applied sciences

MDPI

Article

Combining Human Parsing with Analytical Feature Extraction and Ranking Schemes for High-Generalization Person Reidentification

Nikita Gabdullin [1,2]

[1] Joint Stock "Research and Production Company "Kryptonite", 115114 Moscow, Russia; n.gabdullin@kryptonite.ru
[2] Research and Education Working Group, Chung-Ang University, Seoul 06973, Republic of Korea

Abstract: Person reidentification (re-ID) has been receiving increasing attention in recent years due to its importance for both science and society. Machine learning (particularly Deep Learning (DL)) has become the main re-ID tool that has allowed to achieve unprecedented accuracy levels on benchmark datasets. However, there is a known problem of poor generalization in respect of DL models. That is, models that are trained to achieve high accuracy on one dataset perform poorly on other ones and require re-training. In order to address this issue, we present a model without trainable parameters. This, in turn, results in a great potential for high generalization. This approach combines a fully analytical feature extraction and similarity ranking scheme with DL-based human parsing wherein human parsing is used to obtain the initial subregion classification. We show that such combination, to a high extent, eliminates the drawbacks of existing analytical methods. In addition, we use interpretable color and texture features that have human-readable similarity measures associated with them. In order to verify the proposed method we conduct experiments on Market1501 and CUHK03 datasets, thus achieving a competitive rank-1 accuracy comparable with that of DL models. Most importantly, we show that our method achieves 63.9% and 93.5% rank-1 cross-domain accuracy when applied to transfer learning tasks, while also being completely re-ID dataset agnostic. We also achieve a cross-domain mean average precision (mAP) that is higher than that of DL models in some experiments. Finally, we discuss the potential ways of adding new features to further improve the model. We also show the advantages of interpretable features for the purposes of constructing human-generated queries from verbal descriptions in order to conduct searches without a query image.

Keywords: person reidentification; re-id; human parsing; analytical features; similarity ranking; generalization

Citation: Gabdullin, N. Combining Human Parsing with Analytical Feature Extraction and Ranking Schemes for High-Generalization Person Reidentification. *Appl. Sci.* **2023**, *13*, 1289. https://doi.org/10.3390/app13031289

Academic Editors: Wonjoon Kim, Sekyoung Youm and Sungbum Jun

Received: 1 December 2022
Revised: 11 January 2023
Accepted: 16 January 2023
Published: 18 January 2023

1. Introduction

Person re-identification (re-ID) is becoming one of the most significant research topics in the domains of computer vision and computational intelligence, due to its two-fold importance for both science and society. It focuses on person identification across camera systems by addressing the increasing demand for public safety. The problem regarding person identification commonly resides with the person detection and tracking tasks [1]. In respect of this, the reidentification task is commonly formulated as follows: a person in a query image is to be matched with a person in an image, or images, that are obtained from data streams of different cameras, or a given camera, at various moments in time. A database of images or videos is often used instead of live streams. The main challenge is to assess the similarity between objects while taking into account possible changes in human appearance due to variations in camera viewpoints, lighting conditions, the person's pose, and occlusions. In this paper, we focus on the similarity assessment; while for the

database search techniques, the reader is referred to the literature [2] with a special mention of the emerging graph-based methods [3,4].

In practice, re-ID person similarity evaluation is conducted using images obtained from video frames. Recently, various video re-ID approaches have emerged that effectively use temporal information in order to improve the assessment accuracy [5,6]. Nevertheless, the visual similarity estimation remains the core of re-ID. In addition, the methods that perform well on images can be expected to perform even better when combined with temporal data.

The fast development in the area of convolutional neural networks (CNN), as well as deep learning (DL), that affected numerous areas of computer science also led to the emergence of machine learning and DL-assisted re-ID techniques [7,8]. Re-ID requires CNNs to assess the similarity for image pairs, which is not typical for conventional image classification tasks. This inspired researchers to propose novel network architectures, such as: Siamese networks [9,10]; specialized loss functions, e.g., triplet loss [11,12]; specific attention modules [13]; re-ID graph neural networks [14]; and others. Moreover, the augmented images were often included in training datasets in order to simulate illumination-related effects [15,16]. Many early works focused on using CNNs for either feature extraction or metric learning [17,18], with end-to-end models gradually becoming dominant in the field.

However, generalization of the results is challenging for DL-assisted re-ID. That is, models trained on specific datasets tend to perform poorly on other data, as can be illustrated by the fact that models trained on Market1501 to 98% rank-1 accuracy reach only 38% accuracy on DukeMTMC [19]. The effects of negative transfer learning can be drastic due to significant variations in data contents, e.g., the different clothes people wear, different camera types, different filming environments, etc. A combination of such effects is often referred to as the "Open-World re-ID problem". In addressing this problem, a further complication of the CNN architecture [18,20] is required. Therefore, models become extremely bulky with significant parameter redundancy that reduces speed, as well as increases computational power and storage demands. Thus, generalization implies overparameterization, which increases costs and does not guarantee good performance.

In order to address this issue, we focus on analytical techniques to obtain a compact and fast-performing model. This is performed in order to reduce computational costs and to improve generalization. This is achieved by constructing compact interpretable object descriptors (i.e., feature vectors) that are combined with a similarity ranking scheme. Analytical models, while falling out of favor recently due to growing interest towards DL models, previously showed promising results on re-ID tasks [21,22]. This was mainly achieved by constructing an ensemble of features while utilizing a trainable model that finds feature weights in a manner similar to metric learning [21,23]. This approach is different to CNN techniques that work on images with nearly no preprocessing and generate feature vectors that are incomprehensible for human operators. On the contrary, analytical feature extraction makes it possible to construct human-readable features that improve interpretability of the results. In this work, we focus on color and texture features. When comparing two objects, we calculate vectors with elements representing percentages of similarity with respect to specific features.

A drawback of old analytical models was found in the fact that feature vectors were global, i.e., generated for the whole image while possibly including the background and other objects. Thus, it was impossible to obtain features of specific elements, such as clothes or hair color. Now such separation into different elements can be achieved using human parsing that allows one to divide an object into class-specific subregions. This makes it possible to generate feature vectors for specific classes and to assess their similarity class by class. The overall similarity is obtained by combining class similarities, which are weighted by class importance. It acts as a score for obtaining a similarity ranking for query images by matching them with test images. The overall matching accuracy is evaluated with a rank-r matching rate [7,24,25]. In this paper, we propose a model without trainable parameters,

which is immune to the dataset-dependent negative effects that emerge due to data fitting techniques used by CNNs or trainable analytical models.

The rest of the paper is organized as follows: in Section 2, the methodology is discussed; in Section 3, the experiments and results are summarized; in Section 4, discussions on the topic are presented; and in Section 5, the paper is concluded.

2. Methodology

2.1. Overview of the Method

The proposed method consists of four major steps, as shown in Figure 1: the parsing of the query image, feature extraction for each subregion (class), the similarity score calculation between each query and the test set images that are utilized in order to calculate rank-r ranking results. It should be noted that for the best performance, it is necessary to store parsing masks and feature vectors in the database along with the original images. This allows image parsing and feature extraction operations to be conducted once per image.

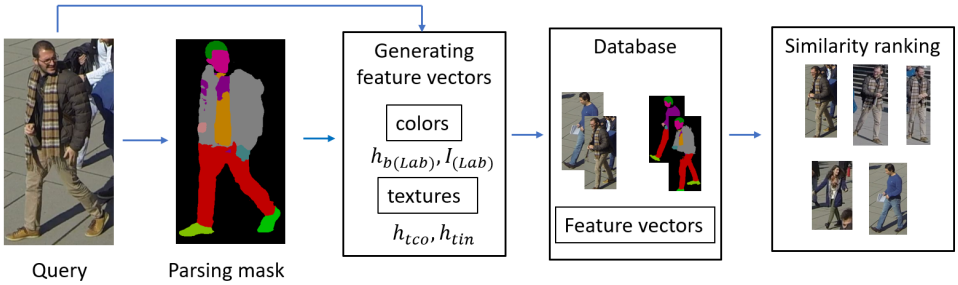


Figure 1. Method overview: parsing of the query image for the per-class feature extraction with consequent comparison with dataset images, which is conducted in order to form a similarity ranking.

2.2. Image Parsing

The original image processing techniques utilized in analytical re-ID were applied to full images, primarily due to the lack of reliable segmentation techniques [21]. This resulted in feature vectors of different subregions becoming mixed, thus complicating the similarity estimation. The implementation of preliminary image parsing—as well as, specifically, human parsing—is a promising technique that allows one to divide an object into subregions that correspond to specific classes. Whereas clustering-based segmentation is well-known, it is also computationally expensive and does not provide sufficient accuracy. Spatiotemporal features, which were based on a graph partitioning of the full image, were proposed to localize feature vectors to specific regions [22]. More recently, a simple approach was proposed that leveraged the fact that body parts in different images ordinarily appeared in the same image sections in re-ID datasets. This allowed one to separate images into three or more subregions, as well as to calculate local feature vectors for horizontal strips [26–28]. In this way, the information about the head, torso, and legs could be obtained. Whereas otherwise being computationally convenient, this method was not robust in scenarios that included pose change and partial occlusion. Furthermore, the information regarding the difference between elements within one strip, such as hands and torso, was lost. This was addressed via several modifications that were based on a more detailed representation of the human body structure [5,29]. Nevertheless, the issues of background subtraction, the exact region boundary estimation, as well as the body part or clothes classification remained a challenge—which human parsing allowed one to address.

The main goal of human parsing is to provide a mask that assigns a specific class to every image pixel. For re-ID purposes, these include body parts, hair, clothes, and other wearables. Whereas this task can also be performed by more sophisticated general purpose image segmentation tools, they are unnecessarily complex due to redundant classes, such as furniture, animals, or vehicles—which are all irrelevant for the purposes of person segmentation [30]. Human parsing is usually performed by neural networks that are trained on human parsing datasets, e.g., LIP or Pascal [31,32]. Several studies have shown that

incorporating human parsing into a re-ID framework significantly improves the prediction accuracy [25,33]. However, existing solutions still utilize additional neural networks for feature extraction, metric learning, feature importance estimation (or attention), and decision making, thereby resulting in complex architectures [25,33–35].

In this paper, we propose to combine recent advances in human parsing with analytical feature extraction. Human parsing naturally handles aspects that are challenging for analytical methods by providing the initial segmentation of an object into subregions. As every subregion is assigned its class, this allows a comparison of feature vectors of the same class without dealing with the noise originating from neighboring subregions. The shape of the subregions mimics the real shape of the elements in the image, which pairs well with methods that do not require one to apply filters to subregions for the purposes of feature extraction.

In this work, we use an out-of-the-box human parser SCHP [36,37]. This parser was trained on LIP dataset for human parsing by authors of the original study. It should be stressed that the parser was used "as is" and that no additional training was performed in this study. For any pair of images that were parsed, feature vectors were compared only for classes present in both images. We merged several semantically similar LIP classes, i.e., upper clothes, dresses, coats, and jumpsuits, which were merged into "upper clothes" (class 5); in addition, pants and skirts were merged into "pants" (class 9). Thus, we worked with fifteen unique classes out of the twenty original LIP ones. The merging was performed in order to avoid a common human parsing problem: the fact that semantically similar classes—e.g., coats and upper clothes—can become easily mixed up. Although knowing whether a person is wearing a coat, or a shirt, is indeed valuable and that additional logic can be built on this knowledge, it cannot be relied upon in the current state. Therefore, we do not require the parsers to make such distinctions correctly.

2.3. Color Similarity

2.3.1. Choice of Color Space and Histogram Modification

Although the RGB color scheme, or color space, is widely used in image encoding, it is not particularly suitable for the image processing that is performed for identification purposes. Therefore, other color spaces are used instead. These commonly include: HSV, CMYK, YCbCr, and others [21,38]. Indeed, HSV color space and, specifically, *H* channel information was found to be the most descriptive by several researchers who worked on color similarity for re-ID [15,38]. However, the highest accuracy results were commonly obtained using an ensemble of features, which could include a variety of different channels from varying color spaces [21].

There is also a problem that these color spaces are perceptually non-uniform with non-metric distances. On the contrary, CIE-Lab (Lab) color space is a uniform color space where Euclidian distance can be used as a metric for color difference calculation [39]. Although less popular in comparison to other color spaces, it was previously, and successfully, applied to image similarity estimations [40]. In this paper, we use Lab color space to leverage the combination of perceptual uniformity with metric distance measurement in order to create a two-fold color similarity estimate, which is achieved by comparing the histograms of Lab channels. It is also essential that the effects of illumination changes that complicate color comparison are localized in the lightness (*L*) channel. We, thus, propose two approaches to illumination change handling based on *L* channel histogram analysis. It should be mentioned that these approaches were developed while working with the Wildtrack dataset [41].

Firstly, the *L* channel histograms (Figure 2b), which are extracted for every subregion class of an input image (Figure 2a), are "stretched". In order to achieved this, the number of bins in the original histogram was first reduced from 256 to 64 by averaging the pixel values of the neighboring bins. Then, Algorithm 1 was used to obtain a histogram, as is shown in Figure 2c.

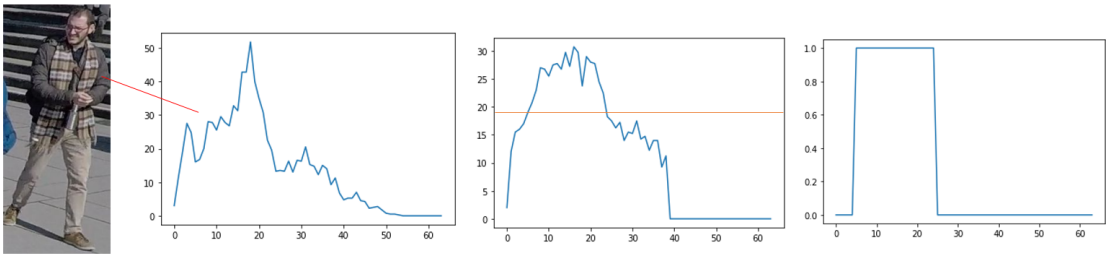
This allows one to normalize the lightness levels among the images that were taken in different lighting conditions. Algorithm 1 is different from conventional histogram stretching in that it does not force the new histogram to occupy the whole range. In addition, the range of "stretching" depends on the extent that the number of pixels in certain bins exceeds the average pixel number. In other words, the range of stretching is proportional to the excess of pixels in certain bins.

Before stretching, a check for over-illumination is also performed. By performing this, it was noticed that for images where the number of pixels in the last bin exceed 1% of all pixels prior to the proposed modification, L channel histograms become unreliable for the purposes of color comparison—which, in turn, was due to a loss of information. The over-illuminated classes were not included into the similarity estimation.

Secondly, shadows can lead to a significant distortion of color similarity estimation. Local shadows appearing due to local occlusions further complicate the situation, as such effects cannot be handled by the first approach. It was noticed that in many cases, local shadows form a distinct peak on L a channel histogram, which precedes the peak related to the most prominent *real* color, as shown in Figure 3. This suggests that removing the first peak may significantly improve the analysis accuracy. However, regions related to shadows are harder to locate for low-resolution or poorly lit images. Moreover, naively removing a part of the histogram results in a worse accuracy in such cases. As such, the search for a comprehensive approach to handling local shadows is still being conducted.

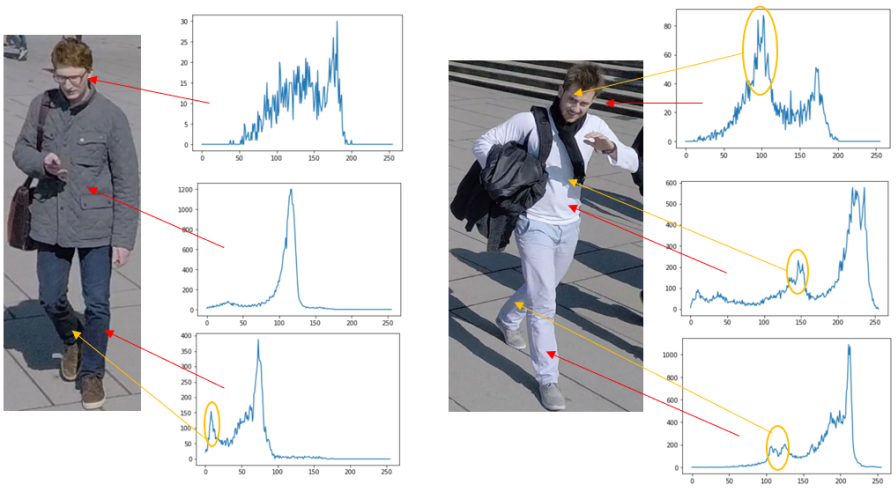
Algorithm 1 L channel histogram modification

Require: h is a 64 bin L channel histogram
1: Calculate h_{av} average of h
2: Zero bin i if $h_i < h_{av}$ for all bins
3: Calculate new average h'_{av}
4: $m \leftarrow argmax(h)$
5: **for** $i > m$ **do**
6: **if** $h(i) > h'_{av}$ **then**
7: $E \leftarrow 0.5(h(i) - h'_{av})$
8: $h(i) \leftarrow h(i) - E$
9: add 0.25E each to the next four bins
10: **end if**
11: **end for**
12: Repeat for $i < m$
13: Zero the first and the last non-zero bins



(a) (b) (c) (d)

Figure 2. Histogram modification and binarization: (**a**) query image; (**b**) L channel histogram of one of the classes (e.g., upper clothes); (**c**) modified histogram; and (**d**) binarized histogram after thresholding.



(a) (b)

Figure 3. The correlation between shadows and peaks in the L channel histograms of face, upper clothes, and pants subregions: (**a**) prominent shadows appear only on pants and (**b**) shadows appear in every subregion. Yellow arrows show correspondence between histogram regions and shadows in images.

2.3.2. Representative Color Intensities and Histogram Thresholding

Possible changes in respect of the view angle present another challenge for color comparison methods. They can lead to the same distinct areas appearing to have different sizes, thus resulting in the Lab channels' histograms for the same classes of the same object to shift dramatically. This renders bin-to-bin histogram comparison methods to be less accurate, which has inspired the development of other comparison metrics [40,42]. However, they still may entail misleading results due to drifts in pixel intensities, which are caused by illumination and point-of-view changes. In order to address this issue, we propose a simple thresholding scheme, with the purpose of finding the most representative pixel intensities.

Firstly, the number of bins in the color channel histograms was reduced from 256 to 64, as was the case for the L channel histograms in Section 2.3.1. This reduces the influence of the drifting effects. Secondly, for every bin j of the original histogram h_o of channel i (where i = L,a,b), where the L channel histogram is "stretched", we assign "1" in its binarized version h_b when the number of pixels in that bin exceeds the threshold k_{bi}. Then, we assign "0" as otherwise shown in Figure 2d, such that:

$$h_{bij} = \begin{cases} 1 \; if \; h_{oij} \geq k_{bi}, \\ 0 \; otherwise. \end{cases} \tag{1}$$

$$k_{bi} = k_{inc} \frac{\sum_{j=0}^{255} h_{bij}}{256}, \tag{2}$$

where k_{inc} is an empirical coefficient that increases the threshold value over a simple average. In our experiments, k_{inc} = 1.5 was found to provide the best results. This approach also allows us to design a comparison scheme for a pair of histograms. Provided we have two binarized histograms h_{bi1} and h_{bi2} of the same class in two images, we find their sum h_{ti}, which can only consist of zeroes, ones, and twos. These thus define the color channel similarity measures as:

$$h_{ti} = h_{bi1} + h_{bi2}, \tag{3}$$

$$S_i = \frac{s_{2i}(h_{ti})}{s_{1i}(h_{ti}) + s_{2i}(h_{ti})}, \tag{4}$$

where s_1 and s_2 are the numbers of times "1" or "2", which appeared in h_t, respectively. The intuition behind this is to find the intensities that are prominent in both histograms, as well as to obtain their fraction relative to the total number of non-zero bins, such that S_i lies in [0, 1] range. This operation is very similar to histogram intersection [42]. However, it does not require h_{bi1} and h_{bi2} to have the same number of non-zero bins and it is commutation-invariant. The former is very important due to the fact that the number of non-zero intensities after thresholding can vary significantly for different images.

2.3.3. Distance in Lab Color Space as Similarity Measure

It was previously mentioned that distances in Lab color space obey Euclidian distance equation. This allows one to judge how similar two colors are depending on how close their corresponding Lab vectors lie. As we deal with histograms rather than actual colors, we first propose to calculate the average intensity for the histograms in every channel, as well as to treat such triplets as the coordinates of a point in Lab space. In this case, the number of pixels h_{oij} in bins j of the original histogram h_o act as weights so that the average intensity in color channel i is calculated as:

$$I_i = \frac{\sum_{j=0}^{255} jh_{oij}}{\sum_{j=0}^{255} h_{oij}}. \tag{5}$$

For the image segments of a uniform color triplet (I_L, I_a, I_b), it is expected to lie close to the main color that is perceived by the human eye. In the case of multiple subregions with different colors, it may not correspond to any real color that is present in the image. However, this is sufficient for distance calculation purposes. As such, for two segments of the same class:

$$d = \sqrt{(I_{L1} - I_{L2})^2 + (I_{a1} - I_{a2})^2 + (I_{b1} - I_{b2})^2}. \tag{6}$$

We then propose a normalized similarity measure S_d with values in [0, 1] range, which is related to d as:

$$S_d = \begin{cases} 1 - \frac{d}{k_d} \; if \; d < k_d, \\ 0 \; otherwise, \end{cases} \tag{7}$$

where k_d is an empirical distance coefficient and, in this work, $k_d = 35$.

2.4. Texture Similarity

Analytical texture descriptors are often separated into two broad categories: shape abstraction and local texture descriptors. Prominent examples of the former are skeletal [43] and morphological [44] filters that deal with the overall geometrical structure, the object's shape, and the subregion boundaries. However, human parsing allows us to segment images and classify subregions, so the shape descriptors provide insufficient information. This is because the two objects that are already classified as "upper clothes" or "pants" have a very similar structure. Furthermore, shape descriptors are not necessarily affine-invariant, which is a disadvantage for re-ID applications.

In this work, we focus on the local texture descriptors that extract subregion-specific information about the object's texture. This is achieved using filters or operations on the pixels and their neighborhoods [45]. In prior work, filter packs that included Gabor filters were found to perform well on re-ID tasks [21]. However, such filters provided desirable translation and affine-invariant texture descriptors. As such, the necessity to fine-tune filter parameters led to a significant number of filters that were required to analyze a sufficiently vast variety of different textures. To avoid this, we have chosen the local binary pattern (LBP) method, due to its ability to generate texture descriptors.

LBP is a local texture descriptor that was found to be extremely useful in texture and face classification. It assigns every pixel of a grayscale image a binary number with bits representing whether the pixel neighbor's intensity is greater than that of a central pixel.

In the simplest case of only checking the closest neighbors (radius $r = 1$), we obtain an 8-bit binary number and convert it to decimal. All decimal numbers of the image are combined into a 256-bin histogram. It is important to note that every bin corresponds to a unique texture type, which can be significantly different for neighboring bins.

Over the years, various modifications to the original LBP have been proposed, including multiscale extensions [46], spatial enhancements [47], and texture uniformity classifications [48]. These methods often include a concatenation of several local histograms in order to encode additional spatial information. However, such modifications are not suitable for re-ID purposes, due to the fact that they are not translation and rotation invariant. This is acceptable for face recognition tasks, where distinct facial features are always localized in the same regions of an image. However, this is not suitable for re-ID images, where the point of view and object orientation can change dramatically. The separation of textures into uniform and non-uniform classes, which was found to be useful in many face recognition tasks, did not perform well here. The reason lies in the significant reduction in the number of non-zero bins and the loss of some descriptive information. Whereas for the face recognition tasks, uniform patterns were found to be the most informative. However, this appears to not be the case for re-ID tasks.

Therefore, we use the original LBP method with one slight modification. It was shown that LBP can be made rotation and affine transformation invariant by performing a circular bit-wise right shift on the binary numbers [49]. Whereas this also reduces the number of possible non-zero bins, these invariances are essential for re-ID. As such, this modification was found to improve the performance in our experiments.

In this work, we calculate two LBP histograms for every class of interest in the image. They correspond to the LBP descriptions of the subregion's contour and its inner area. It was found that separating them positively affects the similarity estimation, due to the fact that contour similarity varies in a wider range. In addition, its variations are not trivial with respect to changes in the inner area's similarity. As a similarity measure for the contour or inner area (channel $m = co, in$), histograms of the two images are used for the histogram intersection. In order to obtain a commutation invariant similarity measure for a pair of histograms, the histograms are first normalized with respect to the total number of points in all bins in the histogram yielding modified histograms h_{tm1} and h_{tm2}, such that values in the bins show the percentages of all textures as corresponding to those bins, so that:

$$S_m = \sum_{j=0}^{255} min(h_{tm1j}, h_{tm2j}) \tag{8}$$

with values in [0, 1] range. In our experiments, similarity estimates obtained using texture and color descriptors did not provide overlapping results, thereby implying that there is no trivial relationship between features. This is essential, because texture descriptors can address cases that are challenging for color similarity estimations and vice versa.

2.5. Similarity Score Calculation for Similarity Ranking

Previous sub-sections discuss six feature channels (S_f), with similarity measures S_L, S_a, S_b, S_d, S_{co}, and S_{in}. In general, for a class c with multiple feature channels, class similarity is calculated as:

$$S_c = \sum_f w_f S_f \tag{9}$$

$$\sum_f w_f = 1 \tag{10}$$

where w_f is the weight of a feature channel f. Table 1 shows the w_f values used in this study.

Table 1. Feature weights w_f that determine feature importance in class similarity evaluations.

Feature	L	a	b	d	t_{in}	t_{co}
Weight	0.13	0.13	0.13	0.31	0.15	0.15

It is worth pointing out that their sum equals one, which ensures that S_c is in [0, 1] range. For a query-test pair with n-shared classes, the total similarity score is calculated as:

$$S_{sim} = \sum_{c}^{n} w_c S_c \tag{11}$$

$$S_{simn} = \frac{S_{sim}}{\sum_{c}^{n} w_c} \tag{12}$$

where w_c are class weights shown in Table 2.

Table 2. Class w_c weights determining the class importance in similarity score calculations.

Parsing Class	Hair, Socks, Face, Legs, Arms	Hat, Gloves, Sunglasses, Shoes	Scarf	Pants	Upper Clothes
Weight	1	2	3	6	8

It should be noted that, unlike channel-wise similarity estimates, the similarity score S_{sim} does not naturally fall in [0, 1] range. In general, S_{sim} depends on the number of shared classes n, which varies from one to the number of classes in the query image. The highest score would be obtained via a pair of images with all possible classes having a perfect similarity among all features—which, for the weights in Table 2, amounts to 34. In practice, it is extremely rare to have objects with all fifteen different classes present in both images and where the similarity scores are much lower. Whereas a percentage similarity measure can be obtained using (12), it is less meaningful when understood as the similarity score due to the fact that an image pair with fewer classes and with low w_c but high S_c will have higher S_{simn} than another pair with many important classes but slightly lower S_c. Thus, we propose to use S_{sim}, rather than S_{simn} as the scores for ranking.

In order to calculate a rank-r matching rate for a query image, its scores S_{sim} with all test images are calculated and sorted in descending order. This is performed with the r highest score test images returned for match checking, as per the following:

$$\text{rank-r} = \begin{cases} 1 \; if \; true \; match \; in \; r \\ 0 \; otherwise \end{cases} \tag{13}$$

Most popular rank-r metrics are rank-1, rank-5, and rank-10. In order to evaluate method performance on a dataset with multiple queries, the rank-r accuracy is averaged for all queries. The main disadvantage of utilizing a rank-r metric is that it requires a single match in r, which is usually the simplest to perform among the possible tests. Thus, it does not contain information about other possible matches and does not reflect the ability of the method to capture complicated matches. As such, in order to address this problem, the mean average precision (mAP) metric is often used along with rank-r when there is more than one correct match in the test set [7,25]. In addition, this is calculated as:

$$mAP = \frac{1}{n_q} \sum_{i} \frac{1}{n_{TPi}} \sum_{j} \frac{m_{ij}}{j} \tag{14}$$

$$m_{ij} = \begin{cases} cumulative \; number \; of \; true \; matches \; for \; query \; i \; found \; at \; step \; j \\ 0 \; otherwise \end{cases} \tag{15}$$

where n_q is the number of queries in a query set, and n_{TPi} is the number of true matches for an ith query in the test set.

3. Experiments

3.1. Details of the Datasets

In order to verify our methods, we conducted experiments on two datasets: Market1501 [28] and CUHK03 [50]. For the latter, we specifically used a version with a clear query-test separation [51]. We conducted experiments on both labeled (L) and detected (D) versions of CUHK03. Table 3 provides an overview of the datasets.

Table 3. The details of the used datasets.

Dataset	Identities	Images	Test	Queries	Clear Queries	Cameras
Market1501	1501	32,668	19,732	3368	3062	6
CUHK03 (L)	1360	13,164	5328	1400	1310	2
CUHK03 (D)	1360	12,697	5332	1400	1294	2

Whereas Table 3 shows that Market1501 is significantly larger than CUHK03, the latter provides a considerably bigger challenge due to poor illumination conditions, which results in different persons looking similar. In addition, CUHK03 is considered to be one of the most challenging re-ID datasets by certain researchers [35]. This would explain the lower performance of our and other methods on CUHK03 when compared to Market1501.

We also inspected SCHP-generated parsing masks of all the queries and marked samples where the parser committed considerable mistakes. Figure 4 illustrates typical parsing errors, where a parsing mask is obviously incorrect, or a significant number of subregions are mislabeled. For certain experiments, which are summarized in Tables 4 and 5, we excluded query images that included parsing errors, while labeling the remaining queries as clear queries (cq). We provide the results for both full and clear queries in order to show the effect of parsing errors on the overall accuracy, as well as to better investigate the accuracy of the proposed feature extraction and similarity ranking approaches. It should be noted that parsing errors have not been removed from the test subsets.

Figure 4. Parsing errors: (**a**) incorrect parsing and labeling of subregions, (**b**) overall correct parsing but most subregions are mislabeled as "jumpsuit".

3.2. Market1501 Experiments

Table 4 shows a comparison regarding the accuracy between our method and other models that also operated on the Market1501 dataset. The comparison shows that we achieve results comparable with unsupervised learning models and DL parsing models in rank-1 and rank-10 categories. However, the mAP accuracy is significantly lower in comparison with other approaches. Figure 5 shows an example of a typical challenge

wherein two different people who are dressed similar have almost the same similarity scores, such their images are mixed obstructing the correct person retrieval. This behavior is understandable, due to the fact that we only use color and texture as features with clothes that provide the greatest contribution. It should be noted that we still outperform modern unsupervised models in terms of rank-1/rank-10 metrics.

Table 4. Performance comparison on the Market1501 dataset.

Model	Backbone	Human Parsing	Learning Type	Rank-1	Rank-10	mAP
SML [52] '19	ResNet-50	No	US *	67.7	-	40
SIV [53] '17	ResNet-50	No	S	79.51	-	59.87
MSCAN [54] '17	Custom	No	S	80.31	-	57.53
CAP [55] '21	ResNet-50	No	US	91.4	97.7	79.2
SSP [34] '18	ResNet-50	Yes	S	92.5	-	80
SPReID [25] '18	Inception	Yes	S	94.63	98.4	90.96
Pyramid [56] '19	ResNet	No	S	95.7	99	88.2
APNet-C [19] '21	ResNet-50	No	S	96.2	-	90.5
CTL-S [57] '21	ResNet-50	No	S	98	99.5	98.3
Ours	ResNet-101 (parser)	Yes	A	91	96	25.2
Ours (cq)	ResNet-101 (parser)	Yes	A	93.5	98.0	25.3

* S—supervised, US—unsupervised, and A—analytical, no learning.

Figure 5. Similarly dressed people (**right**), in a test set of the Market1501 dataset, that have similar scores for a query (**left**), thereby resulting in a low mAP.

3.3. CUHK03 Experiments

Table 5 shows the trends that similar to Table 4. However, the rank-1 accuracy achieved on CUHK03 is lower than that of current cutting-edge models by a bigger margin when compared to Market1501 experiments. This can be explained by poor lighting, thus resulting in color becoming a less reliable criterion. The effects shown in Figure 5 are also common for the CUHK03 dataset. However, we still outperform several older DL models. It should be stressed that the results in Table 5 are obtained by using exactly the same model as the results in Table 4. In regard to this, the importance of this fact is discussed further in Section 4.4.

Table 5. Performance comparison on CUHK03 dataset.

Model	Backbone	Human Parsing	Learning Type	Rank-1 (L)	mAP (L)	Rank-10 (D)	mAP (D)
HA-CNN [58] '18	Inception	No	WS *	44.4	41	41.7	38.6
DaRe [59] '18	DenseNet-201	No	S	56.4	52.2	54.3	50.1
DaRe [59] '18	DenseNet-201	No	S + RR	73.8	74.7	70.6	71.6
SSP [34] '18	ResNet-50	Yes	S	65.6	63.1	66.8	60.5
OSNet [60] '19	OSNet	No	S	-	-	72.3	67.8
Top-DB-Net [35] '20	ResNet-50	Yes	S	79.4	75.4	77.3	73.2
Top-DB-Net [35] '20	ResNet-50	Yes	S + RR	88.5	86.7	86.9	85.7
MPN [61] '21	ResNet-50	No	S	85	81.1	83.4	79.1
Deep Miner [62] '21	RedNet-50	No	S	86.6	84.7	83.5	81.4
LightMBN [63] '21	OSNet	No	S	87.2	85.1	84.9	82.4
Ours	Resnet-101 (parser)	Yes	A	61.1	20.9	59.7	20.2
Ours (cq)	Resnet-101 (parser)	Yes	A	63.9	22.1	62.2	21.4

* S—supervised, WS—Weakly Supervised, RR—re-ranking, and A—analytical, no learning.

4. Discussions

4.1. Performance and Space Requirements

In this paper, we propose a system that features extraction and image comparison modules that are fully analytical without trainable parameters. Such a system has little hardware requirements and does not require a GPU. Feature vectors are compact and require, on average, 3 KB of storage space per image. Due to simplicity of the operations (7)–(12), the analysis takes milliseconds even on an average-grade machine and supports multi-threading on a multi-core CPU. Specifically, feature extraction takes about 30 ms per image and the similarity value calculation for a query-test image pair takes 7 ms on an Intel i9-9900K 3.60GHz CPU. The most time-consuming operation is feature extraction, as it requires one "walk" over the input matrix (query image). However, it is only performed once for every image as feature vectors are stored for future comparison. The accuracy results shown in Section 3, are comparable with certain cutting-edge DL models, while being less hardware-demanding. As such, this fact indicates the viability of the proposed approach. Moreover, we also discuss in Section 4.4, the much higher generalization potential when the results in Tables 4 and 5 are analyzed together.

One might doubt the above claim as our approach still requires a parser. In this paper, we use an SCHP with an ResNet101 backbone (43 million parameters) [36,37], which is larger than the backbones of most models, as is detailed in Tables 4 and 5. However, using such a large parser is not a necessary requirement. A more compact network could be used instead with little loss of accuracy as discussed in [64]. It was shown that rank-1 and mAP decrease by up to 2.5% and 3.7%, when the backbone is changed to MobileNetV2 or OSNet, respectively. At the same time, this leads from ten- to twenty-fold reduction in the number of model's parameters. Therefore, this method can be implemented with a compact parser for high-speed low-demand computations, which is especially promising for the purposes of edge computing applications. Still, this paper presents the results that can be achieved with an "out-of-the-box" parser [37].

4.2. Adding New Features

A major contribution of this paper is the feature extraction and its similarity score calculation scheme, which is based on class similarity considerations. In this paper, we propose two types of features, i.e., color and texture features, as discussed in Section 2. Whereas this is sufficient to achieve high rank-1 accuracy on studied datasets, a low mAP accuracy indicates that there is space for further improvements. This can be achieved by adding new features on feature vector generation steps, as shown in Figure 1 in order to address the problem illustrated in Figure 5.

There are two rules that a new feature should comply with: (a) its associated similarity measure S_f should be in [0, 1] range and (b) its weight w_f should be added in Table 1 and all weights should be adjusted to comply with (10). Class similarity scores are still calculated using (9). Hence, new features can be added with minimal interference with the rest of the model. Future work will need to consider the pattern and shape features in order to improve accuracy.

4.3. Human-Readable Vectors and Human-Generated Queries

In Section 2, we have discussed the fact that all similarity measures are in [0, 1] range and that they can be interpreted by a human operator as a percentage similarity according to a specific feature. Whereas this interpretability is useful for analysis purposes, it can also be used to construct feature vectors while not having an actual image. The possibility to conduct search for an object without query images was investigated in the early days of re-ID [38]. In addition, it can be further improved by using our proposed approach, which includes human parsing.

Constructing feature vectors can be performed as follows: let us assume that colors of the clothes of a person are known. The RGB values of the colors can be converted into Lab in order to instantly obtain triplets (I_L, I_a, I_b), as discussed discussed in Section 2.3.

By using same triplets, binarized histograms can be generated by assigning several "ones" in a region surrounding the main color's intensity. Moreover, the LBP texture features are less intuitive, such that a look-up table of vectors in respect of typical textures can be used instead.

Now, let us consider a situation in which a query is a vague verbal description. Comparing such a query with a database would be impossible for methods that rely on convolutional operations, i.e., all DL models, due to the fact that the image is missing. Generating vectors corresponding to a given verbal query is also impossible, as the feature vectors generated by DL models have no clear interpretation. This would require one to generate an artificial image, which is a very complex task and would require making assumptions of numerous unknown parameters [65].

On the contrary, it is possible to search for matches regarding such a query using our proposed method. Figure 6 illustrates an example of a search for a dark hair person wearing a red shirt, black pants, and black shoes. For the purposes of this search, color descriptions are converted into vectors; in addition, texture features are generated using values typical for clothes in the CUHK03 dataset. Figure 7 illustrates another set of search results for a person wearing a white shirt. Hence, an operator needs to describe only relevant classes without the need to specify other class properties, e.g., the color of the shoes, if it is unknown.

Figure 6. Search results for a human-generated query of a person wearing a red shirt, black pants, and black shoes in the CUHK03 dataset.

Figure 7. Search results for a request to find ten people wearing white shirts.

4.4. Generalization and Potential Application to Open-World Scenarios

Tables 4 and 5 illustrate that the highest accuracy is achieved by models that are trained using supervised learning, especially if re-ranking is applied (see Table 5). This, encourages models to provide higher scores to true matches, thus increasing mAP accuracy. This observation makes it tempting to render the parameters in Tables 1 and 2 to be trainable in order to possibly improve the accuracy in our experiments. However, this would render the generalization worse due to the bias the model would develop towards training datasets. Indeed, such effects are well-known in machine learning.

On the contrary, the absence of trainable parameters renders our approach dataset-agnostic. Our results regarding the Market1501 dataset in Table 4, are obtained using exactly the same model as in case of the CUHK03 results in Table 5. Therefore, this can be considered analogous to transfer learning experiments in context of neural networks.

As such, our 62.9% Market1501–CUHK03 rank-1 "transfer" accuracy is significant when compared to the 50.1% DukeMTMC–Market1501 transfer in [19], as shown in Table 6. Furthermore, Table 6 shows that 20–25% mAP accuracy is common for generalization experiments. In addition, in certain cases, the proposed method results in a transfer-learning mAP that is higher than that of APNet-C. Having said that, it should be noted that the DukeMTMC dataset has since been retracted, which prevents us from conducting exactly the same experiments as in [19]. However, the DukeMTMC dataset can be considered less challenging than the CUHK03 [61,62] dataset.

A number of recent works have suggested domain adaptive clustering techniques as another approach for the purposes of generalization improvement [66,67]. However, such methods still require DL-model training on re-ID datasets and a fine-tuning of clusters on target datasets, which indicates dataset-dependence. On the contrary, the proposed method in this paper is that it is re-ID dataset-agnostic. Furthermore, it is very promising from the perspective of the development of highly generalizable models. This also suggests that when accuracy improvements are achieved on one dataset it is likely to improve the accuracy on other datasets too. Such behavior, although desirable, is not guaranteed for DL models. In contrast with the current trends towards end-to-end DL models, this paper indicates a significant potential in combining machine learning (human parsing) with human intelligence (analytical features) in order to obtain more flexible systems that are more robust in open-world scenarios, while also being easier for human interpretation and understanding.

Table 6. Comparison between single-domain and cross-domain accuracies of supervised DL CNN models with the proposed approach.

Dataset	Market		DukeMTMC		CUHK		Market→X		X→Market	
Metric	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
APNet-C [19]	**96.2**	**90.5**	90.4	81.5	**87.4**	**85.3**	37.7	**22.8**	50.9	23.7
Ours (cq)	93.5	25.3	-	-	63.9	22.1	**63.9**	22.1	**93.5**	**25.3**

X is the DukeMTMC dataset for APNet-C and the CUHK03 dataset in our experiments. Best results are in bold.

5. Conclusions

This paper proposes a re-ID system that combines analytical feature extraction and similarity ranking schemes with human parsing. It shows that parsing masks provide sufficient information to overcome the known limitations of analytical re-ID methods. Two types of features, namely Lab color and LBP texture features, are utilized. In addition, several original pre-processing techniques are proposed. These features are combined with a class similarity evaluation scheme that is used to obtain a similarity ranking with results comparable with conventional rank-r evaluation metrics. The obtained results show a rank-1 accuracy that is comparable with supervised DL models and exceed that of unsupervised ones that were trained on the Market1501 and CUHK03 datasets. The reasons for reduced mAP accuracy are discussed, and potential solutions in adding new feature channels is proposed. It is shown that by having no trainable parameters, the proposed model has significant generalization potential, as illustrated by 93.5% and 63.9% rank-1 "transfer" accuracies achieved by a completely re-ID dataset-agnostic model. Our future work will include studying of additional features in order to improve accuracy, along with parser size minimization in order to realize the proposed system as an edge computing application.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available on request from corresponding author.

Acknowledgments: The author would like to thank Anton Raskovalov and Igor Netay for their fruitful discussions, as well as Vasily Dolmatov for his assistance in problem formulation, choice of methodology, and supervision.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person Re-identification: Past, Present and Future. *arXiv* **2016**. [CrossRef]
2. Iguernaissi, R.; Merad, D.; Aziz, K.; Drap, P. People Tracking in Multi-Camera Systems: A Review. *Multimed. Tools Appl.* **2019**, *78*, 10773–10793. [CrossRef]
3. Kodirov, E.; Xiang, T.; Fu, Z.; Gong, S. Person Re-Identification by Unsupervised l1 Graph Learning. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 178–195.
4. Chen, D.; Xu, D.; Li, H.; Sebe, N.; Wang, X. Group Consistent Similarity Learning via Deep CRF for Person Re-identification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8649–8658. [CrossRef]
5. Wu, Y.; Bourahla, O.E.F.; Li, X.; Wu, F.; Tian, Q.; Zhou, X. Adaptive Graph Representation Learning for Video Person Re-Identification. *IEEE Trans. Image Process.* **2020**, *29*, 8821–8830. [CrossRef] [PubMed]
6. Ye, M.; Ma, A.; Zheng, L.; Li, J.; YUEN, P. Dynamic Label Graph Matching for Unsupervised Video Re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 5152–5160. [CrossRef]
7. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2872–2893. [CrossRef] [PubMed]
8. Lavi, B.; Serj, M.F.; Ullah, I. Survey on Deep Learning Techniques for Person Re-Identification Task. *arXiv* **2018**. [CrossRef]
9. Chicco, D., Siamese Neural Networks: An Overview. In *Artificial Neural Networks*; Springer: New York, NY, USA, 2021; pp. 73–94. [CrossRef]
10. Wu, L.; Shen, C.; Hengel, A.v.d. PersonNet: Person Re-identification with Deep Convolutional Neural Networks. *arXiv* **2016**. [CrossRef]
11. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. 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 (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1487–1495. [CrossRef]
12. Zhu, Z.; Jiang, X.; Zheng, F.; Guo, X.; Huang, F.; Sun, X.; Zheng, W. Viewpoint-Aware Loss with Angular Regularization for Person Re-Identification. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13114–13121. [CrossRef]
13. Schumann, A.; Stiefelhagen, R. Person Re-identification by Deep Learning Attribute-Complementary Information. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1435–1443. [CrossRef]
14. Shen, Y.; Li, H.; Yi, S.; Chen, D.; Wang, X. Person Re-identification with Deep Similarity-Guided Graph Neural Network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
15. Lan, X.; Zhu, X.; Gong, S. Universal Person Re-Identification. *arXiv* **2019**. [CrossRef]
16. Zeng, Z.; Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.Y.; Satoh, S. Illumination-Adaptive Person Re-Identification. *IEEE Trans. Multimed.* **2020**, *22*, 3064–3074. [CrossRef]
17. Xiong, F.; Gou, M.; Camps, O.; Sznaier, M. Person Re-Identification Using Kernel-Based Metric Learning Methods. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; pp. 1–16. [CrossRef]
18. Zheng, W.S.; Gong, S.; Xiang, T. Towards Open-World Person Re-Identification by One-Shot Group-Based Verification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 591–606. [CrossRef]
19. Chen, G.; Gu, T.; Lu, J.; Bao, J.A.; Zhou, J. Person Re-Identification via Attention Pyramid. *IEEE Trans. Image Process.* **2021**, *30*, 7663–7676. [CrossRef] [PubMed]
20. Khan, F.M.; Bremond, F. Person Re-identification for Real-world Surveillance Systems. *arXiv* **2016**. [CrossRef]
21. Gray, D.; Tao, H. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In Proceedings of the Computer Vision—ECCV, Marseille, France, 12–18 October 2008; pp. 262–275.
22. Gheissari, N.; Sebastian, T.; Hartley, R. Person Reidentification Using Spatiotemporal Appearance. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1528–1535. [CrossRef]
23. Nanni, L.; Munaro, M.; Ghidoni, S.; Menegatti, E.; Brahnam, S. Ensemble of different approaches for a reliable person re-identification system. *Appl. Comput. Inform.* **2016**, *12*, 142–153. [CrossRef]
24. Zheng, W.S.; Gong, S.; Xiang, T. Person Re-Identification by Probabilistic Relative Distance Comparison. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 20–25 June 2011; pp. 649–656. [CrossRef]
25. Kalayeh, M.M.; Basaran, E.; Gokmen, M.; Kamasak, M.E.; Shah, M. Human Semantic Parsing for Person Re-identification. *arXiv* **2018**. [CrossRef]
26. Park, H.; Ham, B. Relation Network for Person Re-identification. *arXiv* **2019**. [CrossRef].

27. Quan, R.; Dong, X.; Wu, Y.; Zhu, L.; Yang, Y. Auto-ReID: Searching for a Part-Aware ConvNet for Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
28. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-Identification: A Benchmark. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015. [CrossRef]
29. Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; Huang, T. Horizontal Pyramid Matching for Person Re-Identification. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019. [CrossRef] [PubMed]
30. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef].
31. Gong, K.; Liang, X.; Zhang, D.; Shen, X.; Lin, L. Look Into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
32. Zhao, J.; Li, J.; Cheng, Y.; Sim, T.; Yan, S.; Feng, J. Understanding Humans in Crowded Scenes: Deep Nested Adversarial Learning and A New Benchmark for Multi-Human Parsing. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October2018; pp. 792–800. [CrossRef]
33. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-Driven Deep Convolutional Model for Person Re-identification. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3980–3989. [CrossRef]
34. Quispe, R.; Pedrini, H. Improved person re-identification based on saliency and semantic parsing with deep neural network models. *Image Vis. Comput.* **2019**, *92*, 103809. [CrossRef]
35. Quispe, R.; Pedrini, H. Top-DB-Net: Top DropBlock for Activation Enhancement in Person Re-Identification. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2980–2987. [CrossRef]
36. Li, P.; Xu, Y.; Wei, Y.; Yang, Y. Self-Correction for Human Parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3260–3271. [CrossRef]
37. Li, P.; Xu, Y.; Wei, Y.; Yang, Y. Self Correction for Human Parsing. Available online: https://github.com/GoGoDuck912/Self-Correction-Human-Parsing (accessed on 11 January 2023).
38. Park, U.; Jain, A.; Kitahara, I.; Kogure, K.; Hagita, N. ViSE: Visual Search Engine Using Multiple Networked Cameras. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 1204–1207. [CrossRef]
39. Günther Wyszecki, W.S.S. *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2000. [CrossRef]
40. Rubner, Y.; Tomasi, C.; Guibas, L.J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 1–20. [CrossRef]
41. Chavdarova, T.; Baqué, P.; Bouquet, S.; Maksai, A.; Jose, C.; Bagautdinov, T.; Lettry, L.; Fua, P.; Van Gool, L.; Fleuret, F. WILDTRACK: A Multi-camera HD Dataset for Dense Unscripted Pedestrian Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5030–5039. [CrossRef]
42. Cha, S.H.; Srihari, S.N. On measuring the distance between histograms. *Pattern Recognit.* **2002**, *35*, 1355–1370. [CrossRef] [PubMed]
43. Fatih Demirci, M.; Shokoufandeh, A.; Dickinson, S.J. Skeletal Shape Abstraction from Examples. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 944–952. [CrossRef]
44. Vizilter, Y.; Pyt'ev, Y.; Chulichkov, A.; Mestetskiy, L.M., Morphological Image Analysis for Computer Vision Applications. In *Computer Vision in Control Systems-1: Mathematical Theory*; Springer International Publishing: Cham, Switzerland, 2015; pp. 9–58. [CrossRef]
45. Shu, X.; Wu, X.J. A novel contour descriptor for 2D shape matching and its application to image retrieval. *Image Vis. Comput.* **2011**, *29*, 286–294. [CrossRef]
46. Thewsuwan, S.; Horio, K. Texture-Based Features for Clothing Classification via Graph-Based Representation. *J. Signal Process.* **2018**, *22*, 299–305. [CrossRef] [PubMed]
47. Ahonen, T.; Hadid, A.; Pietikainen, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef]
48. Barkan, O.; Weill, J.; Wolf, L.; Aronowitz, H. Fast High Dimensional Vector Multiplication Face Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1960–1967. [CrossRef]
49. Shekar, B.; Pilar, B. Shape Representation and Classification through Pattern Spectrum and Local Binary Pattern—A Decision Level Fusion Approach. In Proceedings of the Fifth International Conference on Signal and Image Processing, Bangalore, India, 8–10 January 2014; pp. 218–224. [CrossRef]
50. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159. [CrossRef]
51. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking Person Re-identification with k-Reciprocal Encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3652–3661. [CrossRef]

52. Yu, H.X.; Zheng, W.S.; Wu, A.; Guo, X.; Gong, S.; Lai, J.H. Unsupervised Person Re-Identification by Soft Multilabel Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
53. Zheng, Z.; Zheng, L.; Yang, Y. A Discriminatively Learned CNN Embedding for Person Reidentification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 1–20. [CrossRef]
54. Li, D.; Chen, X.; Zhang, Z.; Huang, K. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7398–7407. [CrossRef]
55. Wang, M.; Lai, B.; Huang, J.; Gong, X.; Hua, X.S. Camera-aware Proxies for Unsupervised Person Re-Identification. *arXiv* **2020**. [CrossRef]
56. Zheng, F.; Deng, C.; Sun, X.; Jiang, X.; Guo, X.; Yu, Z.; Huang, F.; Ji, R. Pyramidal Person Re-IDentification via Multi-Loss Dynamic Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8506–8514. [CrossRef]
57. Wieczorek, M.; Rychalska, B.; Dabrowski, J. On the Unreasonable Effectiveness of Centroids in Image Retrieval. *arXiv* **2021**. [CrossRef]
58. Li, W.; Zhu, X.; Gong, S. Harmonious Attention Network for Person Re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2285–2294. [CrossRef]
59. Wang, Y.; Wang, L.; You, Y.; Zou, X.; Chen, V.; Li, S.; Huang, G.; Hariharan, B.; Weinberger, K.Q. Resource Aware Person Re-identification across Multiple Resolutions. *arXiv* **2018**. [CrossRef]
60. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-Scale Feature Learning for Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3701–3711. [CrossRef]
61. Ding, C.; Wang, K.; Wang, P.; Tao, D. Multi-Task Learning with Coarse Priors for Robust Part-Aware Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1474–1488. [CrossRef]
62. Benzine, A.; Seddik, M.E.A.; Desmarais, J. Deep Miner: A Deep and Multi-branch Network which Mines Rich and Diverse Features for Person Re-identification. *arXiv* **2021**. [CrossRef]
63. Herzog, F.; Ji, X.; Teepe, T.; Hörmann, S.; Gilg, J.; Rigoll, G. Lightweight Multi-Branch Network For Person Re-Identification. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, Alaska, USA, 19–22 September 2021; pp. 1129–1133. [CrossRef]
64. Gabdullin, N.; Raskovalov, A. Google Coral-based edge computing person reidentification using human parsing combined with analytical method. *arXiv* **2022**. [CrossRef]
65. Jiang, Y.; Yang, S.; Qju, H.; Wu, W.; Loy, C.C.; Liu, Z. Text2Human: Text-Driven Controllable Human Image Generation. *ACM Trans. Graph.* **2022**, *41*, 1–11. [CrossRef]
66. Xie, H.; Luo, H.; Gu, J.; Jiang, W. Unsupervised Domain Adaptive Person Re-Identification via Intermediate Domains. *Appl. Sci.* **2022**, *12*, 6990. [CrossRef]
67. Zheng, K.; Lan, C.; Zeng, W.; Zhang, Z.; Zha, Z.J. Exploiting Sample Uncertainty for Domain Adaptive Person Re-Identification. *arXiv* **2020**. [CrossRef]