Pattern Recognition Letters 145 (2021) 23–29

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Person re-identification: Implicitly defining the receptive fields of deep learning classification frameworks

Ehsan Yaghoubi[a,*], Diana Borza[b], S.V. Aruna Kumar[c], Hugo Proença[a]

[a] IT: Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal
[b] Technical University of Cluj-Napoca, Romania
[c] University of Beira Interior, Covilhã, Portugal

ARTICLE INFO

Article history:
Received 9 June 2020
Revised 29 November 2020
Accepted 19 January 2021
Available online 3 February 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Person re-identification
Data augmentation
Explicit attention mechanism
Visual surveillance

ABSTRACT

The *receptive fields* of deep learning models determine the most significant regions of the input data for providing correct decisions. Up to now, the primary way to learn such receptive fields is to train the models upon masked data, which helps the networks to ignore any unwanted regions, but also has two major drawbacks: (1) it yields edge-sensitive decision processes; and (2) it augments considerably the computational cost of the inference phase. Having theses weaknesses in mind, this paper describes a solution for implicitly enhancing the inference of the networks' receptive fields, by creating synthetic learning data composed of interchanged segments considered *apriori* important or irrelevant for the network decision. In practice, we use a segmentation module to distinguish between the foreground (important) versus background (irrelevant) parts of each learning instance, and randomly swap segments between image pairs, while keeping the class label exclusively consistent with the label of the segments deemed important. This strategy typically drives the networks to interpret that the identity and clutter descriptions are not correlated. Moreover, the proposed solution has other interesting properties: (1) it is parameter-learning-free; (2) it fully preserves the label information; and (3) it is compatible with the data augmentation techniques typically used. In our empirical evaluation, we considered the person re-identification problem, and the well known RAP, Market1501 and MSMT-V2 datasets for two different settings (*upper-body* and *full-body*), having observed highly competitive results over the state-of-the-art. Under a reproducible research paradigm, both the code and the empirical evaluation protocol are available at https://github.com/Ehsan-Yaghoubi/reid-strong-baseline.

1. Introduction

Person re-identification (re-id) refers to the cross-camera retrieval task, in which a query from a target subject is used to retrieve identities from a gallery set. This process is tied to many difficulties, such as variations in human pose, illumination, partial occlusion, and cluttered background. The primary way to address these challenges is to provide large-scale *labeled* learning data (which are not only hard to collect, but particularly costly to annotate) and expect that the deep model learns the critical parts of the input data autonomously. This strategy is supposed to work for any problem, upon the existence of enough learning data, which might correspond to millions of learning instances in hard problems.

To skim the costly annotation step, various works propose to augment the learning data using different techniques [1]. They either use the available data to synthesize new images or generate new images by sampling from the learned distribution. In both cases, the main objective is to increase the quantity of data, without assisting the model in finding the input regions, so that often the networks find spurious patterns in the background regions that –yet– are matched with the ground truth labels. This kind of techniques shows positive effects in several applications; for example, Dvornik et al. [2] proposes an object detection model, in which the objects are cut out from their original background and pasted to other scenes (e.g., a plane is pasted between different sky images). On the contrary, in the pedestrian attribute recognition and re-identification problems, the background clutter is known as a primary obstacle to the reliability of the inferred models.

Holistic CNN-based re-id models extract global features, regardless of any critical regions in the input data, and typically fail when the background covers most of the input. In particular, when deal-

[*] Corresponding author.
E-mail address: Ehsan.yaghoubi@ubi.pt (E. Yaghoubi).

https://doi.org/10.1016/j.patrec.2021.01.035
0167-8655/© 2021 Elsevier B.V. All rights reserved.

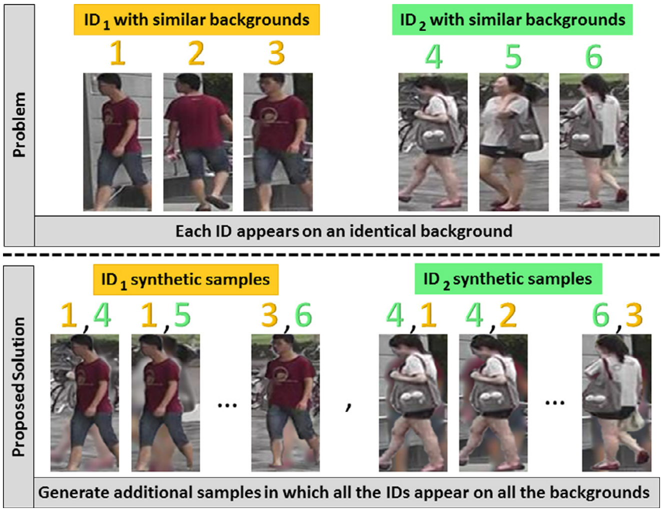
Problem
ID 1 with similar backgrounds
ID 2 with similar backgrounds
1
2
3
4
5
6
Each ID appears on an identical background
Proposed Solution
ID 1 synthetic samples
ID 2 synthetic samples
1,4
1,5
3,6
4,1
4,2
6,3
Generate additional samples in which all the IDs appear on all the backgrounds

Fig. 1. The main challenge addressed in this paper: during the learning phase, if the model sees all samples of one ID in a single scene, the final feature representation of that subject might be entangled with spurious (background) features. By creating synthetic samples with multiple backgrounds, we implicitly *guide* the network to focus on the deemed important (foreground) features.

ing with limited amounts of learning data, three problems emerge: (1) holistic methods may not find the foreground regions automatically; (2) part-based methods [3], Li et al. [4] typically fail to detect the appropriate critical regions; and (3) attention-based models (e.g., Xu et al. [5] and Zhao et al. [6]) face difficulties in case that multiple persons appear in a single bounding box. As an attempt to reduce the classification bias due to the background clutter (caused by inaccurate person detection or crowded scenes), Zheng et al. [7] proposes an alignment method to refine the bounding boxes, while [8] uses a local feature matching technique. As illustrated in Fig. 1, although the alignment-based re-id approaches reduce the amounts of clutter in the learning data, the networks still typically suffer from the remaining background features, particularly if some of the IDs always appear in the same scene (background).

To address the above-described problems, this paper introduces a receptive field implicit definition method based on data augmentation that could be applied to the existing re-id methods as a complementary step. The proposed solution is (1) mask-free for the *test* phase, i.e., it does not require any additional explicit segmentation in test time; and (2) contributes to foreground-focused decisions in the inference phase. The main idea is to generate synthetic data composed of interleaved segments from the original learning set, while using class information only from specific segments. During the learning phase, the newly generated samples feed the network, keeping their label exclusively consistent with the identity from where the region-of-interest was cropped. Hence, as the model receives images of each identity with inconsistent unwanted areas (e.g., background), it naturally pays the most attention to the regions consistent with ground truth labels. We observed that this pre-processing method is equivalent to only learn from the effective receptive fields and ignore the destructive regions. During the test phase, samples are provided without any mask, and the network naturally disregards the detrimental information, which is the insight for the observed improvements in performance.

In particular, when compared to Inoue [9] and Zhong et al. [10], this paper can be seen as a data augmentation technique with several singularities: (1) we not only enlarge the learning data but also implicitly provide the inference model with an attentional decision-making skill, contributing to *ignore* irrelevant image features during the test phase; (2) we generate highly representative samples, making it possible to use our solution along with other data augmentation methods; and (3) our solution allows the on-the-fly data generation, which makes it efficient and easy to be implemented beside the common data augmentation techniques. Our evaluation results point for consistent improvements in performance when using our solution over the state-of-the-art person re-id method.

2. Related work

2.1. Data augmentation

Data augmentation targets the root cause of the over-fitting problem by generating new data samples and preserving their ground truth labels. *Geometrical transformation* (scaling, rotations, flipping, etc.), *color alteration* (contrast, brightness, hue), *image manipulation* (random erasing [10], kernel filters, image mixing [9]), and *deep learning approaches* (neural style transfer, generative adversarial networks) [1] are the common augmentation techniques.

Recently, various methods have been proposed for image synthesizing and data augmentation [1]. For example, Inoue [9] generates n^2 samples from an n-sized dataset by using a sample pairing method, in which a random couple of images are overlaid based on the average intensity values of their pixels. Zhong et al. [10] presents a *random erasing* data augmentation strategy that inflates the learning data by randomly selecting rectangular regions and changing their pixels values. As an attempt to robustify the model against occlusions, increasing the volume of the learning data turned the concept of *random erasing* into a popular data augmentation technique. Dvornik et al. [2] addressed the problem of object detection, in which the background has helpful features for detecting the objects; therefore, authors developed a context-estimator network that places the instances (i.e., cut out objects) with meaningful sizes on the relevant backgrounds.

2.2. Person re-ID

. In general, early person re-id works studied either the descriptors to extract more robust feature representations or metric-based methods to handle the distance between the inter-class and intra-class samples [11]. However, recent re-id studies are mostly based on deep learning neural networks that can be classified into three branches [12]: Convolutional Neural Network (CNN), CNN-Recurrent neural network, and Generative Adversarial Network (GAN).

Among the CNN and CNN-RNN methods, those based on attention mechanisms follow a similar objective to what we pursue in this paper; i.e., they ignore background features by developing attention modules in the backbone feature extractor. Attention mechanism may be developed for either single-shot or multi-shot (video) [13], Zhang et al. [14], Cheng et al. [15] scenarios, both of them aim to learn a distinctive feature representation that focuses on the critical regions of the data. To this end, Yang et al. [16] use the body-joint coordinates to remove the extra background and divide the image into several horizontal pieces to be processed by separate CNN branches. Xu et al. [5] and Zhao et al. [6] propose a body-part detector to re-identify the probe person with matching the bounding boxes of each body-part, while [17] uses the masked out body-parts to ignore the background features in the matching process. In contrast to these works that explicitly implement the attentional process in the structure of the neural network [18], we provide an attentional control ability based on receptive field augmentation detailed in Section 3. Therefore, in some terms, our work is similar to the GAN-based re-id techniques, which usually aim to either increase the quantity of the data [19] or present novel poses of the existing identities [20], Borgia et al. [21] or

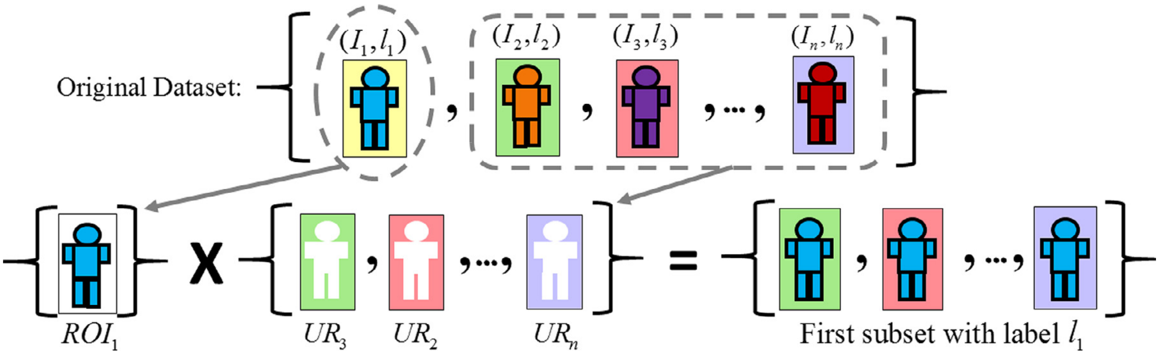

Fig. 2. The proposed full-body attentional data augmentation (best viewed in color). Blue, orange, purple, and red denote the samples 1, 2, 3, and *N*, respectively. The pale-yellow, green, pink, and purple colors represent their cluttered (background) regions, which should be irrelevant for the inference process. Therefore, all the synthetic images labeled as 1 share the blue body region but have different backgrounds, which provides a strong cue for the network to disregard such segments from the decision process.

transfer the camera style [22], Wei et al. [23]. Although GAN-based works present novel features for each individual, they generate some destructive features that are originated from the new backgrounds. Furthermore, these works ignore to handle the problem of co-appearance of multiple identities in one shot.

3. Proposed method

Fig. 2 provides an overview of the proposed image synthesis method, in this case, considering the full-body as the region of interest (ROI). We show the first synthesize subset, in which the new samples comprise of the *ROI* of the 1st sample and the background of the other samples.

3.1. Implicit definition of receptive rields

As an intrinsic behavior of CNNs, in the learning phase, the network extracts a set of essential features in accordance with the image annotations. However, extracting relevant and compressed features is an ongoing challenge, especially when the background[1] changes with person ID. Intuitively, when a person's identity appears with an identical background, some background features are entangled with the useful foreground features and reduce the inference performance. However, if the network sees one person with different backgrounds, it can automatically discriminate between the relevant regions of the image and the ground truth labels. Therefore, to help the *inference model* automatically distinguish between the unwanted features and foreground features, in the *learning phase*, we repeatedly feed synthetically generated, fake images to the network that has been composed of two components: (i) critical parts of the current input image that describe the ground truth labels (i.e., person's identity), and we would like to have an attention on them, and (ii) parts of the other real samples that intuitively are uncorrelated with the current identity –i.e., background and possible body parts (if any) that we would like the network to ignore them. Thus, the model looks through each region of interest, juxtaposed with different unwanted regions –of all the images– enabling the network to learn where to look at in the image and ignores the parts that are changing arbitrarily and are not correlated with ground truth labels. Consequently, during the test phase, the model explores the region of interest and discards the features of unwanted regions that have been trained for.

Formally, let $\boldsymbol{I}_i$ represent the *i*th image in the learning set, $\boldsymbol{l}_i$ its ground truth label (ID) and $\boldsymbol{M}_j$ the corresponding ground-truth binary mask that discriminates between the foreground/background regions. As the available re-id datasets do not provide ground-truth human body masks, we use the Mask R-CNN [24] to obtain such masks (see Section 4). Considering that $\boldsymbol{ROI}_.$ refers the region of interest and $\boldsymbol{UR}_.$ the unwanted regions, the goal is to synthesis the artificial sample $\boldsymbol{S}_{i\neg j}$, using label $\boldsymbol{l}_i$: $\boldsymbol{S}_{i\neg j}(x, y) = \boldsymbol{ROI}_i \cup \boldsymbol{UR}_j$, where $\boldsymbol{ROI}_. = \boldsymbol{I}_.(x, y)$ such that $\boldsymbol{M}_.(x, y) = 1$, $\boldsymbol{UR}_. = \boldsymbol{I}_.(x, y)$ such that $\boldsymbol{M}_.(x, y) = 0$, and (x, y) are the coordinates of the pixels.

Therefore, for an *n*-sized dataset, the *maximum* number of generated images is equal to $n^2 - n$. However, to avoid losing the expressiveness of the generated samples, we consider several constraints. Hence, a combination of the common data transformations (e.g., flipping, cropping, blurring) can be used along with our method. Obviously, since we utilize the ground truth masks, our technique should be used in the first place, before any other augmentation transformation, to avoid extra processing on the binary masks.

3.2. Synthetic image generation

To ensure that the synthetically generated images have a natural aspect, we impose the following constraints:

3.2.1. Size and shape constraint

Considering that human bodies are deformable objects of varying size and alignment within the bounding boxes, any blind image generation process will yield unrealistic results. Therefore, we added a constraint that avoids combining images with significant differences in their aspect ratios of the ROIs to circumvent the unrealistic stretching/shrinking of the replaced content in the generated images. To this end, the ratio between the foreground areas defined by masks $\boldsymbol{M}_j$ and $\boldsymbol{M}_i$ should be more than the threshold $\boldsymbol{T}_s$ (we considered $\boldsymbol{T}_s = 0.8$ in our experiments). Let $\boldsymbol{A}_.$ be the area of the foreground region (i.e., mask $\boldsymbol{M}_.$):$\boldsymbol{A}_j = \sum_{x=0}^{w} \sum_{y=0}^{h} M_j(x, y)$, where w and h are the width and height of the image, respectively.

This constraint translates to $\min(\boldsymbol{A}_i, \boldsymbol{A}_j)/\max(\boldsymbol{A}_i, \boldsymbol{A}_j) > \boldsymbol{T}_s$. Moreover, to ensure the shape similarity, we calculate the Intersection over Union metric (IoU) for masks $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$: $IoU(\boldsymbol{M}_i, \boldsymbol{M}_j) = (\boldsymbol{M}_i \cap \boldsymbol{M}_j)/(\boldsymbol{M}_i \cup \boldsymbol{M}_j)$.

For the IoU calculation, we ought to consider only the rectangular area around the masks (instead of the whole image area); moreover, when calculating the IoU, the size of the masks must match, and in case of resizing the masks, the aspect ratios should be preserved. To fulfill these conditions, we find the contours in the binary masks using [25] and calculate the minimal up-right bounding rectangle of the masks. The width of the rectangular

[1] The terms *(unwanted region/region-of-interest)*, *(undesired/desired) boundaries*, *(background/foreground) areas*, and *(unwanted/wanted) areas* refer to the data segments that are deemed to be irrelevant/relevant to the ground truth label. For example, in a hair color recognition problem, the region-of-interest is the hair area, which can be defined by a binary mask.

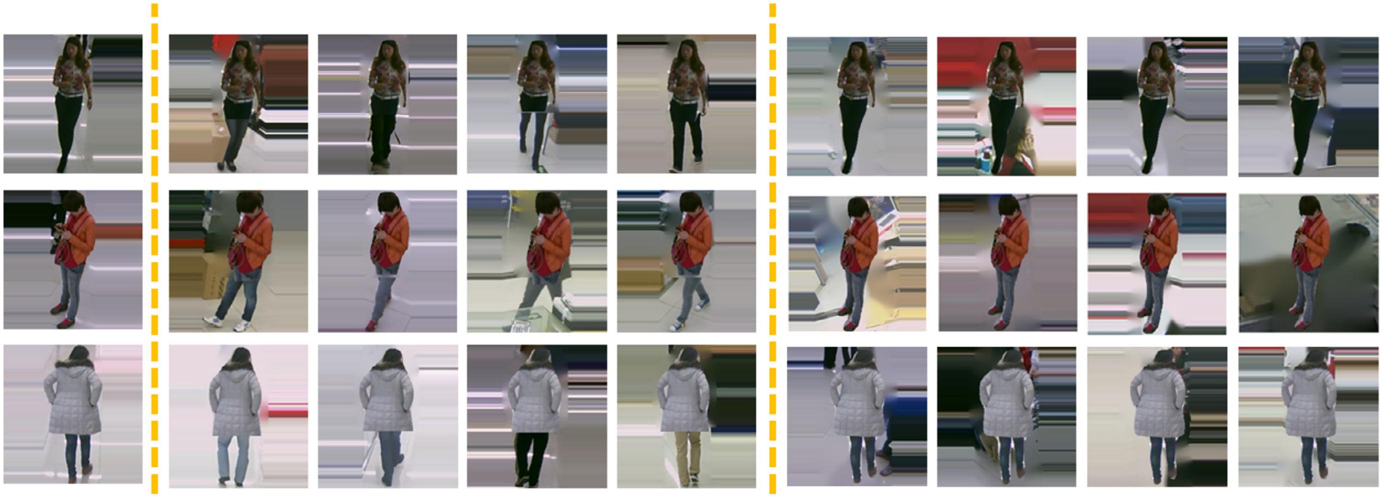

Fig. 3. Examples of synthetic data generated for upper-body (center columns) and full-body (rightmost columns) receptive fields. The leftmost column shows the original images. Additional examples are provided at https://github.com/Ehsan-Yaghoubi/reid-strong-baseline.

masks in all images is set to a fixed size and, afterwards, we apply zero padding to the height of the smaller mask to match the sizes. Finally, if the $IoU(\boldsymbol{M}_i, \boldsymbol{M}_j)$ is higher than a threshold $\boldsymbol{T}_i$, we consider those images for the merging process ($\boldsymbol{T}_i = 0.5$ was used in our experiments).

3.2.2. Smoothness constraint

The transition between the source image and the replaced content should be as smooth as possible to prevent from strong edges. One challenge is that $\boldsymbol{M}_i$ and the body silhouette of the jth person do not match perfectly. To overcome this issue, we enlarge the mask $\boldsymbol{M}_j$ by using the morphological dilation operator with a 5 × 5 kernel: $\boldsymbol{M}_d = \boldsymbol{M}_j \oplus K_{5\times5}$. Next, to guarantee the continuity between the background and the newly added content, we use the image in-painting technique in Telea [26] to remove the undesired area from the source image, as it has been dictated by the enlarged mask $\boldsymbol{M}_d$.

3.2.3. Viewpoint constraint

The proposed method can be used for focusing on a specific region of the body. For example, supposing that the upper-body should be considered the RoI, the generated images will be composed of the 1st sample's upper-body and the remaining segments (background and lower-body regions) of the other images, while keeping the label of the 1st sample. When defining the receptive fields of specific regions (e.g., upper body in Fig. 3), it is important to generate high representative samples. Hence, we consider the body poses of samples and only combine images with the same viewpoint annotations causing to prevent from generating images composed of the anterior upper-body of the ith person and posterior lower-body (and background) of the jth person. One can apply Alphapose [27] to any pedestrian dataset to estimate the body poses and then, uses a clustering method such as [28], MacQueen et al. [29], Sculley [30], or Ng et al. [31] to create clusters of poses as the viewpoint label.

The detailed information for the two experiments carried out is given in Section 5.3. Fig. 3 shows some examples generated by our technique, providing attention to the upper-body or full-body region. When defining the CNN's receptive fields on the upper-body region, fake samples are different in the human lower body and the environment, while they resemble each other in the person's upper body and identity label. By selecting the full-body as the ROI, the generated images will be composed of similar body silhouettes with different surroundings.

4. Implementation details

As the settings and configurations on all the datasets are identical, in the following we only mention the details for the RAP dataset. We based our method on the baseline [32] and selected similar model architecture, parameter settings, and optimizer. In this baseline, authors resized images on-the-fly into 128 × 128 pixels. As the RAP images vary in resolution (from 33 × 81 to 415 × 583), to avoid any data deformation, we first mapped the images to a squared shape, using a *replication* technique, in which the row or column at the very edge of the original image is replicated to the extra border of the image.

The RAP dataset does not provide human body segmentation annotations. To generate the segmentation masks, we first fed the images to Mask-R-CNN model [24] (using its default parameter settings described in https://github.com/matterport/Mask_RCNN). Next, as described in Section 3.2, we generated the synthetic images.

To provide the train and test splits for our model, we followed the instructions of the dataset publishers in Wei et al. [23], Li et al. [33], Zheng et al. [34]. Furthermore, following the configurations suggested in Luo et al. [32], we used the state-of-the-art tricks such as warm-up learning rate [35], random erasing data augmentation [10], label smoothing [36], last stride [37], and BN-Neck [32], alongside the conventional data augmentation transformations (i.e., random horizontal flip, random crop, and 10-pixel-padding and original-size-crop).

5. Experiments and discussion

We evaluate the proposed method under two settings: (1) by defining the upper-body receptive fields, assuming that most of the identity information lies in upper body. In this setting, we generate the synthetic data by modifying the lower-body parts of the subject images. This setting requires both segmentation masks and viewpoint annotations, as the perspective/viewpoint of the upper-body region should be consistent with the perspective of the lower body. In practice, this strategy assures that we do not combine a front-view upper body with a rear-view lower body. (2) by defining the full-body receptive fields, in which the attention of the network is “oriented” towards the entire body. The notion of viewpoint does not apply here, since the method can be seen as a simple background swapping process, where the person is placed in a different environment. In our experiments, we evaluate our model on the earlier setting and RAP dataset for two modes: (a) when human-based annotations are available for four viewpoints, and (b) when the subjects' viewpoint is inferred using a clustering method. Furthermore, we tested our method with the later setting over the RAP, Market1501, and MSMT17 datasets.

5.1. Datasets

The *Richly Annotated Pedestrian* (RAP) benchmark [33] is one of the largest well-known pedestrian dataset composing of around 85,000 samples, from which 41,585 images have been selected manually for identity annotation. The RAP re-id set includes 26,638 images of 2589 identities and 14,947 samples as distractors that have been collected from 23 cameras in a shopping mall. The provided human bounding boxes have different resolutions ranging from 33 × 81 to 415 × 583. In addition to human attributes, the RAP dataset is annotated for camera angle, body-part position, and occlusions. The MSMT17-V2 re-id dataset [23] consists of 4101 identities captured with 15 cameras in outdoor and indoor environment. The total number of person bounding boxes are 126,441 which have been detected using Faster RCNN [38]. The Market1501 dataset [34] used the Deformable Part Model (DPM) detector [39] to extract 32,668 person bounding boxes from 1105 identities using 6 cameras in outdoor scenes. The Market1501 dataset images were normalized to 128 × 64 pixel resolution.

5.2. Baseline

A recent work by Facebook AI [40] mentions that upgrading factors such as the learning method (e.g., Roth et al. [41], Kim et al. [42]), network architecture (e.g., ResNet, GoogleNet, BN-Inception), loss function (e.g., embedding losses [43], Wang et al. [44] and classification losses [45], Qian et al. [46]), and parameter settings may improve the performance of an algorithm, leading to unfair comparison. This way, to be certain that the proposed solution actually contributes to performance improvement, our empirical framework was carefully designed in order to keep constant as many factors as possible with a recent re-id baseline [32] This baseline has advanced the state-of-the-art performance with respect to several techniques such as Kalayeh et al. [47], Zhong et al. [48], and Li et al. [49]. In summary, it is a holistic deep learning-based framework that uses a bag of tricks that are known to be particularly effective for the person re-id problem. Authors employ the ResNet-50 model as the backbone feature extractor

5.3. Re-ID results

5.3.1. Experiments on the RAP dataset

As stated before, the proposed method with the upper-body setting requires viewpoint labels; however, not all pedestrian datasets provide this ground truth information. As annotating a large dataset with this information would be extremely time consuming, we suggested that state of the art pose detectors are used to automatically infer the subjects viewpoint. To test this hypothesis, we have chosen the RAP dataset since it includes manual annotations for the samples viewpoint. Hence, we evaluated our upper-body-based model for two different modes: (1) by considering the human-based viewpoint annotations; and (2) by using Alphapose followed by a clustering method (Balanced Iterative Reducing and Clustering using Hierarchies [28]) to automatically estimate human poses. In the latter case, we used Alphapose with its default settings to extract the body key-points of all the persons in the dataset; next, we applied the BIRCH clustering method and created 8 clusters of body poses. Finally, to swap the unwanted regions in the original image with another sample, the candidate image is selected from the same cluster where the original image is located. In both modes, the network configuration and the hyper-parameters were exactly the same.

Table 1 provides the overall performances based on the mean Average Precision (mAP) metric and Cumulative Match Characteristic (CMC) for ranks 1, 5, and 10, denoting the possibility of retrieving at least one true positive in the top-1, 5, and 10 ranks. We evaluated the proposed method using two sampling methods and observed a slight improvement in the performance of both methods when using the *triplet-softmax* over *softmax* sampler. As previously mentioned, our method could be treated as an augmentation method that requires a paired-process (i.e., exchanging the foreground and background of each pair of images), imposing a computational cost only to the *learning phase*. Moreover, due to increasing the learning samples from n to less than n^2, the network needs more time and the number of epochs to converge. Therefore, learning our method (using *triplet-softmax* sampler) for 280 epochs lasted around 20 h with loss value 1.3, while the baseline method completed 2000 epochs after 37 h of learning with loss value 1.0.

The experimental results of upper-body setting are given in rows 2 and 3 of Table 1, pointing for an optimal performance, when we use 8 cluster of poses instead of the ground truth viewpoint labels; therefore, our method could be used in conjunction with viewpoint estimation models to boost the performance, without requiring viewpoint annotations.

Comparison of the first and second rows of Table 1 shows that our technique with an attention on the human upper-body achieves competitive results, such that retrieval accuracy in rank 1 is 0.3% better than the baseline. However, in higher ranks and mAP metrics, the baseline has better performance.

The fourth row of Table 1 provides the performance of the proposed method with an attention on the human full-body and –not surprisingly– indicates that concentration on the full-body (rather than upper-body) yields more useful features for short-time person re-id. However, comparing four rows of the result table together, we could perceive how much is the lower-body important –as a body-part with most background region? For example, when using full-body region (over the upper-body) with *triplet-softmax* sampler, the rank 1 accuracy improves from 66.8 to 69.0 (i.e., 2.2% improvement), while the accuracy difference of rank 1 between the holistic baseline and full-body method is 2.9%, indicating that 2.2 of our improvement (in rank 1) over the baseline is because of attention on the lower-body and the rest (0.7%) is due to focusing on the upper-body.

During the learning phase, each synthesized sample is generated with a probability between $[0, 1]$, with 0 meaning that no changes will be done in the dataset (i.e., we use the original samples) and 1 indicates that all samples will be transformed (augmented). We studied the effectiveness of our method for different probabilities (from 0.1 to 0.9) and gave the obtained results in Table 2. Overall, the optimal performance of the proposed technique is attained when the augmentation probability lies in the $[0.3, 0.5]$ interval. This leads us to conclude that such intermediate probabilities of augmentation keep the discriminating information of the original data while also guarantee the transformation of enough data for yielding an effective attention mechanism.

5.3.2. Experiments on the Market1501 dataset

Table 3 compares the performance of our method with respect to several state-of-the-art techniques on the Market1501 set [34], and supports the superiority of our method, with 0.4 of rank 1 accuracy over [50] and 1.1 of mAP over [51]. Additionally, we post-processed our results on the Market1501 using the re-ranking method proposed by Zhong et al. [52]. Zhong et al. [52] post-processes the global features of the gallery set and the probe per-

Table 1
Results comparison between the baseline (top row) and our solutions for defining receptive fields, particularly tuned for the *upper body* and *full body*, on the RAP benchmark. mAP and Ranks 1, 5, and 10 are given, for the *softmax* and *triplet-softmax* samplers. The best results per performance measure appear in bold.

	Softmax sampler				Triplet-softmax sampler			
Model	rank = 1	rank = 5	rank = 10	mAP	rank = 1	rank = 5	rank = 10	mAP
Luo et al. [32]*	64.1	81.5	86.8	**45.8**	66.1	81.9	86.3	45.9
Ours (setting 1, mode 1: upper body with viewpoint annotations)	64.4	80.5	85.6	42.5	66.5	81.5	86.0	43.0
Ours (setting 1, mode 2: upper body without viewpoint annotations)	65.1	81.4	86.2	43.3	66.8	82.0	86.5	43.8
Ours (setting 2: full body)	**65.7**	**82.2**	**87.2**	45.0	**69.0**	**83.6**	**88.1**	**46.3**

*The best possible results occurred for *triplet-softmax* sampler in epoch 1120. The other models were trained for 280 epochs which lasted around 20 h.

Table 2
Results of the proposed receptive field definer solution for upper-body and full-body models. Bold and underline styles denote the best and runner-up results. “Aug. Prob.” stands for *augmentation probability*.

Model	Aug. Prob.	Rank 1	Rank 5	Rank 10	Rank 50	mAP
Upper-body	0.1	53.4	72.3	78.9	90.6	34.8
	0.3	63.1	79.8	84.8	**93.2**	**41.1**
	0.5	**64.4**	**80.5**	**85.6**	92.7	42.5
	0.7	62.1	78.3	83.0	91.6	37.7
	0.9	59.0	75.3	80.6	90.2	34.8
Full-body	0.3	**69.0**	**83.6**	**88.1**	**94.8**	**46.3**
	0.5	68.0	82.6	87.0	94.3	44.6

Table 3
Results comparison on the Market1501 benchmark. The top two results are given in **bold**.

Model	Rank 1	Rank 5	Rank 10	mAP
Zeng et al. [53]	88.5	–	–	71.5
Chang et al. [54]	84.7	94.2	96.6	64.7
Ge et al. [55]	90.5	–	–	77.7
Wang et al. [56]	91.3	–	–	76.0
Yuan et al. [57]	91.4	96.6	97.7	76.7
Zhou et al. [51]	91.5	96.8	97.3	85.4
Chang et al. [58]	92.1	96.5	98.6	81.9
Jiang et al. [59]	92.3	–	–	78.2
Liu et al. [60]	93.3	–	–	76.8
Zhang et al. [61]	93.3	97.5	98.4	81.3
Tang et al. [62]	93.4	97.6	98.5	82.2
Chen et al. [63]	93.9	–	–	84.5
Khatun et al. [50]	94.7	95.7	98.5	–
Ours	**95.1**	**98.2**	**99.0**	**86.5**
Ours + re-ranking	**95.8**	**98.0**	**98.5**	**94.3**

Table 4
Results comparison on the MSMT17 benchmark. The best results are given in **bold**.

Model	Rank 1	Rank 5	Rank 10	mAP
Ye et al. [64]	68.3	–	–	**49.3**
Jiang et al. [59]	68.8	–	–	41.0
Yuan et al. [57]	69.4	81.5	85.6	39.2
Ours	**71.7**	**83.6**	**87.4**	45.9

son. This method indicates that the k-reciprocal nearest neighbors to the probe image should have more priorities in the ranking list. Using this technique with settings $k1 = 20$, $k2 = 6$, and $\lambda = 0.3$, the rank 1 and mAP results were improved from 95.1 and 86.5 to 95.8 and 94.3, respectively.

5.3.3. *Experiments on the MSMT17-V2 dataset*

The empirical results over the MSMT17-v2 benchmark [23] are given in Table 4. Results show that the proposed method advances the state-of-the-art methods in ranks 1, 5, and 10 by more than 2 percent, while - based on the mAP metric - our method (45.9%) ranks second best, after [64].

6. Conclusions

CNNs are known to be able to autonomously find the critical regions of the input data and discriminate between foreground-background regions. However, to accomplish such a challenging goal, they demand large volumes of learning data, which can be hard to collect and particularly costly to annotate, in case of supervised learning problems. In this paper, we described a solution based on data segmentation and swapping, that interchanges segments *apriori* deemed to be important or irrelevant for the network responses. The proposed method can be seen as a data augmentation solution that implicitly empowers the network to improve its *receptive fields inference skill*. In practice, during the learning phase, we provide the network with an attentional mechanism derived from prior information (i.e., annotations and body masks), that determines not only the critical regions of the input data but also provides important cues about any useless input segments that should be disregarded from the decision process. Finally, it is important to stress that, in *test* time, samples are provided without any segmentation mask, which lowers the computational burden with respect to previously proposed explicit attention mechanisms. As a proof-of-concept, our experiments were carried out in the highly challenging pedestrian re-identification problem, and the results show that our approach –as a complementary data augmentation technique– could contribute to significant improvements in the performance of the state-of-the-art.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the FCT/MEC through National Funds and Co-Funded by the FEDER-PT2020 Partnership Agreement under Project UIDB/50008/2020, Project POCI-01-0247-FEDER-033395 and in part by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT.

References

[1] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 60.

[2] N. Dvornik, J. Mairal, C. Schmid, On the importance of visual context for data augmentation in scene understanding, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1–1 In press, doi:10.1109/TPAMI.2019.2961896.

[3] R.R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: European Conference on Computer Vision, Springer, 2016, pp. 135–153.

[4] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 384–393.

[5] J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, Attention-aware compositional network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2119–2128.

[6] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3219–3228.

[7] Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, IEEE Trans. Circuits Syst. Video Technol. 29 (10) (2018) 3037–3045.

[8] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, J. Sun, Alignedreid: surpassing human-level performance in person re-identification, arXiv:1711.08184 (2017).

[9] H. Inoue, Data augmentation by pairing samples for images classification, arXiv:1801.02929 (2018).

[10] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, Proc AAAI Conf, 2020. 0–0

[11] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-identification, Image Vis. Comput. 32 (4) (2014) 270–286.

[12] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, D.-S. Huang, Deep learning-based methods for person re-identification: a comprehensive review, Neurocomputing 337 (2019) 354–371, doi:10.1016/j.neucom.2019.01.079.

[13] G. Chen, J. Lu, M. Yang, J. Zhou, Spatial-temporal attention-aware learning for video-based person re-identification, IEEE Trans. Image Process. 28 (9) (2019) 4192–4205.

[14] L. Zhang, Z. Shi, J.T. Zhou, M.-M. Cheng, Y. Liu, J.-W. Bian, Z. Zeng, C. Shen, Ordered or orderless: a revisit for video based person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1–1 In press, doi:10.1109/TPAMI.2020.2976969.

[15] L. Cheng, X.-Y. Jing, X. Zhu, F. Ma, C.-H. Hu, Z. Cai, F. Qi, Scale-fusion framework for improving video-based person re-identification performance, Neural Comput. Appl. 32 (2020) 1–18, doi:10.1007/s00521-020-04730-z.

[16] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, Attention driven person re-identification, Pattern Recognit. 86 (2019) 143–155.

[17] C. Zhou, H. Yu, Mask-guided region attention network for person re-identification, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2020, pp. 286–298.

[18] M. Denil, L. Bazzani, H. Larochelle, N. de Freitas, Learning where to attend with deep architectures for image tracking, Neural Comput. 24 (8) (2012) 2151–2184.

[19] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3754–3762.

[20] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, J. Hu, Pose transferrable person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4099–4108.

[21] A. Borgia, Y. Hua, E. Kodirov, N. Robertson, GAN-based pose-aware regulation for video-based person re-identification, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1175–1184.

[22] Y. Lin, Y. Wu, C. Yan, M. Xu, Y. Yang, Unsupervised person re-identification via cross-camera similarity exploration, IEEE Trans. Image Process. 29 (2020) 5481–5490.

[23] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer GAN to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 79–88.

[24] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proc IEEE ICCV, 2017, pp. 2961–2969.

[25] S. Suzuki, et al., Topological structural analysis of digitized binary images by border following, Comput. Vis. Graph. Image Process. 30 (1) (1985) 32–46.

[26] A. Telea, An image inpainting technique based on the fast marching method, J. Graph. Tools 9 (1) (2004) 23–34.

[27] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, RMPE: regional multi-person pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2334–2343.

[28] T. Zhang, R. Ramakrishnan, M. Livny, Birch: an efficient data clustering method for very large databases, ACM Sigmod Rec. 25 (2) (1996) 103–114.

[29] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, Oakland, CA, USA, 1967, pp. 281–297.

[30] D. Sculley, Web-scale k-means clustering, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 1177–1178.

[31] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in: Advances in Neural Information Processing Systems, 2002, pp. 849–856.

[32] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019. 0–0

[33] D. Li, Z. Zhang, X. Chen, K. Huang, A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios, IEEE Trans. Image Process. 28 (4) (2018) 1575–1590.

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[35] X. Fan, W. Jiang, H. Luo, M. Fei, Spherereid: deep hypersphere manifold embedding for person re-identification, J. Vis. Commun. Image Represent. 60 (2019) 51–58.

[36] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned CNN embedding for person reidentification, ACM TOMM 14 (1) (2018) 13.

[37] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), in: Proc IEEE ECCV, 2018, pp. 480–496.

[38] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[39] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2009) 1627–1645.

[40] K. Musgrave, S. Belongie, S.-N. Lim, A metric learning reality check, arXiv:2003.08505 (2020).

[41] K. Roth, B. Brattoli, B. Ommer, MIC: mining interclass characteristics for improved metric learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8000–8009.

[42] W. Kim, B. Goyal, K. Chawla, J. Lee, K. Kwon, Attention-based ensemble for deep metric learning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 736–751.

[43] F. Cakir, K. He, X. Xia, B. Kulis, S. Sclaroff, Deep metric learning to rank, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1861–1870.

[44] X. Wang, X. Han, W. Huang, D. Dong, M.R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.

[45] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.

[46] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, R. Jin, Softtriple loss: deep metric learning without triplet sampling, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6450–6458.

[47] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: Proc IEEE CVPR, 2018, pp. 1062–1071.

[48] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: a novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2018) 1176–1190.

[49] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: Proc IEEE CVPR, 2018, pp. 2285–2294.

[50] A. Khatun, S. Denman, S. Sridharan, C. Fookes, Semantic consistency and identity mapping multi-component generative adversarial network for person re-identification, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 2267–2276.

[51] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, B. Zhang, R. Ji, Fine-grained spatial alignment model for person re-identification with focal triplet loss, IEEE Trans. Image Process. 29 (2020) 7578–7589.

[52] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.

[53] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, S. Satoh, Illumination-adaptive person re-identification, IEEE Trans. Multimed. 22 (2020) 3064–3074 12, doi:10.1109/TMM.2020.2969782.

[54] Y.-S. Chang, M.-Y. Wang, L. He, W. Lu, H. Su, N. Gao, X.-A. Yang, Joint deep semantic embedding and metric learning for person re-identification, Pattern Recognit. Lett. 130 (2020) 306–311.

[55] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, et al., FD-GAN: pose-guided feature distilling GAN for robust person re-identification, in: Advances in Neural Inf. Process. Syst., 2018, pp. 1222–1233.

[56] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, S. Satoh, Learning sparse and identity-preserved hidden attributes for person re-identification, IEEE Trans. Image Process. 29 (1) (2019) 2013–2025.

[57] Y. Yuan, W. Chen, Y. Yang, Z. Wang, In defense of the triplet loss again: learning robust person re-identification with fast approximated triplet loss and label distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 354–355.

[58] Z. Chang, Z. Qin, H. Fan, H. Su, H. Yang, S. Zheng, H. Ling, Weighted bilinear coding over salient body parts for person re-identification, Neurocomputing 407 (2020) 454–464.

[59] M. Jiang, C. Li, J. Kong, Z. Teng, D. Zhuang, Cross-level reinforced attention network for person re-identification, J. Vis. Commun. Image Represent. (2020) 102775.

[60] S. Liu, T. Si, X. Hao, Z. Zhang, Semantic constraint GAN for person re-identification in camera sensor networks, IEEE Access 7 (2019) 176257–176265.

[61] W. Zhang, L. Huang, Z. Wei, J. Nie, Appearance feature enhancement for person re-identification, Expert Syst. Appl. (2020) 113771.

[62] Y. Tang, X. Yang, N. Wang, B. Song, X. Gao, Person re-identification with feature pyramid optimization and gradual background suppression, Neural Netw. 124 (2020) 223–232.

[63] F. Chen, N. Wang, J. Tang, D. Liang, H. Feng, Self-supervised data augmentation for person re-identification, Neurocomputing 415 (2020) 48–59, doi:10.1016/j.neucom.2020.07.087.

[64] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: a survey and outlook, arXiv:2001.04193 (2020).