is used to compute the BCT_l, BCT_u, and BCT_f values for each frame $f \in S$, via the following equations:

$$BCT_l = \frac{\theta_{ankle_{left}}}{\theta_{hips_{middle}}} - \frac{\theta_{ankle_{right}}}{\theta_{hips_{middle}}}, \tag{19}$$

$$BCT_u = \frac{\theta_{wrist_{left}}}{\theta_{neck}} - \frac{wrist_{right}}{\theta_{neck}}, \tag{20}$$

$$BCT_f = \frac{\theta_{heel_{left}}}{\theta_{nose}} - \frac{heel_{right}}{\theta_{nose}}, \tag{21}$$

where $ankle_{left}$, $ankle_{right}$, and $hips_{middle}$, that is, OpenPose joints 14, 11, 8, correspond to the lower-body triangle; $wrist_{left}$, $wrist_{right}$, and $neck$, that is, joints 7, 4, 1, denote the upper-body triangle; while $heel_{left}$, $heel_{right}$, and $nose$, that is, joints 21, 24, 0, indicate the full-body triangle. Summarizing, BCT is a local meta-feature (i.e., $BCT \in MF_l$) denoting 3 quantities, that is, lower, upper, and full-body convex triangulation.

3.2.8. Relative Movement

Relative Movement (RM) describes the amount of change a given body part has with respect to the whole body, as conceived in Reference [36]. This meta feature is computed for the head, left/right arm, and left/right leg, defined RM_h, RM_{al}, RM_{ar}, RM_{ll}, and RM_{lr}, respectively. Intuitively, RM first analyses position and velocity changes for each body component, then computes the ratio between a single body part position (or velocity) variation amount, and the sum of position (or velocity) changes for all body components. Formally, given a video sequence S, the set of components to be analysed $C = \{h, al, ar, ll, lr\}$, and the set containing all joints of a given body part $J_c, c \in C$, the average change of a component (AC_c), over the entire recording S, is computed as follows:

$$AC_c = \frac{\sum_{f=0}^{|S|-2} \sum_{j \in J_c} \frac{|\Delta(j_f, j_{f+1})|}{|J_c|}}{|S|-2}, \tag{22}$$

where $c \in C = \{h, al, ar, ll, lr\}$ and h, al, ar, ll, lr, correspond to $head$, arm_{left}, arm_{right}, leg_{left}, and leg_{right} components, composed by joints (0, 1, 15, 16, 17, 18), (5, 6, 7), (2, 3, 4), (12, 13, 14), and (9, 10, 11), respectively. Finally, the RM_c, $c \in C$, is derived using the following equation:

$$RM_c = \frac{AC_c}{\sum_{c \in C} AC_c}. \tag{23}$$

Summarizing, RM is a global meta-feature (i.e., $RM \in MF_g$) computed for both position and velocity changes of head, arms, and legs; thus resulting in 10 distinct values describing the entire recording S.

3.2.9. Limp

Limp (L) denotes whether or not a leg is moved less than the other one when walking. This meta feature is calculated using the average velocity difference between left and right legs, over the entire video sequence. Intuitively, a limping person generally has much lower velocity in either leg. Thus, this aspect can be captured by exploiting L, where a limp in the right or left leg is denoted via a negative or positive L value. Formally, given a video sequence S, the Limp L is computed as follows:

$$L = \sum_{f=0}^{|S|-2} \sum_{j \in J_{leg_{left}}} \frac{|v(j)|}{|S|-2} - \sum_{f=0}^{|S|-2} \sum_{j' \in J_{leg_{right}}} \frac{|v(j')|}{|S|-2}, \tag{24}$$

where $J_{leg_{left}}$ and $J_{leg_{right}}$ represent the joint sets of left and right leg, defined by OpenPose joints (12, 13, 14) and (9, 10, 11), respectively; while $v(\cdot)$ is the joint velocity which can be easily computed using position variations and video frame per second (FPS). Summarizing, L is a single value global meta-feature (i.e., $L \in MF_g$) indicating limp over the whole video sequence S.

3.2.10. Bones Extension

Bones Extension (BE) describes left/right foot, left/right leg, chest, left/right arm, and head bones extension, defined BE_{fl}, BE_{fr}, BE_{ll}, BE_{lr}, BE_{ct}, BE_{al}, BE_{ar}, and BE_h respectively. Intuitively, these meta-features provide a bone size estimation by exploiting the maximum distance between the two end-points of a bone, over the entire video sequence. Formally, given a video sequence S, the set of limb bones B and the sets of joints describing the bones $J_b, b \in B$; BE_b is computed as follows:

$$BE_b = \sum_{\substack{j,j' \in J_b \\ s.t\ j \sim j'}} \max_{f \in S} d(j, j'), \tag{25}$$

where $d(\cdot,\cdot)$ is the Euclidean distance; $\sim$ identifies adjacent joints of a given bone; while BE_{fl}, BE_{fr}, BE_{ll}, BE_{lr}, BE_{ct} , BE_{al}, BE_{ar}, and BE_h denote the left foot, right foot, left leg, right leg, chest, left arm, right arm, and head, via OpenPose joint sets (19, 21), (22, 24), (12, 13, 14), (9, 10, 11), (1, 8), (5, 6, 7), (2, 3, 4), (0, 1), respectively. Summarizing, BE is global meta-feature (i.e., $BE \in MF_g$) defining bone length over feet, legs, chest, arms, and head, for a total of 8 distinct values over the entire recording S.

3.3. Bodyprint LSTM Hashing

The last step of the proposed framework, is the construction of bodyprints through the binary coding technique. In this last module, the set MF_l is analysed through a LSTM so that time variations of local meta-features can be fully captured. The LSTM output is then concatenated to the set MF_g, so that time-invariant body characteristics are merged together with time-varying ones. Finally, two dense layers are employed to implement the binary coding technique, so that a unique representation for a given person is built, ultimately allowing re-identification of that person.

3.3.1. LSTM

All local meta-features MF_l are generated for each frame of the input video sequence containing skeleton joint positions. While the proposed features provide a detailed local description of the motion (i.e., for each specific frame), they do to not account for possible time correlation between two different frames of the input sequence. To fully exploit this information, a single layer LSTM network was chosen due to its inherent ability to handle input sequences [53]. This network leverages forget gates and peep-hole connections so that non relevant information is gradually ignored, thus improving its ability to correlate both close and distant information in a given sequence. Formally, a generic LSTM cell at time t is described by the following equations:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i), \tag{26}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f), \tag{27}$$

$$\mathbf{c}_t = \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{W}b_c), \tag{28}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o), \tag{29}$$

$$\mathbf{W}_t = \sigma_t \tanh(\mathbf{c}_t), \tag{30}$$

where **i**, **f**, **o**, **c**, and **h** represent input gate, forget gate, output gate, cell activation, and hidden vectors, respectively. Moreover, $\mathbf{W}_{xi}$, $\mathbf{W}_{xf}$, $\mathbf{W}_{xo}$, and $\mathbf{W}_{xc}$ are weights connecting the various components to the input, while $\mathbf{W}_{ci}$, $\mathbf{W}_{cf}$, and $\mathbf{W}_{co}$ correspond to diagonal weights for peep-hole connections. Additionally $\mathbf{b}_i$, $\mathbf{b}_f$, $\mathbf{b}_o$, and $\mathbf{b}_c$ denote the bias associated to input gate, forget gate, output gate, and cell. Finally, the Hadamard product is used for vector multiplication. To conclude, the LSTM output $\mathbf{h}_T$, that is, the last hidden state summarizing motion characteristics of a T-length video sequence, is concatenated to MF_g, consequently producing a **z** vector of size $|\mathbf{h}_T| + |MF_g|$, representing a body motion.

3.3.2. Bodyprint Hashing

The body motion characterization **z** can be transformed into a unique identifier via binary coding hashing, so that a bodyprint is ultimately built and used for person re-identification. Following the approach in Reference [54], Supervised Deep Hashing (SDH) with a relaxed loss function is exploited to obtain a bodyprint binary code, produced by feeding the **z**-vector to two dense layers. The first layer is used to merge the concatenation of global and local meta-features, while the second one is used to obtain a binary code of dimension k. Intuitively, through SDH, similar body descriptions should result in similar codes and vice-versa, ultimately enabling a bodyprint to be used in re-identification since it is associated to a specific person.

The key aspect of this SDH approach is the relaxed loss function, where the Euclidean distance between the hash of two samples is used in conjunction with a regularizer to relax the binary constraint. This relaxation is necessary due to the sign function, usually used to obtain binary codes, leading to a discontinuous, non-differentiable and non-treatable problem via back-propagation. Formally, given two input sequences S_1 and S_2 and their corresponding bodyprints $\mathbf{b}_1, \mathbf{b}_2 \in \{-1, +1\}^k$, it is possible to define $y = 0$ in case the sequences are similar (i.e., they are derived from the same person), and $y = 1$ otherwise. Consequently, the relaxed loss L_r with respect to the two sequences is computed as follows:

$$\begin{aligned} L_r(\mathbf{b}_1, \mathbf{b}_2, y) = &\frac{1}{2}(1-y)\|\mathbf{b}_1 - \mathbf{b}_2\|_2^2 \\ &+ \frac{1}{2}y\max(m - \|\mathbf{b}_1 - \mathbf{b}_2\|_2^2,\ 0) \\ &+ \alpha(\||\mathbf{b}_1| - \mathbf{1}\|_1 + \||\mathbf{b}_2| - \mathbf{1}\|_1), \end{aligned} \tag{31}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ represent L1 and L2-norm, respectively; $\mathbf{1}$ denotes an all-ones vector; $|\cdot|$ indicates the element-wise absolute value operation; $m > 0$ is a similarity threshold; while α is a parameter modulating the regularizer weight. In detail, the first term penalises sequences generated from the same person and mapped to different bodyprints; the second term punishes sequences of different persons encoded to close binary codes, according to the threshold m; while the third term is the regularizer exploited to relax the binary constraint. Supposing there are N pairs randomly selected from the training sequences $\{S_{i,1}, S_{i,2}, y_i | i = 1, \ldots, N\}$, the resulting loss function to be minimised is:

$$\begin{aligned} \mathcal{L}_r = &\sum_{i=1}^{N}\{\frac{1}{2}(1-y_i)\|\mathbf{b}_{i,1} - \mathbf{b}_{i,2}\|_2^2 \\ &+ \frac{1}{2}y_i\max(m - \|\mathbf{b}_{i,1} - \mathbf{b}_{i,2}\|_2^2,\ 0) \\ &+ \alpha(\||\mathbf{b}_{i,1}| - \mathbf{1}\|_1 + \||\mathbf{b}_{i,2}| - \mathbf{1}\|_1\}, \\ s.t.\ &\mathbf{b}_{i,j} \in \{-1, +1\}^k,\ i \in \{1, \ldots, N\},\ j \in \{1, 2\}. \end{aligned} \tag{32}$$

While this function can be applied via the back-propagation algorithm with mini-batch gradient descent method, the subgradients of both max and absolute value operations are non-differentiable at certain points. Thus, as described in Reference [54], the partial derivatives at those points are defined to be **1** and are computed, for the three terms of Equation (32) T_1, T_2, T_3, as follows:

$$\begin{aligned}
&\frac{\partial T_1}{\partial \mathbf{b}_{i,j}} = (-1)^{j+1}(1-y_i)(\mathbf{b}_{i,1}-\mathbf{b}_{1,2}).\\
&\frac{\partial T_2}{\partial \mathbf{b}_{i,j}} = \begin{cases} (-1)^j y_i(\mathbf{b}_{i,1}-\mathbf{b}_{i,2}), & \|\mathbf{b}_{i,1}-\mathbf{b}_{i,2}\|_2^2 < m;\\ \mathbf{0}, & \textit{otherwise}. \end{cases}\\
&\frac{\partial T_3}{\partial \mathbf{b}_{i,j}} = \alpha\delta(\mathbf{b}_{i,j}),\ \delta(x) = \begin{cases} 1, & -1 \le x \le 0 \textit{ or } x \ge 1;\\ -1, & \textit{otherwise}. \end{cases}
\end{aligned} \tag{33}$$

To conclude, bodyprints can be computed by applying $sign(\mathbf{b})$, and this final representation is then exploited for person re-identification as extensively shown in the experimental section.

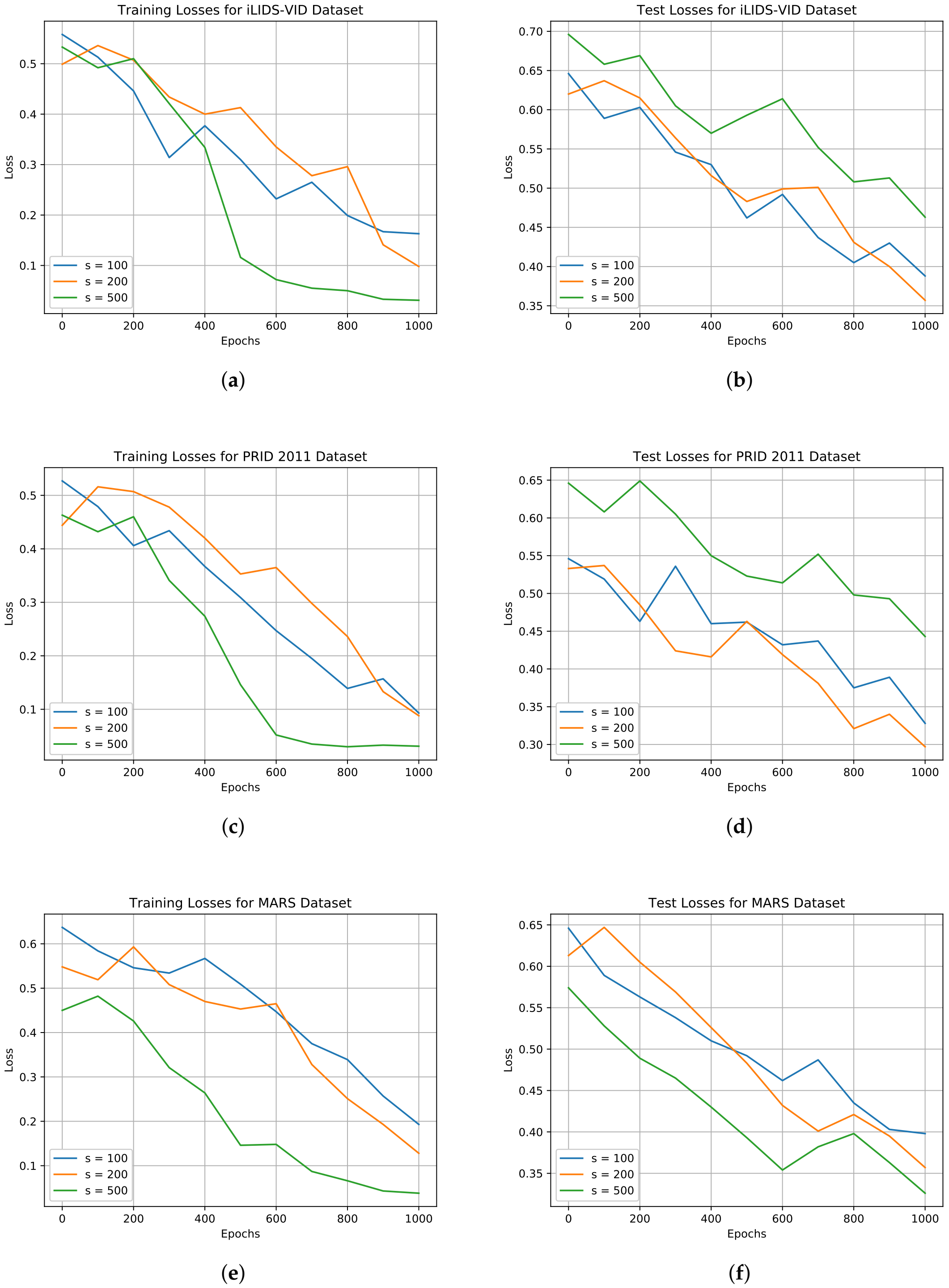
4. Results and Discussion

In this section, the results obtained with the proposed method are discussed and compared with other state-of-the-art approaches. The comparisons are performed on three public video re-identification datasets, which are discussed below.

4.1. Datasets and Settings

Concerning the datasets, the experiments were performed on iLIDS-VID [38], PRID 2011 [39], and MARS [40]. The iLIDS-VID dataset is comprised of 600 image sequences belonging to 300 people acquired by two non-overlapping cameras. Each sequence has a number of frame ranging from 23 to 192, with an average of 73. The PRID 2011 dataset consists of 400 image sequences belonging to 200 people acquired by two adjacent cameras. Each sequence has a number of frame ranging from 5 to 675, with an average of 100. Finally, the MARS dataset consists of 1261 videos acquired by 2 to 6 cameras. The entire dataset contains 20,175 tracklets, of which 3248 are distractors. The chosen datasets are challenging for the following reasons. iLIDS-VID presents clothing similarities, occlusions, cluttered background and variations across camera views. PRID 2011 main challenges are the lighting and, as for iLIDS-VID, the variations across camera views. MARS, instead, has the above mentioned challenges besides the distractors.

Concerning the model, it was implemented in Pytorch and the following parameters were used. For the LSTM input, a tensor having a shape of $[256, 50, 21]$ was employed, where 256 is the batch size, 50 are the frames used for each subject and 21 are the local meta-features MF_l. For the LSTM output a tensor having a shape of $[256, s]$ is instead utilized, where $s \in [100, 200, 500]$. The value of s was chosen empirically during the training of the model. As depicted in Figure 3, for iLIDS-VID and PRID 2011, that is, the dataset with the smaller number of identities, we used 200 as value for s to have better results over $s = 100$ and to avoid the overfitting, obtained with $s = 500$. Instead, for MARS dataset s was set to 500 due to the high number of identities. A value of s higher than 500 led to a high training time with a negligible accuracy improvement. In relation to the dense layers, the first one had an input size of $s + |MF_g|$, namely 119, 219, and 519, while the second dense layer used a dimension $k \in [16, 32, 64]$ to create the bodyprint representation. Regarding the tensor precision, since the used meta-features comprise ratios, the tensor data type was set as a 32-bit float number. For the relaxation parameter α and similarity threshold m, 0.01 and 2 were chosen, respectively. The model was trained on a NVidia RTX 2080 GPU for 1000 epochs with a learning rate of 0.001.



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 3. Train (left column) and test (right column) losses of the proposed method obtained on: iLIDS-VID, PRID 2011, and MARS datasets. Performances for these three datasets are summarized in (**a**,**b**), (**c**,**d**), and (**e**,**f**), respectively. By increasing the LSTM representational power (i.e., *s* size) convergence on the training set is reached much faster (i.e., (**a**,**c**,**e**) figures). However, due to the low number of identities in iLIDS-VID and PRID 2011 datasets, the highest *s* amount (i.e., 500) might result in overfitting scenarios (i.e., (**b**,**d**) figures).

4.2. Experiment Analysis

In Table 1, the results obtained with the proposed method are compared with current key works of the state-of-the-art. In particular, the comparison was performed with deep network based methods, namely, RNN [42], CNN + XQDA + MQ [40], Spatial and Temporal RNN (SPRNN) [43], Attentive Spatial-Temporal Pooling Network (ASTPN) [44], Deep Siamese Network (DSAN) [50], PersonVLAD + XQDA[45], VRNN + KissME [47], and Superpixel-Based Temporally Aligned Representation (STAR) [49]. Regarding the proposed method, 5 different versions were used for comparison. The $Bodyprint_{local}$ uses only local features, $Bodyprint_{global}$ uses only global features, while $Bodyprint_k$, $k \in [16, 32, 64]$, uses both local and global features but with different size of the bodyprint hashing. By first analysing the local-only and global-only version of bodyprint, it is possible to observe that for the iLIDS-VID dataset performances are consistent with the state-of-the-art. For the PRID 2011 and MARS datasets, instead, there is a noticeable drop in the performance. This can be associated with the higher number of identities and to the fact that the proposed model was designed to use synergistically local and global features. In general, we have that the local-only version of the model performs better with respect to the global-only. This is amenable to the fact that due their granularity, the local features have a better description power, while the global features can result similar for different subjects. By considering, instead, the full bodyprint model, we have that starting from the 16-bits size hashing vector, the obtained ranking can overcome many state-of-the-art works.

Table 1. Quantitative comparison between the proposed method and the current state-of-the-art methods on the chosen datasets. The best results are highlighted in bold.

Method	**iLIDS-VID**			**PRID 2011**			**MARS**		
	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20
RNN [42]	58	84	96	70	90	97	-	-	-
CNN + XQDA + MQ [40]	53	81.4	95.1	77.3	93.5	99.3	68.3	82.6	89.4
SPRNN [43]	55.2	86.5	97	79.4	94.4	99.3	70.6	90	97.6
ASTPN [44]	62	86	94	77	95	99	44	70	81
DSAN [50]	61.9	86.8	98.6	77	96.4	99.4	73.5	85	97.5
PersonVLAD + XQDA [45]	70.7	88.2	**99.2**	**88**	96.2	**99.7**	82.8	**94.9**	**99**
VRNN + KissME [47]	64.6	90.2	97.9	84.2	96.9	98.9	61.2	79.5	96.9
STAR [49]	67.5	91.7	98.8	69.2	94.9	99.1	80	89.3	95.1
Bodyprint$_{\text{local}}$	58.7	80.1	92.1	67.4	85.3	96.5	55.7	67	72.3
Bodyprint$_{\text{global}}$	56	78.4	90.6	65.9	83.2	95.5	55.4	64.9	72.2
Bodyprint$_{16}$	67.9	88.5	94.3	75.4	90.8	98	76.5	84.4	90.2
Bodyprint$_{32}$	70.3	90.1	95.6	79.5	92.3	98.7	77	87.4	93.9
Bodyprint$_{64}$	**73.4**	**94.2**	99.1	82.7	**97**	99.2	**86.5**	92.6	95.3

Concerning the iLIDS-VID and the state-of-the-art algorithm that has the best rank-1 performance on it, that is, PersonVLAD, we have that Bodyprint$_{16}$ performs 2.8% worst with respect to it. However, with higher bodyprint vector sizes, that is, 32 and 64 bits, there are an in line performance between Bodyprint$_{32}$ and PersonVLAD, and a 2.7% improvement by using the 64 bits bodyprint version. Moreover, by considering the second best algorithm, that is, ASTPN, the gain obtained with 32 and 64 bits bodyprint vectors is 8.3% and 11.4%, respectively, which is an impressive result. For the PRID 2011 dataset, we have that Bodyprint$_{16}$ and Bodyprint$_{32}$ rank-1 results are in line with the other methods. In detail, Bodyprint$_{16}$ is 4% below SPRNN (i.e., the second best performing rank-1 method on this dataset) while Bodyprint$_{32}$ is 0.1% over it. For rank-5 and rank-20, we have that both Bodyprint$_{16}$ and Bodyprint$_{32}$ are slightly below SPRNN results. Regarding Bodyprint$_{64}$, we have that it is the best algorithm at rank-5, while it is the third best result for rank-1 and rank-20, in which PersonVLAD has the best results. Finally, considering the MARS dataset, we have that at rank-1 Bodyprint$_{16}$ and Bodyprint$_{32}$ are in line with the state-of-the-art, while Bodyprint$_{64}$ substantially outperforms other literature works. Conversely, for the rank-5, we have that Bodyprint$_{64}$ is the second best method after PersonVLAD, which has obtained a score of 94.9%. For the rank-20, instead, Bodyprint$_{64}$ is in line with the other works, by obtaining a value of 95.3%. Despite the method performing generally well, there are some limitations that can influence the accuracy. These limitations are discussed in the next section.

4.3. Limitations

The proposed model presents two major limitations: occlusions and static frames (i.e., only one frame per subject). These two situations strongly influence the feature and meta-feature extraction and computation, leading to a worse performance of the model. Regarding the occlusions, we have that for the global features the average value is lowered with respect to the number of the occluded frames. For the local features, instead, we have two cases. The first case occurs when the lower part of a subject is occluded, hence only the upper local features are available. In this case, 9 local features are available. On the contrary, for the second case we have that the upper part of the subject is occluded, allowing the extraction of the lower local features only. In this case, 5 local features are available. Despite in the first case there are 4 more local features, the performance of both cases are almost identical. Since the proposed model has been designed to consider the whole body of a subject, we have that some global features cannot be computed in case of occlusions, contributing to the lowering of the performance. Concerning the static frames, we have that all the meta-features that need more than one frame to be computer are set to 0. This means that those features lose their descriptive potential and, as for the occlusions, there is a drop in the performance. In detail, when used with static frames or in sequences with a lot of occlusions the rank-1 value is around 55%.

A final remark must be made on the quality of the analysed images. Since the proposed method strongly relies on OpenPose framework, an important requirement is that the frames containing the subjects must not be too small in terms of spatial resolution. Otherwise, OpenPose will not extract all the skeleton parts, leading to the same problems encountered with occlusions. The same does not hold for data acquired with depth cameras, since the skeleton is directly provided and not estimated from RGB images. Anyway, considering the very good results of the proposed system and considering also that it is not based on visual features, thus overcoming a wide range of drawbacks of these systems, we can conclude that the system proposed in this paper can be considered a real contribute to the scientific community about this topic.

5. Conclusions

In this paper, a novel meta-feature based LSTM hashing model for person re-identification in RGB video sequences is presented. The proposed method is not based on visual features, but on meta-features extracted from the 2D skeleton models present in the scene. The meta-features are designed to catch movements, gaits, and bone proportions of a human body, thus providing an abstraction useful to overcome a wide range of drawbacks of the common competitors, including long-term re-identification and camouflage. The usefulness of the proposed method was tested on three benchmark dataset, that is, iLIDS-VID, PRID 2011, and MARS; thus demonstrating a step forward in the current literature.

Author Contributions: Conceptualization, D.A., A.F., D.P. and C.P.; methodology, D.A., A.F., D.P. and C.P.; software, A.F. and D.P.; validation, D.A., C.P.; writing—original draft preparation, D.A., A.F., D.P. and C.P.; writing—review and editing, D.A., L.C., A.F., G.L.F., D.P. and C.P.; supervision, L.C. and G.L.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This work was supported in part by the MIUR under grant "Departments of Excellence 2018-2022" of the Department of Computer Science of Sapienza University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khan, A.; Ali, S.S.A.; Anwer, A.; Adil, S.H.; Mériaudeau, F. Subsea Pipeline Corrosion Estimation by Restoring and Enhancing Degraded Underwater Images. *IEEE Access* **2018**, *6*, 40585–40601. [CrossRef]
2. Piciarelli, C.; Avola, D.; Pannone, D.; Foresti, G.L. A Vision-Based System for Internal Pipeline Inspection. *IEEE Trans. Ind. Inform.* **2019**, *15*, 3289–3299. [CrossRef]
3. Fang, X.; Guo, W.; Li, Q.; Zhu, J.; Chen, Z.; Yu, J.; Zhou, B.; Yang, H. Sewer Pipeline Fault Identification Using Anomaly Detection Algorithms on Video Sequences. *IEEE Access* **2020**, *8*, 39574–39586. [CrossRef]
4. Placidi, G.; Avola, D.; Iacoviello, D.; Cinque, L. Overall design and implementation of the virtual glove. *Comput. Biol. Med.* **2013**, *43*, 1927–1940. [CrossRef] [PubMed]
5. Avola, D.; Cinque, L.; Foresti, G.L.; Marini, M.R.; Pannone, D. VRheab: A fully immersive motor rehabilitation system based on recurrent neural network. *Multimed. Tools Appl.* **2018**, *77*, 24955–24982. [CrossRef]
6. Avola, D.; Cinque, L.; Foresti, G.L.; Marini, M.R. An interactive and low-cost full body rehabilitation framework based on 3D immersive serious games. *J. Biomed. Inform.* **2019**, *89*, 81–100. [CrossRef]
7. Cermeño, E.; Pérez, A.; Sigüenza, J.A. Intelligent video surveillance beyond robust background modeling. *Expert Syst. Appl.* **2018**, *91*, 138–149. [CrossRef]
8. Wang, Y.; Zhu, L.; Yu, Z.; Guo, B. An Adaptive Track Segmentation Algorithm for a Railway Intrusion Detection System. *Sensors* **2019** *19*, 2594. [CrossRef]
9. Ahmad, K.; Conci, N.; De Natale, F.G.B. A saliency-based approach to event recognition. *Signal Process. Image Commun.* **2018**, *60*, 42–51. [CrossRef]
10. Zhang, J.; Wu, C.; Wang, Y. Human Fall Detection Based on Body Posture Spatio-Temporal Evolution. *Sensors* **2020**, *20*, 946. [CrossRef]
11. Avola, D.; Foresti, G.L.; Cinque, L.; Massaroni, C.; Vitale, G.; Lombardi, L. A multipurpose autonomous robot for target recognition in unknown environments. In Proceedings of the IEEE International Conference on Industrial Informatics (INDIN) 2016, Poitiers, France, 19–21 July 2016; pp. 766–771.
12. Zhang, W.; Zhong, S.; Xu, W.; Wu, Y. Motion Correlation Discovery for Visual Tracking. *IEEE Signal Process. Lett.* **2018**, *25*, 1720–1724. [CrossRef]
13. Piciarelli, C.; Foresti, G.L. Drone patrolling with reinforcement learning. In Proceedings of the International Conference on Distributed Smart Cameras (ICDSC) 2019, Trento, Italy, 9–11 September 2019; pp. 1–6.
14. Avola, D.; Foresti, G.L.; Martinel, N.; Micheloni, C.; Pannone, D.; Piciarelli, C. Aerial video surveillance system for small-scale UAV environment monitoring. In Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) 2017, Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
15. Avola, D.; Cinque, L.; Fagioli, A.; Foresti, G.L.; Massaroni, C.; Pannone, D. Feature-based SLAM algorithm for small scale UAV with nadir view. In Proceedings of the International Conference on Image Analysis and Processing (ICIAP) 2019, Trento, Italy, 9–13 September 2019; pp. 457–467.
16. Avola, D.; Cinque, L.; Foresti, G.L.; Martinel, N.; Pannone, D.; Piciarelli, C. A UAV Video Dataset for Mosaicking and Change Detection From Low-Altitude Flights. *IEEE Trans. Syst. Man Cybern. Syst.* **2020**, *50*, 2139–2149. [CrossRef]
17. Hasan, M.; Paul, S.; Mourikis, A.I.; Roy-Chowdhury, A.K. Context-Aware Query Selection for Active Learning in Event Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 554–567. [CrossRef]
18. Fan, Y.; Wen, G.; Li, D.; Qiu, S.; Levine, M.D.; Xiao, F. Video anomaly detection and localization via Gaussian Mixture Fully Convolutional Variational Autoencoder. *Comput. Vis. Image Underst.* **2020**, *195*, 1–12. [CrossRef]
19. Ye, O.; Deng, J.; Yu, Z.; Liu, T.; Dong, L. Abnormal Event Detection via Feature Expectation Subgraph Calibrating Classification in Video Surveillance Scenes. *IEEE Access* **2020**, *8*, 97564–97575. [CrossRef]
20. Avola, D.; Cinque, L.; Foresti, G.L.; Massaroni, C.; Pannone, D. A keypoint-based method for background modeling and foreground detection using a PTZ camera. *Pattern Recognit. Lett.* **2017**, *96*, 96–105. [CrossRef]
21. Avola, D.; Bernardi, M.; Cinque, L.; Foresti, G.L.; Massaroni, C. Adaptive bootstrapping management by keypoint clustering for background initialization. *Pattern Recognit. Lett.* **2017**, *100*, 110–116. [CrossRef]
22. Liang, D.; Pan, J.; Sun, H.; Zhou, H. Spatio-Temporal Attention Model for Foreground Detection in Cross-Scene Surveillance Videos. *Sensors* **2019**, *19*, 5142 . [CrossRef]

23. Ammar, S.; Bouwmans, T.; Zaghden, N.; Neji, M. Deep detector classifier (DeepDC) for moving objects segmentation and classification in video surveillance. *IET Image Process.* **2020**, *14*, 1490–1501. [CrossRef]
24. Avola, D.; Cascio, M.; Cinque, L.; Fagioli, A.; Foresti, G.L.; Massaroni, C. Master and rookie networks for person re-identification. In Proceedings of the International Conference on Computer Analysis of Images and Patterns (CAIP) 2019, Salerno, Italy, 3–5 September 2019; pp. 470–479.
25. Gohar, I.; Riaz, Q.; Shahzad, M.; Zeeshan Ul Hasnain Hashmi, M.; Tahir, H.; Ehsan Ul Haq, M. Person Re-Identification Using Deep Modeling of Temporally Correlated Inertial Motion Patterns. *Sensors* **2020**, *20*, 949. [CrossRef] [PubMed]
26. Almasawa, M.O.; Elrefaei, L.A.; Moria, K. A Survey on Deep Learning-Based Person Re-Identification Systems. *IEEE Access* **2019**, *7*, 175228–175247. [CrossRef]
27. Leng, Q.; Ye, M.; Tian, Q. A Survey of Open-World Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1092–1108. [CrossRef]
28. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-guided contrastive attention model for person re-identification. In Proceedings of the International IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1179–1188.
29. Zhou, S.; Wang, F.; Huang, Z.; Wang, J. Discriminative feature learning with consistent attention regularization for person re-identification. In Proceedings of the International IEEE/CVF International Conference on Computer Vision (ICCV) 2019, Seoul, Korea, 27–28 October 2019; pp. 8039–8048.
30. Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; Yang, Y. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the International IEEE/CVF International Conference on Computer Vision (ICCV) 2019, Seoul, Korea, 27–28 October 2019; pp. 542–551.
31. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. VRSTC: Occlusion-free video person re-identification. In Proceedings of the International IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019, Long Beach, CA, USA, 15–21 June 2019; pp. 7176–7185.
32. Li, J.; Zhang, S.; Wang, J.; Gao, W.; Tian, Q. Global-local temporal representations for video person re-identification. In Proceedings of the International IEEE/CVF International Conference on Computer Vision (ICCV) 2019, Seoul, Korea, 27–28 October 2019; pp. 3957–3966.
33. Huang, Y.; Xu, J.; Wu, Q.; Zhong, Y.; Zhang, P.; Zhang, Z. Beyond Scalar Neuron: Adopting Vector-Neuron Capsules for Long-Term Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [CrossRef]
34. Avola, D.; Bernardi, M.; Cinque, L.; Foresti, G.L.; Massaroni, C. Exploiting Recurrent Neural Networks and Leap Motion Controller for the Recognition of Sign Language and Semaphoric Hand Gestures. *IEEE Trans. Multimed.* **2019**, *21*, 234–245. [CrossRef]
35. Avola, D.; Cascio, M.; Cinque, L.; Foresti, G.L.; Massaroni, C.; Rodolà, E. 2D Skeleton-Based Action Recognition via Two-Branch Stacked LSTM-RNNs. *IEEE Trans. Multimed.* **2019**. [CrossRef]
36. Avola, D.; Cinque, L.; Fagioli, A.; Foresti, G.L.; Massaroni, C. Deep Temporal Analysis for Non-Acted Body Affect Recognition. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
37. Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; Sheikh, Y.A. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell. arXiv* **2018**, arXiv:1812.08008.
38. Wang, T.; Gong, S.; Zhu, X.; Wang, S. Person re-identification by video ranking. In Proceedings of the European Conference on Computer Vision (ECCV) 2014, Zurich, Switzerland, 6–12 September 2014; pp. 688–703.
39. Hirzer, M.; Beleznai, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In Proceedings of the Scandinavian Conference on Image Analysis (SCIA) 2011, Ystad, Sweden, 23–27 May 2011; pp. 91–102.
40. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. MARS: A video benchmark for large-scale person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV) 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 868–884.
41. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by Local Maximal Occurrence representation and metric learning. In Proceedings of the International IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
42. McLaughlin, N.; Martinez del Rincon, J.; Miller, P. Recurrent convolutional network for video-based person re-identification. In Proceedings of the International IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 1325–1334.

43. Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; Tan, T. See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification. In Proceedings of the International IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 6776–6785.
44. Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; Zhou, P. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In Proceedings of the International IEEE/CVF International Conference on Computer Vision (ICCV) 2017, Venice, Italy, 22–29 October 2017; pp. 1–10.
45. Wu, L.; Wang, Y.; Shao, L.; Wang, M. 3-D PersonVLAD: Learning Deep Global Representations for Video-Based Person Reidentification. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3347–3359. [CrossRef]
46. Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; Russell, B. ActionVLAD: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 3165–3174.
47. Wu, L.; Wang, Y.; Yin, H.; Wang, M.; Shao, L. Few-Shot Deep Adversarial Learning for Video-Based Person Re-Identification. *IEEE Trans. Image Process.* **2020**, *29*, 1233–1245. [CrossRef]
48. Chung, J.; Kastner, K.; Dinh, L.; Goel, K.; Courville, A.; Bengio, Y. A recurrent latent variable model for sequential data. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS) 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 2980–2988.
49. Gao, C.; Wang, J.; Liu, L.; Yu, J.G.; Sang, N. Superpixel-Based Temporally Aligned Representation for Video-Based Person Re-Identification. *Sensors* **2019**, *19*, 3861. [CrossRef]
50. Wu, L.; Wang, Y.; Gao, J.; Li, X. Where-and-When to Look: Deep Siamese Attention Networks for Video-Based Person Re-Identification. *IEEE Trans. Multimed.* **2019**, *21*, 1412–1424. [CrossRef]
51. Nguyen, T.N.; Huynh, H.H.; Meunier, J. Skeleton-based abnormal gait detection. *Sensors* **2016**, *16*, 1792. [CrossRef]
52. Nambiar, A.; Bernardino, A.; Nascimento, J.C. Gait-based Person Re-identification: A Survey. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–34. [CrossRef]
53. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
54. Liu, H.; Wang, R.; Shan, S.; Chen, X. Deep supervised hashing for fast image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 2064–2072.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).