A Quadruplet Loss for Enforcing Semantically Coherent Embeddings in Multi-output Classification Problems

Hugo Proença, *Senior Member, IEEE*, Ehsan Yaghoubi and Pendar Alirezazadeh

***Abstract*—This paper describes one objective function for learning semantically coherent feature embeddings in multi-output classification problems, i.e., when the response variables have dimension higher than one. Such coherent embeddings can be used simultaneously for different tasks, such as identity retrieval and soft biometrics labelling. We propose a generalization of the triplet loss [34] that: 1) defines a metric that considers the number of agreeing labels between pairs of elements; 2) introduces the concept of *similar* classes, according to the values provided by the metric; and 3) disregards the notion of *anchor*, sampling four arbitrary elements at each time, from where two pairs are defined. The distances between elements in each pair are imposed according to their *semantic similarity* (i.e., the number of agreeing labels). Likewise the triplet loss, our proposal also privileges small distances between *positive* pairs. However, the key novelty is to additionally enforce that the distance between elements of any other pair corresponds inversely to their semantic similarity. The proposed loss yields embeddings with a strong correspondence between the classes centroids and their semantic descriptions. In practice, it is a natural choice to jointly infer coarse (soft biometrics) + fine (ID) labels, using simple rules such as *k-neighbours*. Also, in opposition to its triplet counterpart, the proposed loss appears to be agnostic with regard to demanding criteria for mining learning instances (such as the *semi-hard* pairs). Our experiments were carried out in five different datasets (BIODI, LFW, IJB-A, Megaface and PETA) and validate our assumptions, showing results that are comparable to the state-of-the-art in both the identity retrieval and soft biometrics labelling tasks.**

***Index Terms*—Feature embedding, Soft biometrics, Identity retrieval, Convolutional neural networks, Triplet loss.**

I. Introduction

CHaracterizing pedestrians in crowds has been attracting growing attention, with soft biometrics (e.g., *gender*, *ethnicity* or *age*) being particularly important to determine the identities in a scene. This kind of labels is closely related to human perception and describes the visual appearance of subjects, with applications in identity retrieval [40][36] and person re-identification [15][27].

Deep learning frameworks have been repeatedly improving the state-of-the-art in many computer vision tasks, such as object detection and classification [25][41], action recognition [19][6], semantic segmentation [24][44] and soft biometrics inference [32]. In this context, the triplet loss [34] is a popular concept, where three learning elements are considered at a time, two of them of the same class and a third one of a different class. By imposing larger distances between the elements of the *negative* than of the *positive* pair, the intra-class compactness and inter-class discrepancy in the destiny space are enforced. This strategy was successfully applied to various problems, upon the mining of the *semi-hard* negative input pairs, i.e., cases where the negative element is farther to the anchor than the positive, but still provides a positive loss due to an imposed margin.

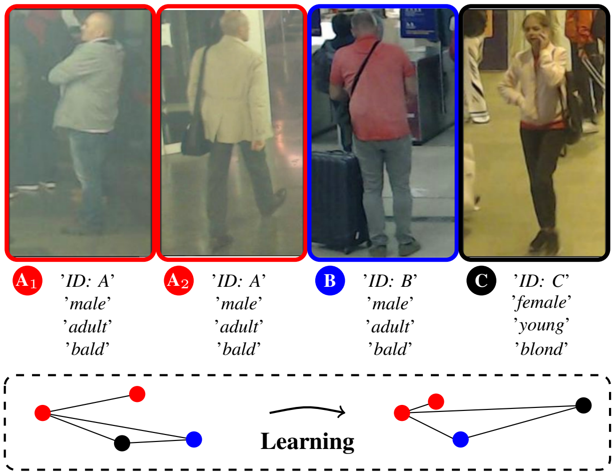

Fig. 1. Likewise the triplet loss [34], the proposed **quadruplet** formulation minimizes the distances between elements of *positive* pairs $\{A_1,\ A_2\}$. However, the key novelty is to additionally consider the semantic similarity between classes (A, B and C). In this example, assuming that A and B are semantically similar, our proposal privileges embeddings where the distances between (A$_{.}$, B) elements are smaller than the distances between (A$_{.}$, C) and between (B, C) elements.

This paper describes one objective function that is a generalization of the triplet loss. Instead of dividing the learning pairs into *positive/negative*, we define a metric to perceive the semantic similarity between two classes (IDs). In learning time, four elements are considered at a time and the margins between the pairwise distances yield from the number of agreeing labels in each pair (Fig. 1). Under this formulation, elements of *similar* classes (e.g., two ”*young, black, bald, male*” subjects) are projected into adjacent regions of the destiny space. Also, as we impose different margins between (almost) all *negative* pairs, we leverage the difficulties in mining appropriate learning instances, which is one of the main difficulties in the triplet loss formulation.

Authors are with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Covilhã, Portugal, E-mail: hugomcp@di.ubi.pt, {D2401, D2389}@di.ubi.pt.

Manuscript received: January, 2020.

The proposed loss function is particularly suitable for *coarse-to-fine* classification problems, where some labels are easier to infer than others and the global problem can be decomposed into more tractable sub-components. This hierarchical paradigm is known to be an efficient way of organizing object recognition, not only to accommodate a large number of hypotheses, but also to systematically exploit the shared attributes. Under this paradigm, the identity retrieval problem is of particular interest, where the finest labels (IDs) are seen as the leaves of hierarchical structures with roots such as the *gender* or *ethnicity* features. However, note that the proposed formulation does not appropriately handle soft labels that vary among different images of a subject (e.g., *hairstyle*). Also, it does not take into account the varying difficulty of estimating the different labels, allowing further improvements based in metric learning concepts.

The remainder of this paper is organized as follows: Section II summarizes the most relevant research in the scope of our work. Section III describes the proposed objective function. In Section IV we discuss the obtained results and the conclusions are given in Section V.

II. Related Work

Deep learning methods for biometrics can be roughly divided into two major groups: 1) methods that directly learn multi-class classifiers used in identity retrieval and soft biometrics inference; and 2) methods that learn low-dimensional feature embeddings, where inference yields from nearest neighbour search.

A. Soft Biometrics and Identity Retrieval

Bekele *et al.* [2] proposed a residual network for multi-output inference that handles classes-imbalance directly in the cost function, without depending of data augmentation techniques. Almudhahka *et al.* [1] explored the concept of comparative soft biometrics and assessed the impact of automatic estimations on face retrieval performance. Guo *et al.* [12] studied the influence of distance in the effectiveness of body and facial soft biometrics, introducing a joint density distribution based rank-score fusion strategy [13]. Vera-Rodriguez *et al.* [31] used hand-crafted features extracted from the distances between key points in body silhouettes. Martinho-Corbishley *et al.* [29] introduced the idea of *super-fine* soft attributes, describing multiple concepts of one trait as multi-dimensional perceptual coordinates. Also, using joint attribute regression and deep residual CNNs, they observed substantially better retrieval performance in comparison to conventional labels. Schumann and Specker used an ensemble of classifiers for robust attributes inference [35], extended to full body search by combining it with a human silhouette detector. He *et al.* [17] proposed a weighted multi-task CNN with a loss term that dynamically updates the weight for each task during the learning phase.

Several works regarded the semantic segmentation as a tool to support labels inference: Galiyawala *et al.* [10] described a deep learning framework for person retrieval using the height, clothes' color, and gender labels, with a segmentation module used to remove clutter. Similarly, Cipcigan and Nixon [3] obtained semantically segmented regions of the body that fed two CNN-based feature extraction and inference modules.

Finally, specifically designed for handheld devices, Samangouei and Chellappa [32] extracted various facial soft biometric features, while Neal and Woodard [26] developed a human retrieval scheme based on thirteen demographic and behavioural attributes from mobile phones data, such as calling, SMS and application data, having authors positively concluded about the feasibility of this kind of recognition.

A comprehensive summary of the most relevant research in soft biometrics is given in [38].

B. Feature Embeddings and Loss Functions

Triplet loss functions were motivated by the concept of *contrastive* loss [14], where the rationale is to penalize distances between *positive* pairs, while favouring distances between *negative* pairs. Kang *et al.* [21] used a deep ensemble of multi-scale CNNs, each one based on triplet loss functions. Song *et al.* [37] learned semantic feature embeddings that lift the vector of pairwise distances within the batch to the matrix of pairwise distances, and described a structured loss on the lifted problem. Liu and Huan [28] proposed a triplet loss learning architecture composed of four CNNs, each one learning features from different body parts that are fused at the score level.

A posterior concept was the *center* loss [42], which finds a center for each class and penalizes the distances between the projections and their corresponding class center. Jian *et al.* [20] combined additive margin *softmax* with center loss to increase the inter-classes distances and avoid overconfidence on classifications. Ranjan *et al.*'s *crystal* loss [30] restricts the features to lie on a hypersphere of a fixed radius, adding a constraint on the features projections such that their ℓ_2-norm is constant. Chen *et al.* [4] used deep representations to feed a Bayesian metrics learning module that maximizes the log-likelihood ratio between intra- and inter-classes distances. Deng *et al.*'s *Sphereface* [8] proposes an additive angular margin loss, with a clear geometric interpretation due to the correspondence to the geodesic distance on the hypersphere.

Observing that CNN-based methods tend to overfit in person re-identification tasks, Shi *et al.* [36] used siamese architectures to provide a joint description to a metric learning module, regularizing the learning process and improving the generalization ability. Also, to cope with large intra-class variations, they suggested the idea of *moderate positive mining*, again to prevent overfitting. Motivated by the difficulties in generate learning instances for triplet loss frameworks, Su *et al.* [39] performed adaptive CNN fine-tuning, along with an adaptive loss function that relates the maximum distance among the positive pairs to the margin demanded for separate *positive* from *negative* pairs. Hu *et al.* [18] proposed an objective function that generalizes the Maximum Mean Discrepancy [33] metric, with a weighting scheme that favours good quality data. Duan *et al.* [9] proposed the *uniform* loss to learn deep equi-distributed representations for face recognition. Finally, observing the typical unbalance between positive and negative pairs, Wang *et al.* [41] described an adaptive margin list-wise loss, in which learning data are provided with a set of negative pairs divided into three classes (*easy*, *moderate*, and *hard*), depending of the distance rank with respect to the query.

Finally, we note the differences between our loss function and the (also *quadruplet*) loss described by Chen *et al.* [5]. These authors attempt to augment the inter-classes margins and the intra-class compactness without explicitly using any semantical constraint. As in the original triplet loss formulation, the concept of *similar* class doesn't exist in [5], and there is no rule to explicitly enforce the projection of identities that share most of the labels into neighbour regions of the latent space. In opposition, our method concerns essentially about such kind of semantical coherence, i.e., assures that similar classes are projected into adjacent regions of the embedding. Also, even the idea behind the loss formulation is radically different in both methods, in the sense that [5] still considers the concept of *anchor* (as the triplet-loss), which is also in opposition to our proposal.

III. Proposed Method

A. Quadruplet Loss: Definition

Consider a supervised classification problem, where t is the dimensionality of the response variable $\boldsymbol{y}_i$ associated to the input element $\boldsymbol{x}_i \in [0, 255]^n$. Let $f(.)$ be one embedding function that maps $\boldsymbol{x}_i$ into a d-dimensional space Ψ, with $\boldsymbol{f}_i = f(\boldsymbol{x}_i) \in \Psi$ being the projected vector. Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_b\}$ be a batch of b images from the learning set. We define $\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) \in \mathbb{N}, \forall i, j \in \{1, \dots, b\}$ as the function that measures the semantic similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_0, \tag{1}$$

with $||.||_0$ being the ℓ_0-norm operator.

In practice, $\phi(., .)$ counts the number of disagreeing labels between the $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ pair, i.e., $\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) = t$ when the ith and jth elements have fully disjoint classes membership (e.g., one "*black, adult, male*" and another "*white, young, female*" subjects), while $\phi(\boldsymbol{y}_1, \boldsymbol{y}_2) = 0$ when they have the exact same label (class) across all dimensions, i.e., when they constitute a *positive* pair.

Let $\{i, j, p, q\}$ be the indices of four images in the batch. The corresponding quadruplet loss value $\ell_{i,j,p,q}$ is given by:

$$\begin{aligned}\ell_{i,j,p,q} = sgn\Big(\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) - \phi(\boldsymbol{y}_p, \boldsymbol{y}_q)\Big)& \\ \Big[\big(\|\boldsymbol{f}_p - \boldsymbol{f}_q)\|_2^2 - \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2\big) + \alpha\Big]&, \end{aligned}\tag{2}$$

where $sgn()$ is the sign function, $||\boldsymbol{x}||_2^2$ denotes the square of the ℓ_2-norm of $\boldsymbol{x}$ $\big(||\boldsymbol{x}||_2 = (x_1^2 + \dots x_n^2)^{\frac{1}{2}}$, i.e., $||\boldsymbol{x}||_2^2 = x_1^2 + \dots x_n^2\big)$ and α is the desired margin ($\alpha = 0.1$ was used in our experiments). Evidently, the loss value will be zero when both image pairs have the same number of agreeing labels (as $sgn(0) = 0$ in these cases). In any other case, the sign function will determine the pair which distance in the embedding should be minimized. As an example, if the (p, q) elements are semantically closer to each other than the (i, j) elements $\big(\phi(\boldsymbol{y}_p, \boldsymbol{y}_q) < \phi(\boldsymbol{y}_i, \boldsymbol{y}_j)\big)$, we want to ensure that $\|\boldsymbol{f}_p - \boldsymbol{f}_q\|_2^2 < \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2$.

The accumulated loss in the batch is given by the truncated mean of a sample (of size s) randomly taken from the subset of the $\binom{b}{4}$ individual loss values where $\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) \neq \phi(\boldsymbol{y}_p, \boldsymbol{y}_q)$:

$$\mathcal{L} = \frac{1}{s} \sum_{\boldsymbol{z}=1}^{s} \Big[\ell_{\boldsymbol{z}}\Big]_+, \tag{3}$$

where $\boldsymbol{z} \in \{1, \dots, s\}^4$ denotes the zth composition of four elements in the batch and $[.]_+$ is the $\max(., 0)$ function. Even considering that a large fraction of the combinations in the batch will be invalid (i.e., with $\phi(., .) = 0$), large values of b will result in an intractable number of combinations at each iteration. In practical terms, after filtering out those invalid combinations, we randomly sample a subset of the remaining instances, which is designated as the *mini-batch*.

B. Quadruplet Loss: Training

Consider four indices $\{i, j, p, q\}$ of elements in the mini-batch, with $\phi(\boldsymbol{y}_i, \boldsymbol{y}_j) > \phi(\boldsymbol{y}_p, \boldsymbol{y}_q)$. Let Δ_ϕ denote the difference between the number of disagreeing labels of the $\{i, j\}$ and $\{p, q\}$ pairs:

$$\Delta_\phi = \phi(\boldsymbol{y}_i, \boldsymbol{y}_j) - \phi(\boldsymbol{y}_p, \boldsymbol{y}_q). \tag{4}$$

Also, let $\Delta_{\boldsymbol{f}}$ be the distance between the elements of the most alike pair minus the distance between the elements of the least alike pair in the destiny space (plus the margin):

$$\Delta_{\boldsymbol{f}} = \|\boldsymbol{f}_p - \boldsymbol{f}_q)\|_2^2 - \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2^2 + \alpha. \tag{5}$$

Upon basic algebraic manipulation, the gradients of $\mathcal{L}$ with respect to the quadruplet terms are given by:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_i} = \sum_{\boldsymbol{z}} \begin{cases} 2(\boldsymbol{f}_j - \boldsymbol{f}_i) & \text{, if } \Delta_\phi > 0 \ \wedge \Delta_{\boldsymbol{f}} \geq 0 \\ 0 & \text{, otherwise} \end{cases} \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_j} = \sum_{\boldsymbol{z}} \begin{cases} 2(\boldsymbol{f}_i - \boldsymbol{f}_j) & \text{, if } \Delta_\phi > 0 \ \wedge \Delta_{\boldsymbol{f}} \geq 0 \\ 0 & \text{, otherwise} \end{cases} \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_p} = \sum_{\boldsymbol{z}} \begin{cases} 2(\boldsymbol{f}_p - \boldsymbol{f}_q) & \text{, if } \Delta_\phi > 0 \ \wedge \Delta_{\boldsymbol{f}} \geq 0 \\ 0 & \text{, otherwise} \end{cases} \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_q} = \sum_{\boldsymbol{z}} \begin{cases} 2(\boldsymbol{f}_q - \boldsymbol{f}_p) & \text{, if } \Delta_\phi > 0 \ \wedge \Delta_{\boldsymbol{f}} \geq 0 \\ 0 & \text{, otherwise} \end{cases} \tag{9}$$

In practice terms, the model weights are adjusted only when pairs have different number of agreeing labels (i.e., $\Delta_\phi > 0$) and when the distance in the destiny space between the elements of the most similar pair is higher than the distance between the elements of the least similar pair (plus the margin, $\Delta_{\boldsymbol{f}} \geq 0$). According to this idea, using (6)-(9), the deep learning frameworks supervised by the proposed quadruplet loss are trainable in a way similar to its counterpart triplet

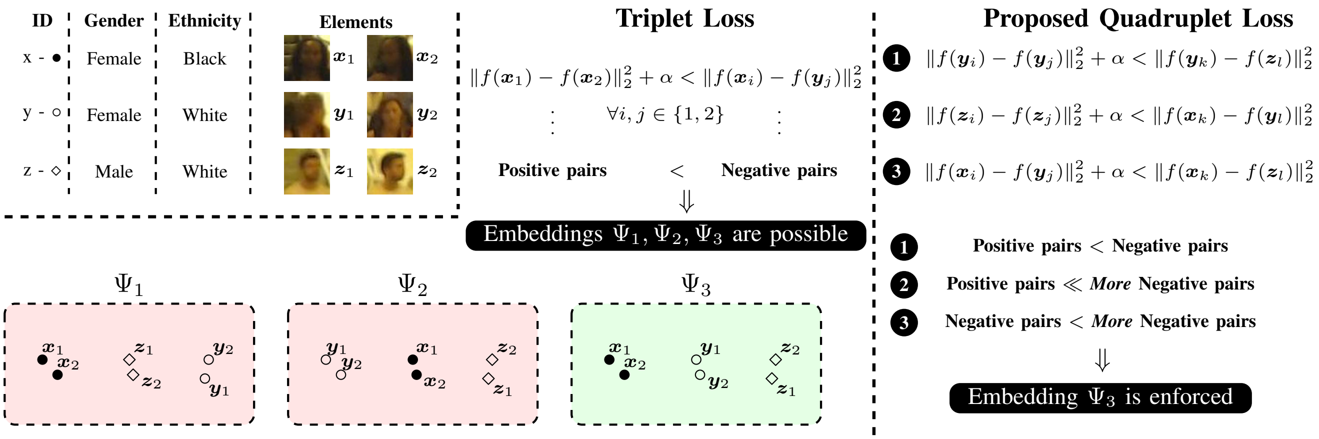
ID
Gender
Ethnicity
Elements
x - ●
Female
Black
x_1
x_2
y - ○
Female
White
y_1
y_2
z - ◇
Male
White
z_1
z_2
Triplet Loss
$\|f(x_1) - f(x_2)\|_2^2 + \alpha < \|f(x_i) - f(y_j)\|_2^2$
$\forall i, j \in \{1, 2\}$
Positive pairs < Negative pairs
Embeddings Ψ_1, Ψ_2, Ψ_3 are possible
Ψ_1
Ψ_2
Ψ_3
Proposed Quadruplet Loss
1 $\|f(y_i) - f(y_j)\|_2^2 + \alpha < \|f(y_k) - f(z_l)\|_2^2$
2 $\|f(z_i) - f(z_j)\|_2^2 + \alpha < \|f(x_k) - f(y_l)\|_2^2$
3 $\|f(x_i) - f(y_j)\|_2^2 + \alpha < \|f(x_k) - f(z_l)\|_2^2$
1 Positive pairs < Negative pairs
2 Positive pairs ≪ More Negative pairs
3 Negative pairs < More Negative pairs
Embedding Ψ_3 is enforced

Fig. 2. Key difference between the triplet loss [34] formulation and the solution proposed in this paper. Using a loss function that analyzes the semantic similarity (in terms of soft biometrics) between the different identities, we enforce embeddings (Ψ_3) that are semantically coherent, i.e., where: 1) elements of the same class appear near each other; but additionally 2) elements of similar classes appear closer to each other than elements with no labels in common. This is in opposition to the original formulation of the triplet loss, that relies mostly in image appearance to define the geometry of the destiny space, obtaining - in case of noisy image features - semantically incoherent embeddings (e.g., in Ψ_1 and Ψ_2, classes are compact and discriminative, but the $\boldsymbol{x}/\boldsymbol{z}$ centroids are too close to each other).

loss and can be optimized according to the standard Stochastic Gradient Descend (SGD) algorithm, which was done in all our experiments.

For clarity purposes, Algorithm 1 gives a pseudocode description of the learning phase and of the batch/mini-batch definition processes.

Algorithm 1 Pseudocode description of the learning phase and of the batch/mini-batch definition processes.

```
Precondition: $\boldsymbol{M}$: CNN, $t_e$: Tot. epochs, $s$: mini-batch size, $b$: batch size, $\boldsymbol{I}$: Learning set, $n$ images
  for 1 to $t_e$ do
    for 1 to $\lfloor \frac{n}{s} \rceil$ do
      $\boldsymbol{b} \leftarrow$ randomly sample $b$ out of $n$ images from $\boldsymbol{I}$
      $\boldsymbol{c} \leftarrow$ create $\binom{b}{4}$ quadruplet combinations from $\boldsymbol{b}$
      $\boldsymbol{c}* \leftarrow$ filter out invalid elements from $\boldsymbol{c}$
      $\boldsymbol{s} \leftarrow$ randomly sample $s$ elements from $\boldsymbol{c}^*$
      $\boldsymbol{M} \leftarrow$ update weights($\boldsymbol{M}, \boldsymbol{s}$) (eqs. (6-9))
    end for
  end for
  return $\boldsymbol{M}$
```

C. Quadruplet Loss: Insight and Example

Fig. 2 illustrates our rationale in the proposed loss. By defining a metric that analyses the similarity between two classes, we create the concept of *semantically similar* class. This enables to explicitly enforce that elements of the *least similar* classes (with no common labels) are at the farthest distances in the embedding. During the learning phase, we sample the image pairs in a stochastic way and enforce projections in a way that resembles the human perception of *semantic similarity*.

As an example, Fig. 3 compares the bidimensional embeddings resulting from the triplet and the quadruplet losses, for the LFW identities with more than 15 images in the dataset (using $t = 2$: {'ID', 'Gender'} labels). This plot yielded from the projection of a 128-dimensional embedding down to two dimensions, according to the Neighbourhood Component Analysis (NCA) [11] algorithm.

It can be seen that the triplet loss provided an embedding where the positions of elements are exclusively determined by their appearance, where 'females' appear nearby 'male tennis players' (upper left corner). In opposition, the quadruplet loss established a large margin between both genders, while keeping the compactness per ID. This kind of embedding is interesting: 1) for identity retrieval, to guarantee that all retrieved elements have soft labels equal to the query; 2) upon a semantic description of the query (e.g., "*find adult white males similar to this image*"), to guarantee that all retrieved elements meet the semantic criteria; and 3) to use the same embedding to directly infer fine (ID) + coarse (soft) labels, in a simple *k-neighbours* fashion.

IV. Results and Discussion

A. Experimental Setting and Preprocessing

Our empirical validation was conducted in one proprietary (BIODI) and four freely available datasets (LFW, PETA, IJB-A and Megaface) well known in the biometrics and re-identification literature.

The BIODI[1] dataset is proprietary of *Tomiworld*®[2], being composed of 849,932 images from 13,876 subjects, taken from 216 indoor/outdoor video surveillance sequences. All images were manually annotated for 14 labels: gender, age, height, body volume, ethnicity, hair color and style, beard, moustache, glasses and clothing (x4). The Labeled Faces in the Wild (LFW) [16] dataset contains 13,233 images from 5,749 identities, collected from the web, with large variations in pose, expression and lighting conditions. PETA [7] is a combination of 10 pedestrian re-identification datasets, composed

[1] http://di.ubi.pt/~hugomcp/BIODI/
[2] https://tomiworld.com/

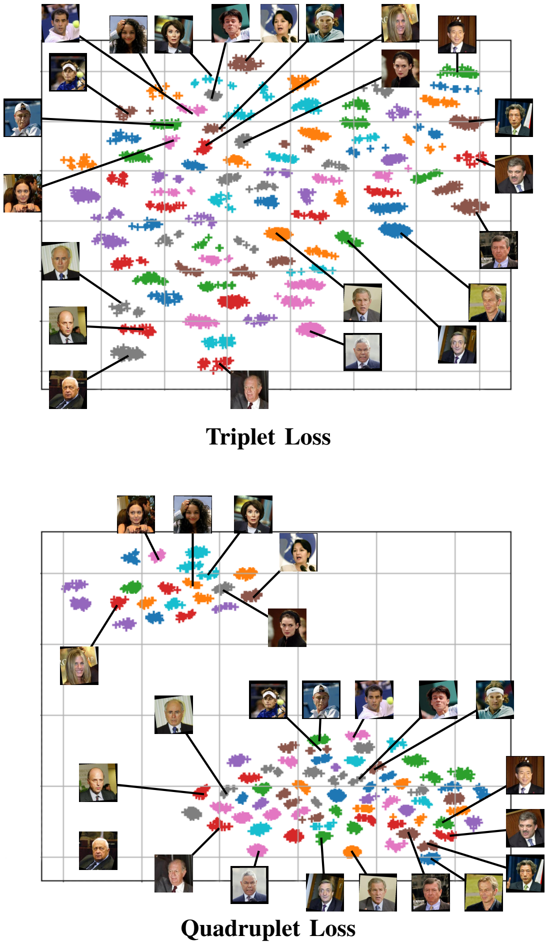

Fig. 3. Comparison between the 2D embeddings resulting from the triplet loss [34] (top plot), and from the proposed quadruplet loss (bottom plot). Results are given for $t = 2$ features {'ID', 'Gender'} for the LFW identities with at least 15 images (89 elements).

of 19,000 images from 8,705 subjects, each one annotated with 61 binary and 4 multi-output atributes. The IIJB-A [23] dataset contains 5,397 images plus 20,412 video frames from 500 individuals, with large variations in pose and illumination. Finally, the Megaface [22] set was released to evaluate face recognition performance at the million scale, and consists of a gallery set and a probe set. The gallery set is a subset of Flickr photos from Yahoo (more than 1,000,000 images from 690,000 subjects). The probe dataset includes FaceScrub and FGNet sets. FaceScrub has 100,000 images from 530 individuals and FGNet contains 1,002 images of 82 identities. Some examples of the images in each dataset are given in Fig. 4.

B. Convolutional Neural Networks

Two CNN architectures were considered: the *VGG* and *ResNet* models (Fig. 5). Here, the idea was not only to compare the performance of the quadruplet loss with respect to the baselines, but also to perceive the variations in performance with respect to different CNN architectures. A *TensorFlow* implementation of both architectures is available at[3].

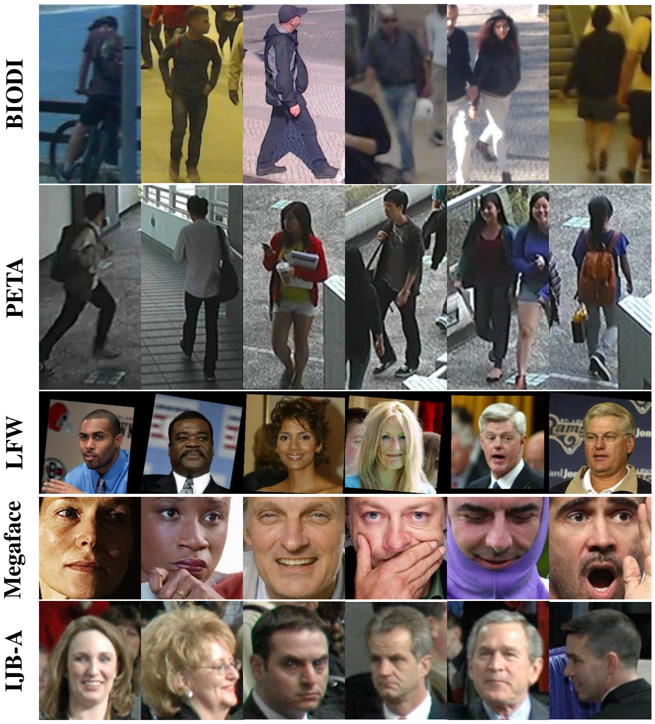

Fig. 4. Datasets used in the empirical validation of the method proposed in this paper. From top to bottom rows, images of the BIODI, PETA, LFW, Megaface and IJB-A sets are shown.

All the models were initialized with random weights, from zero-mean Gaussian distributions with standard deviation 0.01 and bias 0.5. Images were resized to 256 × 256, adding lateral white bands when needed to keep constant ratios. A batch size of 64 was defined, which results in too many combinations of pairs for the triplet/quadruplet losses. At each iteration, we filtered out the invalid triplets/quadruplets instances and randomly selected the mini-batch elements, composed of 64 instances in all cases. For every baseline, 64 pairs were also used as a batch. The learning rate started from 0.01, with momentum 0.9 and weight decay $5e^{-4}$. In the *learning-from-scratch* paradigm, we stopped the learning process when the validation loss didn't decrease for 10 iterations (i.e., *patience*=10).

We initially varied the dimensionality of the embedding (d) to perceive the sensitivity of the proposed method with respect to this parameter. Considering the LFW set, the average AUC values with respect to d are provided in Fig. 6 (the shadowed regions denote the ± standard deviation performance, after 10 trials). As expected, higher values for d were directly correlated to performance, even though results stabilised for dimensions higher than 128. In this regard, we assumed that using higher dimensions would require much more training data, having resorted from this moment to d=128 in all subsequent experiments.

Interestingly, the absolute performance observed for very low d values was not too far of the obtained for much higher

[3] https://github.com/hugomcp/quadruplets

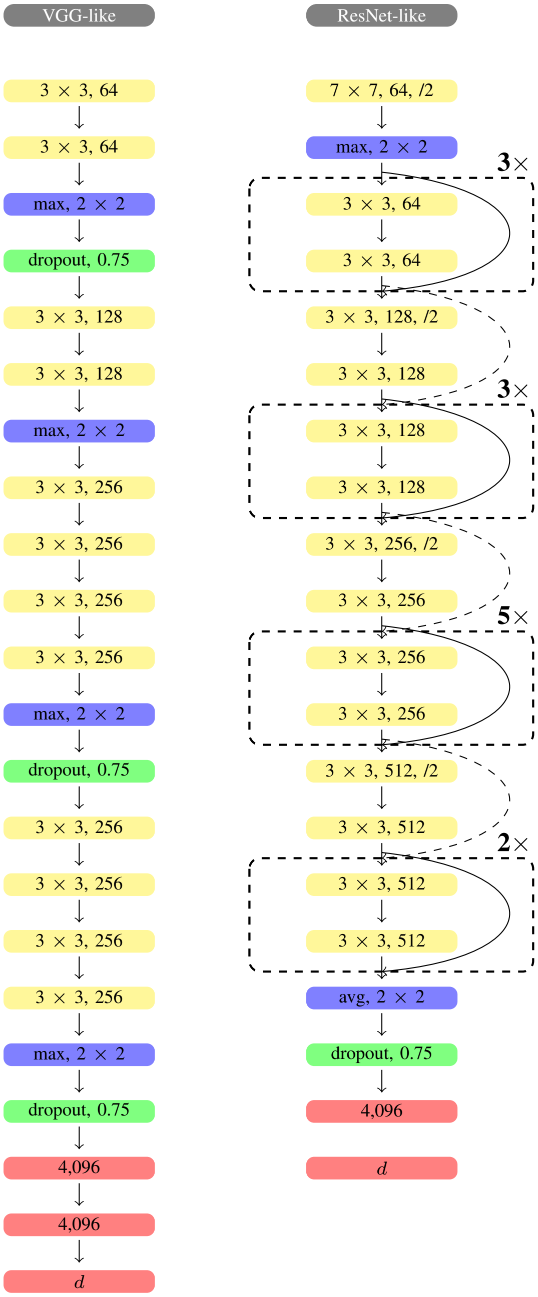

Fig. 5. Architectures of the CNNs used in the experiments. The yellow boxes represent convolutional layers, and the blue and green boxes represent pooling and dropout (keeping probability 0.75) layers. Finally, the red boxes denote fully connected layers. In the ResNet architecture, the dashed skip connections represent convolutions with stride 2 × 2, yielding outputs with half of the spatial input size. The '/2' symbol denotes stride 2 × 2 (the remaining layers use stride 1 × 1).

dimensions, which raises the possibility of using the position of the elements in the destiny space directly for classification and visualization, without the need of any dimensionality reduction algorithm (MDS, LLE or PCA algorithms are frequently seen in the literature for this purpose).

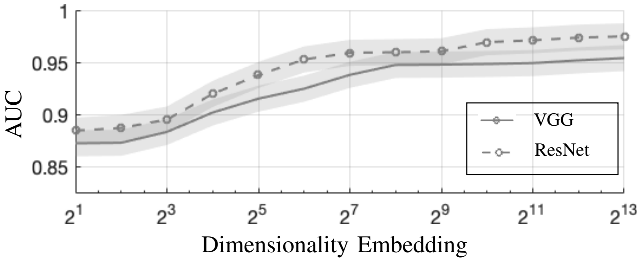

Fig. 6. Variations in the mean AUC values (± the standard deviations after 10 trials, given as shadowed regions) with respect to the dimensionality of the embedding. Results are shown for the LFW validation set, when using the VGG-like (solid line) and ResNet-like (dashed line) CNN architectures.

C. Single- vs. Multi-Output Embeddings Learning: Semantical Coherence

To compare the semantical coherence of the embeddings resulting from single-output (triplet and Chen *et al.*'s losses) and multi-output (ours) learning formulations, we measured the distances (ℓ_2-norm) between each element in an embedding and all the others, grouping values into two sets: 1) *intra-label* observations, when two elements share a specific label (e.g., 'male'/'male' or 'asian'/'asian'); and 2) *inter-labels* observations, in case of different labels in the pair (e.g., 'male'/'female' or 'asian'/'black'). In practice, we measured the distances between elements of the same/different ID, gender, ethnicity and joint gender+ethnicity labels. Note that, in all cases, a unique embedding was obtained for each method, using the $\{$ID$\}$ as feature for the triplet and Chen *et al.* methods, and the $\{$ID, Gender, Ethnicity$\}$ ($t = 3$) for the proposed method, with the annotations for the IJB-A set provided by the Face++ algorithm and subjected to human validation. The VGG-like architecture was considered, as described in Section IV-B.

The results are given in Fig. 7 (LFW, Megaface and IJB-A sets). The green color represents the statistics of the *intra-label* values, while the red color represents the *inter-labels* values. Box plots show the median of the distance values (horizontal solid lines) and the first and third quartiles (top and bottom of the box marks). The upper and lower whiskers are denoted by the horizontal lines outside each box. All outliers are omitted, for visualisation purposes.

The leftmost group in each dataset is the root for the ID retrieval performance, and compares the distances in the embeddings between elements that have the same/different IDs. The remaining cases are the most important for our purposes, and provide the distances between elements that share (or not) some label: the second group compares the 'male'/'male and 'female'/'female' distances (green boxes) to 'male'/'female' values (red boxes). The third group provides the corresponding results for the *ethnicity* label, while the rightmost group provides the distances when jointly considering the *gender* and *ethnicity* features, i.e., when two elements constitute an *intra-label* pair *iff* they have the same gender *and* ethnicity labels.

These results turn evident the different properties of the embeddings yielding from the proposed loss with respect to

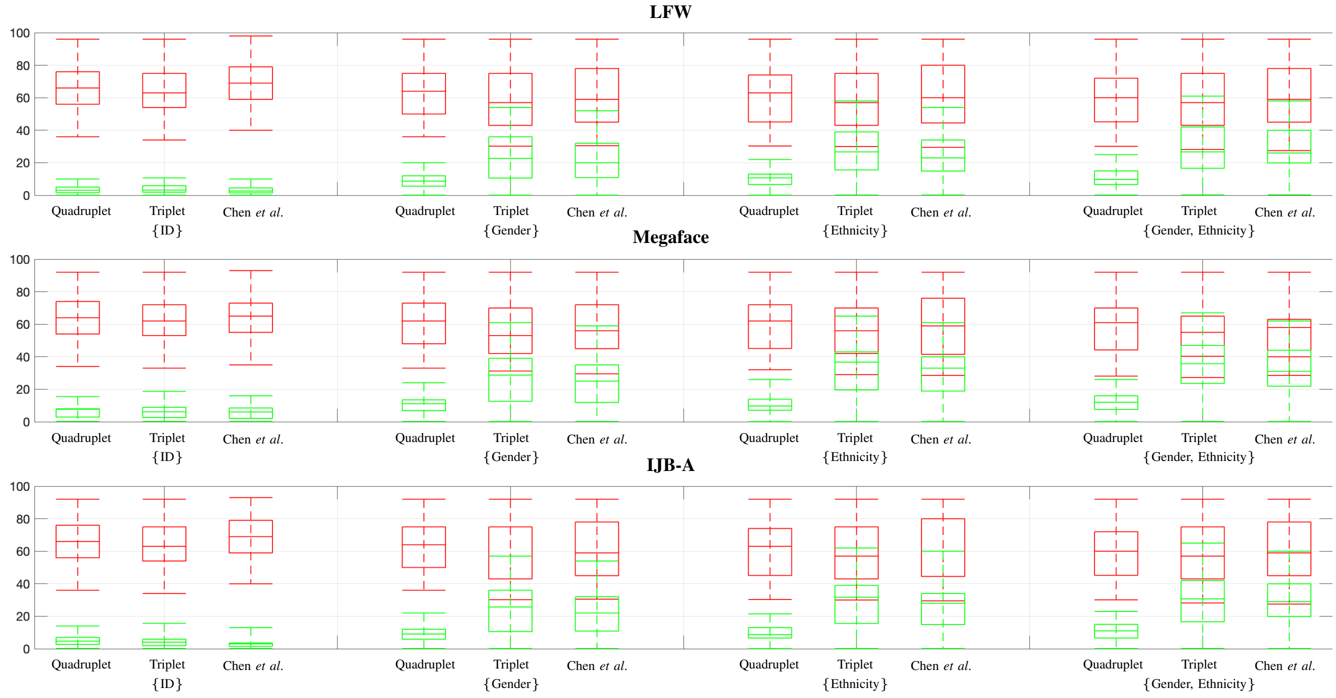

Fig. 7. Box plots of the distances between each element in the embedding with respect to others that share the same (green color) or different (red color) labels. We compare the multi-output learning solution proposed in this paper (Quadruplet), with respect to the single-output learning methods (Triplet [34] and Chen *et al.* [5]). Values regard the LFW (top plot), Megaface (center plot) and IJB-A (bottom plot) sets, measuring the {ID}, {Gender}, {Ethnicity} and {Gender, Ethnicity} same/different label distances.

the baselines. If we consider exclusively the ID to measure the distances between elements, the results almost do not vary among all methods. However, a different conclusion can be drawn when measuring the distances between the same/different *gender*, *ethnicity* and *gender/ethnicity* labels. Here, the proposed quadruplet loss was the unique method where the intra-label/inter-labels whiskers provided disjoint intersections, by a solid margin in all cases, i.e., the difference between the intra-label/inter-labels distances was far larger than in the remaining losses. Of course, such differences are due to the fact that the triplet and Chen *et al*.methods have not considered additional soft labels to define the topology of the embeddings, having exclusively resorted to the ID labels and images appearance for such purpose.

In practice, these experiments turn evident that single-label learning formulation yield embeddings that are semantically incoherent from other labels' perspectives, in the sense that 'males' are often nearby 'females', or 'white' nearby 'asian' elements. In this setting, using such embeddings for simultaneously ID retrieval and soft biometrics labelling is risky, and errors will often occur. In opposition, the proposed loss guarantees large margins between groups of intra-label/inter-labels observations, typically corresponding to *clusters* in the embeddings with respect to the set of learning labels considered.

D. Identity Retrieval

Even considering that the goals of our proposal are beyond the ID retrieval performance, it is important to compare the performance of the quadruplet loss with respect to the baselines in this task. As in the previous experiment, note that all the baselines (triplet loss, center loss, *softmax* and Chen *et al*. [5]) considered exclusively the ID to infer the embeddings, while the proposed loss used all the available labels for that purpose.

Fig. 8 provides the Cumulative Match curves (CMC, outer plots) and the Detection and Identification rates at rank-1 (DIR, inner plots). The results are also summarized in Table I, reporting the rank-1, top-10% values and the mean average precision (mAP) scores, given by:

$$\text{mAP} = \frac{\sum_{q=1}^{n} \bar{P}(q)}{n}, \tag{10}$$

where n is the number of queries, $\bar{P}(q) = \sum_{k=1}^{n} P(k)\Delta r(k)$, $P(k)$ is the precision at cut-off k and $\Delta r(k)$ is the change in recall from $k-1$ to k.

For the LFW set experiment, the BLUFR[4] evaluation protocol was chosen. In the verification (1:1) setting, the test set contained 9,708 face images of 4,249 subjects, which yielded over 47 million matching scores. For the open-set identification problem, the genuine probe set contained 4,350 face images of 1,000 subjects, the impostor probe set had 4,357 images of 3,249 subjects, and the gallery set had 1,000 images. This evaluation protocol was the basis to design, for the other sets, as close as possible experiments, in terms of the number of matching scores, gallery and probe sets.

[4] http://www.cbsr.ia.ac.cn/users/scliao/projects/blufr/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3023304, IEEE Transactions on Information Forensics and Security

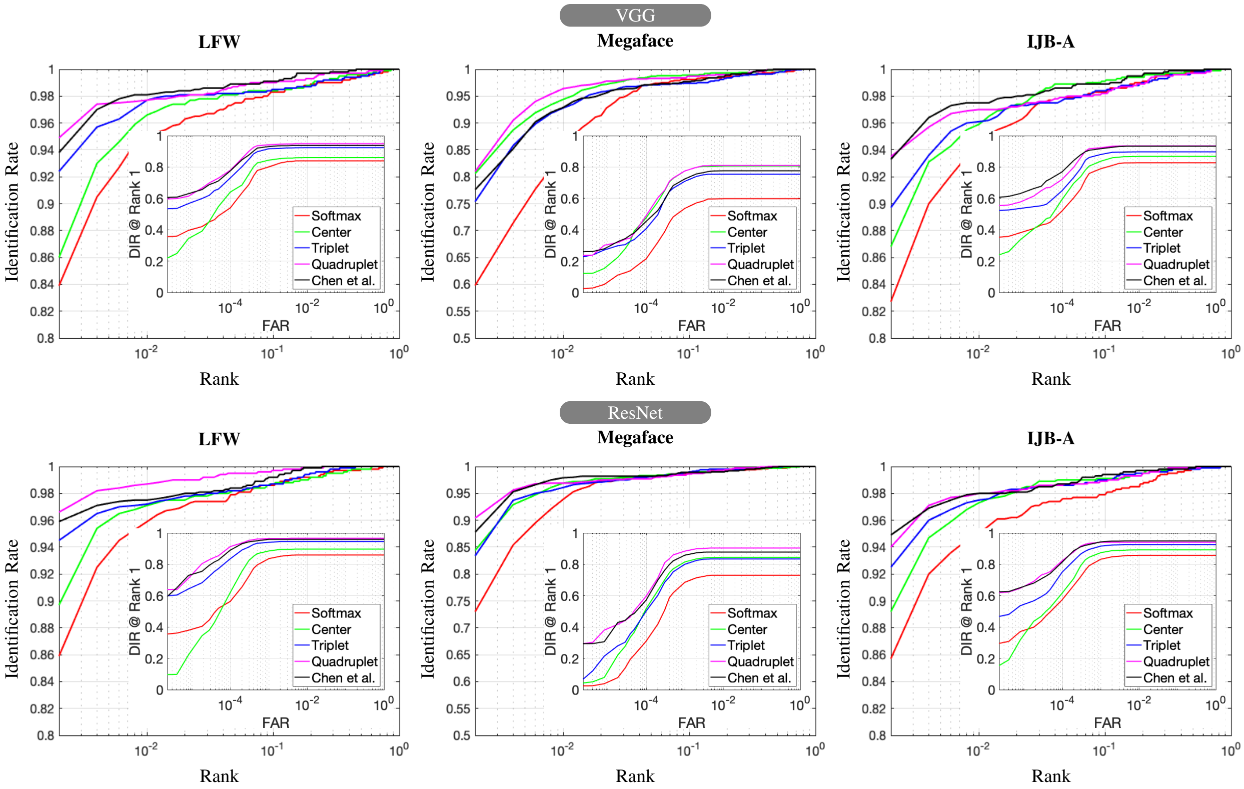

Fig. 8. Identity retrieval results. The outer plots provide the closed-set identification (CMC) curves for the LFW, Megaface and IJB-A sets, using the VGG and ResNet architectures. Inside each plot, the inner regions show the corresponding detection and identification rate (DIR) values at rank-1. Results are shown for the quadruplet loss function (purple color), and four baselines: the *softmax* (red color), center loss (green color), triplet loss (blue color) and Chen *et al*. [5]'s (black color) method.

Generally, we observed that the proposed quadruplet loss outperforms the other loss functions, which might be the result of having used additional information for learning. These improvements in performance were observed in most cases by a consistent margin for both the verification and identification tasks, not only for the VGG but also for the ResNet architecture.

In terms of the errors per CNN architecture, the ResNet-like error rates were roughly 0.9 × (90%) of the observed for the VGG-like networks (higher margins were observed for the *softmax* loss). Not surprisingly, the Chen *et al*. [5]' method outperformed the remaining competitors, followed by the triplet loss function, which is consistent with most of the results reported in the literature. The *softmax* loss got repeatedly the worst performance among the five functions considered.

Regarding the performance per dataset, the values observed for Megaface were far worse for all objective functions than the values for LFW and IJB-A. In the Megaface set, we followed the protocol of the *small* training set, using 490,000 images from 17,189 subjects (images overlapping with Facescrub dataset were discarded). Also, note that the relative performance between the loss functions was roughly the same in all sets. Degradations in performance were slight from the LFW to the IJB-A set and much more visible in case of the Megaface set. In this context, the *softmax* loss produced the most evident degradations, followed by the center loss.

E. Soft Biometrics Inference

As stated above, the proposed loss can also be used for learning a soft biometrics estimator. In test time, the position to where one element is projected is used to infer the soft labels, in a simple nearest neighbour fashion. In these experiments, we considered only 1-NN, i.e., the label inferred for each query was given by the closest gallery element. Better results would be possibly attained if more neighbours had been considered, even though the computational cost of classification will also increase. All experiments were conducted according to a bootstrapping-like strategy: having n test images available, the bootstrap randomly selected (with replacement) $0.9 \times n$ images, obtaining samples composed of 90% of the whole data. Ten test samples were created and the experiments were conducted independently on each trail, which enabled to obtain the mean and the standard deviation at each performance value.

As baselines we used two commercial off-the-shelf (COTS) techniques, considered to represent the state-of-the-art [38]: the Matlab SDK for *Face++*[5] and the *Microsoft Cognitive Toolkit Commercial*[6]. Face++ is a commercial face recognition system, with good performance reported for the LFW face recognition competition (second best rate). Microsoft Cognitive Toolkit is a deep learning framework that provides useful

[5] http://www.faceplusplus.com/
[6] https://www.microsoft.com/cognitive-services/

TABLE I
IDENTITY RETRIEVAL PERFORMANCE OF THE PROPOSED LOSS WITH RESPECT TO THE BASELINES: *softmax*, CENTER AND TRIPLET LOSSES, AND CHEN *et al.* [5]'S METHOD. THE AVERAGE PERFORMANCE ± STANDARD DEVIATION VALUES ARE GIVEN, AFTER 10 TRIALS. INSIDE EACH CELL, VALUES REGARD (FROM TOP TO BOTTOM) THE LFW, MEGAFACE AND IJB-A DATASETS. THE BOLD FONT HIGHLIGHTS THE BEST RESULT PER DATASET AMONG ALL METHODS.

Method	mAP	rank-1	top-10%
VGG			
Quadruplet loss	0.958 ± $3e^{-3}$	**0.951** ± 0.020	**0.979** ± $6e^{-3}$
	0.877 ± 0.011	**0.812** ± 0.053	**0.960** ± $9e^{-3}$
	0.953 ± $5e^{-3}$	**0.939** ± 0.037	0.958 ± $6e^{-3}$
Softmax loss	0.897 ± $4e^{-3}$	0.842 ± 0.034	0.953 ± 0.011
	0.727 ± 0.014	0.615 ± 0.060	0.863 ± 0.017
	0.849 ± 0.010	0.823 ± 0.039	0.941 ± 0.014
Triplet loss [34]	0.934 ± $4e^{-3}$	0.929 ± 0.033	0.964 ± $8e^{-3}$
	0.854 ± $9e^{-3}$	0.758 ± 0.059	0.946 ± 0.017
	0.917 ± $5e^{-3}$	0.901 ± 0.040	0.950 ± 0.011
Center loss [43]	0.918 ± $3e^{-3}$	0.863 ± 0.020	0.962 ± $6e^{-3}$
	0.850 ± 0.013	0.773 ± 0.052	0.939 ± 0.012
	0.862 ± 0.010	0.867 ± 0.041	0.944 ± 0.012
Chen *et al.* [5]	**0.961** ± $2e^{-3}$	0.945 ± 0.022	0.976 ± $6e^{-3}$
	0.864 ± 0.012	0.772 ± 0.061	0.947 ± $9e^{-3}$
	0.948 ± $6e^{-3}$	0.936 ± 0.055	**0.970** ± $4e^{-3}$
ResNet			
Quadruplet loss	**0.968** ± $2e^{-3}$	**0.966** ± 0.012	0.981 ± $4e^{-3}$
	0.902 ± $9e^{-3}$	**0.906** ± 0.048	**0.972** ± $8e^{-3}$
	0.959 ± $3e^{-3}$	0.947 ± 0.021	0.980 ± $4e^{-3}$
Softmax loss	0.912 ± $4e^{-3}$	0.861 ± 0.029	0.960 ± $8e^{-3}$
	0.730 ± 0.010	0.745 ± 0.051	0.899 ± 0.011
	0.841 ± $9e^{-3}$	0.860 ± 0.030	0.958 ± $8e^{-3}$
Triplet loss [34]	0.947 ± $4e^{-3}$	0.948 ± 0.026	0.968 ± $9e^{-3}$
	0.872 ± $8e^{-3}$	0.839 ± 0.052	0.957 ± $9e^{-3}$
	0.919 ± $5e^{-3}$	0.937 ± 0.031	0.961 ± 0.011
Center loss [43]	0.939 ± $3e^{-3}$	0.898 ± 0.016	0.967 ± $6e^{-3}$
	0.847 ± $9e^{-3}$	0.845 ± 0.048	0.945 ± $9e^{-3}$
	0.877 ± $7e^{-3}$	0.893 ± 0.035	0.963 ± $9e^{-3}$
Chen *et al.* [5]	0.966 ± $2e^{-3}$	0.959 ± 0.015	**0.983** ± $4e^{-3}$
	0.916 ± $8e^{-2}$	0.880 ± 0.050	0.975 ± $8e^{-3}$
	0.952 ± $4e^{-3}$	**0.960** ± 0.022	**0.986** ± $6e^{-3}$

information based on vision, speech and language. Also, in order to highlight the distinct properties of the embeddings generated by our proposal with respect to the state-of-the-art, we also measured the soft labelling effectiveness that can be attained by the Triplet loss [34] and Chen*et al.* [43] embeddings if a simple 1-NN rule is used to infer soft biometrics labels.

We considered exclusively the 'Gender', 'Ethnicity' and 'Age' labels ($t = 3$), quantised respectively into two classes for Gender ({'male', 'female'}), three classes for Age ({'young', 'adult', 'senior'}), and three classes for Ethnicity ({'white', 'black', 'asian'}). The average and standard deviation performance values are reported in Table II for the BIODI, PETA and LFW sets.

Overall, the results achieved by the quadruplet loss can be favourably compared to the baseline techniques for most labels, particularly for the BIODI and LFW datasets. Regarding the PETA set, Face++ invariably outperformed the other techniques, even if at a reduced margin in most cases. This was justified by the extreme heterogeneity of image features in this set, in result of being the concatenation of different databases. This should had reduced the representativity of the learning data with respect the test set, being the Face++ model apparently the least sensitive to this covariate. Note that the 'Ethnicity' label is only provided by the Face++ framework. Regarding the Triplet [34] and Chen *et al.* [43] baselines, it is important to note that the reported values were obtained in embeddings that were inferred exclusively based in ID information. Under such circumstances, we confirmed that both solutions produce semantically inconsistent embeddings, in which elements with similar appearance but different soft labels are frequently projected to adjacent regions.

Globally, these experiments supported the possibility of using such the proposed method to estimate soft labels in a *single-shot* paradigm, which is interesting to reduce the computational cost of using specialized third-party solutions for soft labelling.

Finally, we analysed the variations in performance with respect to the number of labels considered, i.e., the value of the t parameter. At first, to perceive how the identity retrieval performance depends of the number of soft labels, we used the annotations provided by the ATVS group [38] for the LFW set, and measured the rank-1 variations for $1 \leq t \leq 4$, starting by the 'ID' label alone and then adding iteratively the 'Gender' → 'Ethnicity' → 'Age' labels. The results are shown in the left plot of Fig. 9. In a complementary way, to perceive the overall labelling effectiveness for large values of t, the BIODI dataset was used (the one with the largest number of annotated labels), and the values obtained for $t \in \{2, \ldots, 14\}$. In all cases, $d = 128$ was kept, with the average labelling error in the test set $\boldsymbol{X}$ given by:

$$e(\boldsymbol{X}) = \frac{1}{n.t} \sum_{i=1}^{n} ||\boldsymbol{p}_i - \boldsymbol{g}_i||_0, \qquad (11)$$

with $\boldsymbol{p}_i$ denoting the t labels predicted for the i^{th} image and $\boldsymbol{g}_i$ being the ground-truth. $|| \, ||_0$ denotes the ℓ_0-norm.

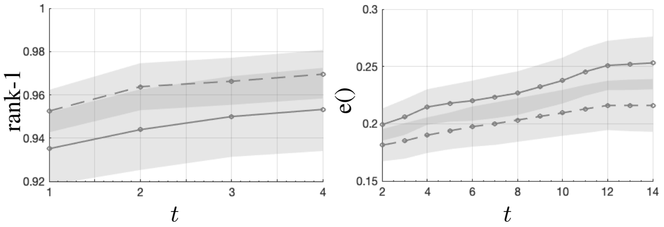

Fig. 9. At left: rank-1 identification accuracy in the LFW dataset, for $1 \leq t \leq 4$. At right: soft biometrics performance in the BIODI test set, for $2 \leq t \leq 14$, for the VGG (solid line) and ResNet (dashed line) architectures.

It is interesting to observe the apparently contradictory results in both plots: at first, a positive correlation between

TABLE II
SOFT BIOMETRICS LABELLING PERFORMANCE (MAP) ATTAINED BY THE PROPOSED METHOD, WITH RESPECT TO TWO COMMERCIAL-OFF-THE-SHELF SYSTEMS (FACE++ AND MICROSOFT COGNITIVE) AND TWO OTHER BASELINES. THE AVERAGE PERFORMANCE ± STANDARD DEVIATION VALUES ARE GIVEN, AFTER 10 TRIALS. INSIDE EACH CELL, THE TOP VALUE REGARDS THE VGG-LIKE PERFORMANCE, AND THE BOTTOM VALUE CORRESPONDS TO THE RESNET-LIKE VALUES.

Method	Gender	Age	Ethnicity
		BIODI	
Quadruplet loss	$0.816 \pm 6e^{-3}$ $\mathbf{0.834} \pm 5e^{-3}$	0.603 ± 0.014 **0.649** ± 0.011	0.777 ± 0.011 $0.786 \pm 9e^{-3}$
Triplet loss [34]	0.684 ± 0.022 0.690 ± 0.019	0.581 ± 0.034 0.584 ± 0.025	0.599 ± 0.028 0.600 ± 0.017
Chen *et al.* [43]	0.693 ± 0.020 0.697 ± 0.015	0.602 ± 0.032 0.604 ± 0.012	0.613 ± 0.019 0.618 ± 0.018
Face++	$0.760 \pm 8e^{-3}$	0.588 ± 0.019	**0.788** ± 0.017
Microsoft Cognitive	$0.738 \pm 7e^{-3}$	0.552 ± 0.026	-
		PETA	
Quadruplet loss	0.862 ± 0.024 0.882 ± 0.018	0.649 ± 0.061 0.658 ± 0.057	0.797 ± 0.053 0.810 ± 0.036
Triplet loss [34]	0.720 ± 0.036 0.722 ± 0.024	0.611 ± 0.038 0.625 ± 0.022	0.612 ± 0.038 0.628 ± 0.026
Chen *et al.* [43]	0.723 ± 0.034 0.731 ± 0.027	0.613 ± 0.037 0.630 ± 0.030	0.636 ± 0.025 0.668 ± 0.021
Face++	0.870 ± 0.028	0.653 ± 0.062	**0.812** ± 0.054
Microsoft Cognitive	**0.885** ± 0.020	**0.660** ± 0.057	-
		LFW	
Quadruplet loss	0.939 ± 0.021 **0.944** ± 0.017	0.702 ± 0.059 0.709 ± 0.049	0.801 ± 0.044 0.817 ± 0.041
Triplet loss [34]	0.794 ± 0.028 0.799 ± 0.022	0.631 ± 0.032 0.636 ± 0.020	0.652 ± 0.022 0.670 ± 0.017
Chen *et al.* [43]	0.794 ± 0.030 0.801 ± 0.021	0.639 ± 0.030 0.659 ± 0.018	0.728 ± 0.027 0.747 ± 0.022
Face++	0.928 ± 0.041	0.527 ± 0.063	**0.842** ± 0.061
Microsoft Cognitive	0.931 ± 0.037	**0.710** ± 0.051	-

the labelling errors and the values of t is evident, which was justified by the difficulty of inferring some of the hardest labels in the BIODI set (e.g., the *type of shoes*). However, the average rank-1 identification accuracy also increased when more soft labels were used, even if the results were obtained only for small values of t (i.e., not considering the particularly hard labels, in result of no available ground truth). Overall, we concluded that the proposed loss obtain *acceptable* performance (i.e., close to the state-of-the-art) when a small number of soft labels is available (≥ 2), but also when a few more labels should be inferred (up to $t \approx 8$). In this regard, we presume that even higher values for t ($t \gg 8$) would require substantially more amounts of learning data and also higher values for d (dimension of the embedding).

F. Semantic Identity Retrieval

Finally, we considered the *semantic identity retrieval* problem, where - along with the query image - semantic criteria are used to filter the retrieved elements (i.e., "*Find this person*" → "*Find this female*", Fig. 10). In this setting, it is assumed that the ground-truth soft labels of the gallery IDs are known, even though the same does not apply for the queries.

We considered the hardest identity retrieval dataset (Megaface) and compared our results to Chen *et al.*'s (the most frequent runner-up in previous experiments). The soft label '*Gender*' (provided by the Microsoft Cognitive Toolkit for the queries) was used as additional semantic data, to filter the retrieved identities. The bottom plot in Fig. 10 provides the results in terms of the hit/penetration rates, being notorious the similar levels of performance of both methods in this setting ('semantic' data series), with Chen *et al.*'s method slightly outperforming up to the top-20 identities, and getting worse results than our solution for the remaining penetration values.

It can be concluded that - when coarse labels are available - our method and Chen *et al.*'s attain similar quality embeddings in terms of compactness and discriminability. However, the key point is that the baseline version of the proposed loss is a way to approximate the results attained by state-of-the-art methods when using semantic information to filter the retrieved identities.

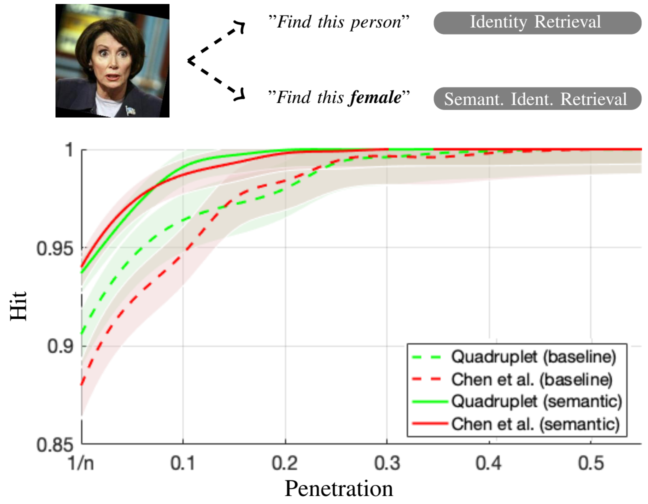

Fig. 10. Comparison between the hit/penetration rates of the proposed loss and Chen *et al.* [5]'s method, when disregarding (baseline) or considering semantic additional information to filter the retrieved results. Values are given for the *ResNet* architecture and Megaface dataset. The '*Gender*' was the semantic criterium in each query and "n" is the number of enrolled identities.

V. CONCLUSIONS AND FURTHER WORK

In this paper we proposed a loss function for multi-output classification problems, where the response variables have dimension greater than one. Our function is a generalization of the well known triplet loss, replacing the *positive/negative* binary division of pairs and the notion of *anchor*, by: i) a metric that considers the *semantic similarity* between any two classes; and ii) a quadruplet term that imposes different distances between pairs of elements according to that similarity.

In particular, we considered the identity retrieval and soft biometrics problems, using the ID and three soft labels ('Gender', 'Age' and 'Ethnicity') to obtain semantically coherent embeddings. In such spaces, not only the intra-class compactness is guaranteed, but also the broad families of classes (e.g., ”white young males” or ”black senior females”) appear in adjacent regions. This enables a direct correspondence between the ID centroids and their semantic descriptions, allowing that simple rules such as k-neighbours are used to jointly infer the identity/soft label information. The insight of the proposed loss is in opposition to single-label loss formulations, where elements are projected into the destiny space based uniquely in ID information and image appearance, being assumed that semantical coherence yields naturally upon the similarity of image features.

As future directions for this work, we are exploring the possibility of fusing the concept described in this paper to the original triplet and Chen *et al*.formulations. In this line of research, the concept of *anchor* will still be disregarded and all images in a triplet will regard different classes (IDs), with the margins imposed according to the soft biometrics similarity between pairs of elements. Also, two other possibilities are: 1) to differently weight the contribution of each soft label in defining the embedding topology; and 2) to consider the conceptual distance inside each label (e.g., ’young’ is closer to ’adult’ than to ’senior’). Both possibilities should also improve the overall ID+soft biometrics labelling performance.

Acknowledgements

This work is funded by FCT/MCTES through national funds and co-funded by EU funds under the project UIDB/EEA/50008/2020. Also, this research was funded by FEDER, Fundo de Coesão e Fundo Social Europeu, under the PT2020 program, in the scope of the POCI-01-0247-FEDER-033395 project. We acknowledge the support of NVIDIA Corporation®, with the donation of one *Titan X* GPU board.

References

[1] N. Almudhahka, M. Nixon and J. Hare. Automatic Semantic Face Recognition. Proceedings of the *IEEE 12th International Conference on Automatic Face & Gesture Recognition*, doi: 10.1109/FG.2017.31, 2017. 2

[2] E. Bekele, C. Narber and W. Lawson. Multi-attribute Residual Network (MAResNet) for Soft-biometrics Recognition in Surveillance Scenarios. Proceedings of the *IEEE 12th International Conference on Automatic Face & Gesture Recognition*, doi: 10.1109/FG.2017.55, 2017. 2

[3] E. Cipcigan and M. Nixon. Feature Selection for Subject Ranking using Soft Biometric Queries. Proceedings of the *15th IEEE International Conference on Advanced Video and Signal-based Surveillance*, doi: 10.1109/AVSS.2018.8639319, 2018. 2

[4] J-C. Chen, R. Ranjan, A. Kumar, C-H. Chen, V. Patel and R. Chellappa. An End-to-End System for Unconstrained Face Verification with Deep Convolutional Neural Networks. Proceedings of the *IEEE International Conference on Computer Vision Workshops*, doi: 10.1109/ICCVW.2015.55, 2015. 2

[5] W. Chen, X. Chen, J. Zhang and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. Proceedings of the *IEEE International Conference on Computer Vision*, doi: 10.1109/CVPR.2017.145, 2017. 3, 7, 8, 9, 10

[6] V. Choutas, P. Weinzaepfel, J. Revaud and C. Schmid. PoTion: Pose MoTion Representation for Action Recognition. Proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, pag. 7024–7033, doi: 1109/CVPR.2018.00734, 2018. 1

[7] Y. Deng, P. Luo, C. Loy and X. Tang. Pedestrian attribute recognition at far distance. Proceedings of the *ACM International Conference on Multimedia*, pag. 789–792, doi: 10.1145/2647868.2654966, 2014. 4

[8] J. Deng, J. Guo, N. Xue and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. https://arxiv.org/abs/1801.07698, 2019. 2

[9] Y. Duan, J. Lu and J. Zhou. UniformFace: Learning Deep Equidistributed Representation for Face Recognition. https://arxiv.org/abs/1801.07698, 2019. 2

[10] H. Galiyawala, K. Shah, V. Gajjar and M. Raval. Person Retrieval in Surveillance Video using Height, Color and Gender. https://arxiv.org/abs/1810.05080, 2018. 2

[11] J. Goldberger, G. Hinton, S. Roweis and R. Salakhutdinov. Neighbourhood Components Analysis. Proceedings to the *Advances in Neural Information Processing Systems* Conference, vol. 17, pag. 513–520, doi: 10.5555/2976040.2976105, 2005. 4

[12] B. Guo, M. Nixon and J. Carter. Fusion Analysis of Soft Biometrics for Recognition at a Distance. Proceedings of the *IEEE 4th International Conference on Identity, Security, and Behavior Analysis*, doi: 10.1109/ISBA.2018.8311457, 2018. 2

[13] B. Guo, M. Nixon and J. Carter. A Joint Density Based Rank-Score Fusion for Soft Biometric Recognition at a Distance. Proceedings of the *International Conference on Pattern Recognition*, pag. 3457–, 3460, doi: 10.1109/ICPR.2018.8546071, 2018. 2

[14] R. Hadsell, S. Chopra and Y. LeCun. Dimensionality reduction by learning an invariant mapping. Proceedings of the *IEEE International Conference on Computer Vision*, doi: 10.1109/CVPR2006.100, 2006. 2

[15] M. Halstead, S. Denman, C. Fookes, Y. Tan and M. Nixon. Semantic Person Retrieval in Surveillance Using Soft Biometrics: AVSS 2018 Challenge II. Proceedings of the *IEEE International Conference on Advanced Video Signal-based Surveillance*, doi: 10.1109/AVSS.2018.8639379, 2018. 1

[16] E. Learned-Miller, G. Huang, A. RoyChowdhury, H. Li and G. Hua. Labeled Faces in the Wild: A Survey. In *Advances in Face Detection and Facial Image Analysis*, Michal Kawulok, M. Emre Celebi and Bogdan Smolka (eds.), Springer, pag. 189–248, doi: 10.1007/978-3-319-25958-1_8, 2016. 4

[17] K. He, Z. Wang, Y. Fu, R. Feng, Y-G. Jiang and X. Xue. Adaptively Weighted Multi-task Deep Network for Person Attribute Classification. Proceedings of the *4th ACM Multimedia* conference, doi: 10.1145/3123266.3123424, 2017. 2

[18] Y. Hu, X. Wu and R. He. Attention-Set Based Metric Learning for Video Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pag. 2827–2840, 2018. 2

[19] S. Ji, W. Xu, M. Yang, K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pag. 221–231, 2019. 1

[20] M. Jiang, Z. Yang, W. Liu and X. Liu. Additive Margin Softmax with Center Loss for Face Recognition. Proceedings of the *2nd International Conference on Video and Image Processing*, pag. 1–8, doi: 10.1145/3301506.3301511, 2018. 2

[21] B-N. Kang, Y. Kim and D. Kim. Deep Convolution Neural Network with Stacks of Multi-scale Convolutional Layer Block using Triplet of Faces for Face Recognition in the Wild. Proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPRW.2017.89, 2017. 2

[22] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pag. 4873–4882, doi: 10.1109/CVPR.2016.527, 2016. 5

[23] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pag. 1931–1939, doi: 10.1109/CVPR.2015.7298803, 2015. 5

[24] F. Lateef and Y. Ruichek Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, vol. 338, pag. 321–348, 2019. 1

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C-Y. Fu and A. Berg. SSD: Single Shot MultiBox Detector. Proceedings of the *European Conference on Computer Vision*, pag. 21–37, doi: 10.1007/978-3-319-46448-0_2, 2016. 1

[26] T. Neal and D. Woodard. You Are Not Acting Like Yourself: A Study on Soft Biometric Classification, Person Identification, and Mobile Device Use. *IEEE Transactions on Biometrics, Behaviour and Identity Science*, vol. 1, no. 2, pag. 109–122, 2019. 2

[27] D. Li, Z. Zhang, X. Chen and K. Huang. A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios *IEEE Transactions on Image Processing*, vol. 28, no. 4, pag. 1575–1590, 2019. 1

[28] H. Liu and W. Huang. Body Structure Based Triplet Convolutional Neural Network for Person Re-Identification. Proceedings of the *IEEE International Conference on Acoustics, Speech and Signal Processing*, doi: 10.1109/ICASSP.2017.7952461, 2017. 2

[29] D. Martinho-Corbishley, M. Nixon and J. Carter. Super-Fine Attributes with Crowd Prototyping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pag. 1486–1500, 2019. 2

[30] R. Ranjan, A. Bansal, S. Sankaranarayanan, J-C. Chen, C. Castillo and R. Chellappa. Crystal Loss and Quality Pooling for Unconstrained Face Verification and Recognition. https://arxiv.org/abs/1804.01159, 2018. 2

[31] R. Vera-Rodriguez, P. Marin-Belinchon, E. Gonzalez-Sosa, P. Tome and J. Ortega-Garcia. Exploring Automatic Extraction of Body-based Soft Biometrics. Proceedings of the IEEE International Carnahan Conference on Security Technology, doi: 10.1109/CCST.2017.8167841, 2017. 2

[32] P. Samangouei and R. Chellappa. Convolutional neural networks for attribute-based active authentication on mobile devices. Proceedings of the IEEE 8^{th} International Conference on Biometrics Theory, Applications and Systems, 1–8, doi: 10.1109/BTAS.2016.7791163, 2016. 1, 2

[33] A. Gretton, K. Borgwardt, M. Rasch, B. Schlkopf and J. Smola A kernel method for the two-sample-problem. Proceedings of the *Advances in Neural Information Processing Systems* Conference, pag. 513–520, doi: 10.5555/2976456.2976521, 2006. 2

[34] F. Schroff, D. Kalenichenko, J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. Proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2015.7298682, 2015. 1, 4, 5, 7, 9, 10

[35] A. Schumann and A. Specker. Attribute-based person retrieval and search in video sequences. Proceedings of the *IEEE International Conference on Advanced Video Signal-based Surveillance*, doi: 10.1109/AVSS.2018.8639114, 2018. 2

[36] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang and S. Li. Constrained Deep Metric Learning for Person Re-identification. https://arxiv.org/abs/1511.07545, 2015. 1, 2

[37] H. Song, Y. Xiang, S. Jegelka and S. Savarese. Deep Metric Learning via Lifted Structured Feature Embedding. Proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2016.434, 2016. 2

[38] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez and F. Alonso-Fernandez. Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation, and COTS Evaluation. *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pag. 2001-2014, 2018. 2, 8, 9

[39] C. Su, Y. Yan, S. Chen and H. Wang. An efficient deep neural networks framework for robust face recognition. Proceedings of the *IEEE International Conference on Image Processing*, doi: 10.1109/ICIP.2017.8296993, 2017. 2

[40] F. Wang, W. Zuo, L. Lin, D. Zhang and L. Zhang. Joint learning of single- image and cross-image representations for person re-identification. Proceedings of the *IEEE International Conference on Computer Vision and Pattern Recognition*, doi: 10.1109/CVPR.2016.144, 2016. 1

[41] J. Wang, Z. Wang, C. Gao, N. Sang and R. Huang. DeepList: Learning Deep Features With Adaptive List-wise Constraint for Person Re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pag. 513–524, 2017. 1, 3

[42] Y. Wen, K. Zhang, Z. Li and Y. Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition Proceedings of the *14^{th} European Conference on Computer Vision*, doi: 10.1007/978-3-319-46478-7_31, 2016. 2

[43] Y. Wen, K. Zhang, Z. Li and Y. Qiao. A Comprehensive Study on Center Loss for Deep Face Recognition. *International Journal of Computer Vision*, vol. 127, pag. 668–683, 2019. 9, 10

[44] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba and A. Oliva. Learning deep features for scene recognition using places database. Proceedings of the *Advances in Neural Information Processing Systems* Conference, pag. 487–495, doi: 10.5555/2968826.2968881, 2014. 1