Doppelgänger Saliency: Towards More Ethical Person Re-Identification

Brandon RichardWebster*, Brian Hu*, Keith Fieldhouse, Anthony Hoogs
Kitware, Inc.
1712 Route 9, Suite 300, Clifton Park, NY 12065
{brandon.richardwebster, brian.hu, keith.fieldhouse, anthony.hoogs}@kitware.com

Abstract

Modern surveillance systems have become increasingly dependent on artificial intelligence to provide actionable information for real-time decision making. A critical question relates to how these systems handle difficult ethical dilemmas, such as the re-identification of similar looking individuals. Potential misidentification of individuals can have severe negative consequences, as evidenced by recent headlines of individuals who were wrongly targeted for crimes they did not commit based on false matches. A computer vision-based saliency algorithm is proposed to help identify pixel-level differences in pairs of images containing visually similar individuals, which we term "doppelgängers." The computed saliency maps can alert human users of the presence of doppelgängers and provide important visual evidence to reduce the potential of false matches in these high-stakes situations. We show both qualitative and quantitative saliency results on doppelgängers found in a video-based person re-identification dataset (MARS) using three different state-of-the-art models. Our results suggest that this novel use of visual saliency can improve overall outcomes by helping human users in the person re-identification setting, while assuring the ethical and trusted operation of surveillance systems.

1. Introduction

Despite the widespread use of surveillance technologies, recent headlines have begun to highlight some of their potential harms. One prominent example is Robert Williams, who was wrongly arrested for a crime he did not commit based on facial recognition software that incorrectly matched him to a different individual [24]. This was one of the first known cases where technology misidentified an individual, with real negative consequences for Williams and his family. He is now suing the Detroit police department for damages, while advocating for greater transparency about the use of such technologies in policing efforts. Unfortunately, additional cases of wrongful arrest and imprisonment due to false facial recognition matches have also been reported [23].

Figure 1. Example doppelgänger saliency. Image regions that differ between the two individuals (e.g. face, shirt logo, pants, and shoes) are highlighted in green. For illustration purposes, colored arrows pointing to corresponding image regions are shown. Note that a region does not have to be highlighted in both images to be considered a difference. In a full person re-identification system, the user can view the highlighted regions to quickly spot visual differences in the doppelgänger pair. Figures best viewed in color.

Modern surveillance systems increasingly rely on artificial intelligence (AI), specifically computer vision, for their automated reasoning capabilities. Given access to large amounts of training data, machine learning algorithms can learn to accurately (although not perfectly) re-identify individuals based on either face or whole-body images. This has resulted in the exponential growth of the field of biometrics, spurring new research in areas such as face recognition [4, 5, 31, 39] and person re-identification (ReID) [22,30,36,53]. Typically lagging behind this work is research addressing the ethical concerns of such technologies, and how to best ensure their appropriate and trusted operation. More recently, several works have tried to tackle ethical issues of using these types of technologies [7,41,42].

Ethical AI studies ethical considerations in the design and use of AI systems [15,19,47,51]. Closely related to the

*denotes equal contribution

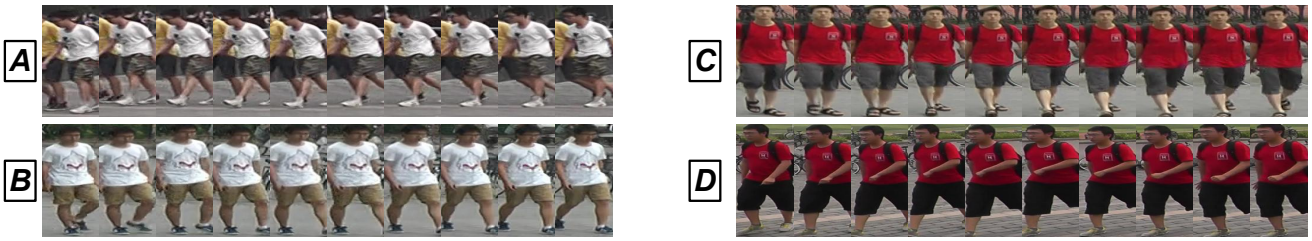

Figure 2. Four example tracklets from the MARS dataset [59] illustrate naturally occurring doppelgängers. Tracklets A and B are fairly similar: both individuals are wearing white shirts, shorts, and glasses, and both have dark hair. However, they have shirt logos that differ in color/style, and one pair of shoes is white while the other pair is black. Tracklets C and D represent a pair of individuals that are almost indistinguishable: same shirt, same shorts (short of lighting differences), both have backpack straps, and both have dark hair. There is only one strong visual difference at this resolution — the shoes are different in type (sandals vs. athletic shoes) and color (black vs. tan).

field of ethical AI is explainable AI (XAI). Explainable AI is a set of tools and resources that seeks to provide human-interpretable explanations of AI models [17,45]. In the specific case of surveillance applications, human-machine interaction is often critical– the machine generates a set of potential match results and a human user must adjudicate these results. However, the person ReID algorithms used are typically "black boxes," which provide users little insight into how they arrived at their final output decisions. In high-stakes situations such as these, there is a need for XAI tools that can help address the interpretability, traceability, reliability, and governability of these systems [6]. This will provide users and model developers the ability to diagnose model failure modes or inherent biases in the underlying data used to train such models, while also potentially improving model operation and ensuring the trusted and responsible use of models.

In this paper, we examine the use of saliency maps, a form of visual XAI, to help users distinguish between visually similar individuals which we term "doppelgängers" (Fig. 1). These saliency maps provide important information to users when adjudicating potential matches identified by the system, highlighting salient visual differences in pairs of images that can support ReID decisions. This additional information can help to reduce the uncertainty associated with high-stakes ReID and improve the overall operation of the system. We make the following contributions in our work:

- We develop a public benchmark of pairs of naturally occurring, visually similar individuals (*i.e.*, doppelgängers), and use this to assess different models and their associated explainability.
- We apply a novel form of classification-based saliency to highlight pixel-level differences in pairs of images containing visually similar individuals, which can help human users avoid false matches in high-stakes ReID.
- We quantitatively evaluate the quality of the generated saliency maps using an automated set of causal metrics that measures the impact of the identified salient regions on downstream model ReID of individuals.

2. Related Work

Person Re-Identification (ReID). Person ReID seeks to accurately identify individuals over time and across various environmental changes such as different camera views, indoor/outdoor settings, etc. Research in this area has increasingly made use of deep learning algorithms, which learn to classify individuals based on large labeled training datasets. Examples of commonly used person ReID datasets include CUHK-03 [32], Market-1501 [60], MARS [59], and MSMT-17 [55]. These datasets typically contain on the order of hundreds to thousands of unique identities, and range from cropped single-image views of different individuals to video-based tracklets of individuals across multiple frames.

Person ReID is a challenging task due to many factors, including occlusion, misaligned frames, and the presence of similar-appearance identities. Example tracklets from the MARS dataset [59] containing visually similar individuals are shown in Fig. 2. Traditional person ReID approaches operate on single images and largely make use of convolutional neural networks [22, 30, 36, 53]. More recently, person ReID models that make use of video-based information have also been proposed, with the ability to better model the appearance and background surroundings of individuals over time [2,14,21,54]. Although the focus of our current work is on appearance-based modeling, other work has explored using additional forms of information such as gait [40,49] or soft biometric fusion [18,35,52].

More recently, there has also been research on incorporating forms of attention or saliency as part of the underlying person ReID algorithm. This includes work that uses forms of attention that learn to highlight relevant image regions for matching either during training or at test time [9, 33, 34, 58]. These attention mechanisms help the model focus on relevant image features for each identity, while removing potential confounds such as the background, occlusion, or multiple individuals in a given frame that often naturally occur in datasets. These approaches can

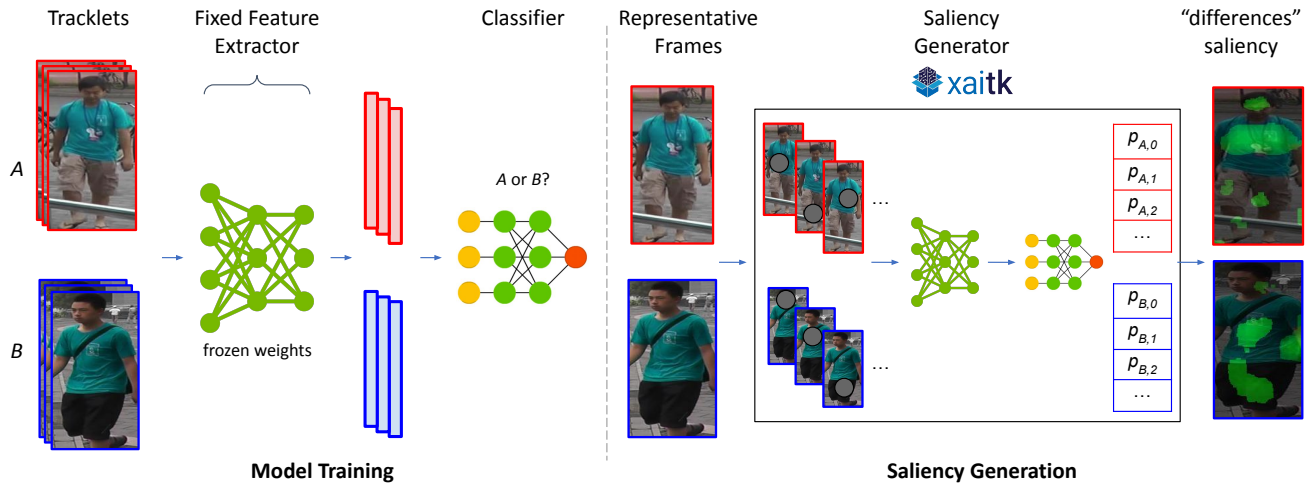

Figure 3. Model training and saliency generation. (Left) Tracklets A and B are used as input to a fixed feature extractor to generate chip-level features. A binary classifier is trained on these features to predict the probability of a chip belonging to tracklet A or B. (Right) A representative chip from each tracklet is fed into the saliency generator provided by the Explainable AI Toolkit (XAITK) [26], which outputs the final "differences" saliency. At a high-level, the saliency generator masks out regions of the input image and inputs those masked images into the trained model and classifier. The computed output probabilities are then weighted-averaged to identify salient regions in each chip that represent the strongest visual differences between the two identities (shown overlaid as green regions).

often provide a basic form of interpretability that is intrinsic to the model, although other forms of post-hoc explainability also exist for evaluating models that have already been trained. Examples of these methods will be described in more detail in Section 2 below.

Ethical and explainable AI. Towards this end, there is also work on ethical and explainable AI (XAI). More broadly, ethical AI focuses on the ethical development and use of AI technologies. Ethical AI touches upon the areas of research, social concern, and public policy, making it rather unique in the field of AI. Some of the most infamous cases are gender and race biases in face recognition systems created by major internet companies [7, 41, 42]. Recently, the Department of Defense also released several ethical principles related to the development of AI technology: responsible, equitable, traceable, reliable, and governable [6]. XAI can be considered a subset of ethical AI, and seeks to provide methods to help users better understand and appropriately trust AI. XAI is critical in high-stakes situations such as autonomous driving, criminal justice, and healthcare [12, 25, 44], where the outputs of the model can negatively impact humans and need to be reliable and trustworthy.

Explanations of AI models typically fall into two different categories based on their scope and mechanism. Local explanations provide interpretations of individual exemplars or data points (*e.g.* images), while global explanations seek to explain models at the entire dataset level. Explanations can either be white-box or black-box based on how much access to the underlying model being explained is required. Black-box methods are model agnostic and can be applied more generally since they only require access to the inputs and outputs of a model. In contrast, white-box methods often require the computation of model gradients which require knowledge of the model's internal architecture and parameters. In addition to post-hoc explanation methods, models can also be made more inherently interpretable [12, 44]. Recent techniques include prototypical part networks [8] and concept bottleneck networks [29], where models are made more interpretable via learning of "prototypical" examples or specialized loss functions.

Visual saliency. We focus specifically on the use of visual explanations in the form of saliency maps, which are heatmaps that provide users insight into image regions the model paid attention to when making its output prediction. The majority of XAI techniques involving saliency have been developed for image classification tasks [16, 43, 46, 57]. Zeiler and Fergus [57] proposed a black-box method for computing classification-based saliency maps using occlusion. By sliding a box across the image, they measured changes in classification confidence to indicate salient image regions that contributed most to the model's classification. Related to this idea is work from Ribeiro et al. *et al*. [43], which proposed local interpretable model-agnostic explanations (LIME). The model uses super-pixels precomputed on the input image, which define correlated regions of the input, and measures the influence of removing these super-pixels on the classification model through a surrogate linear model. In contrast to these black-box approaches, methods can also use the internal activations and associated gradients of the model for a given prediction, such as Grad-CAM [46]. These methods reveal coarse input regions which are associated with the output predic-

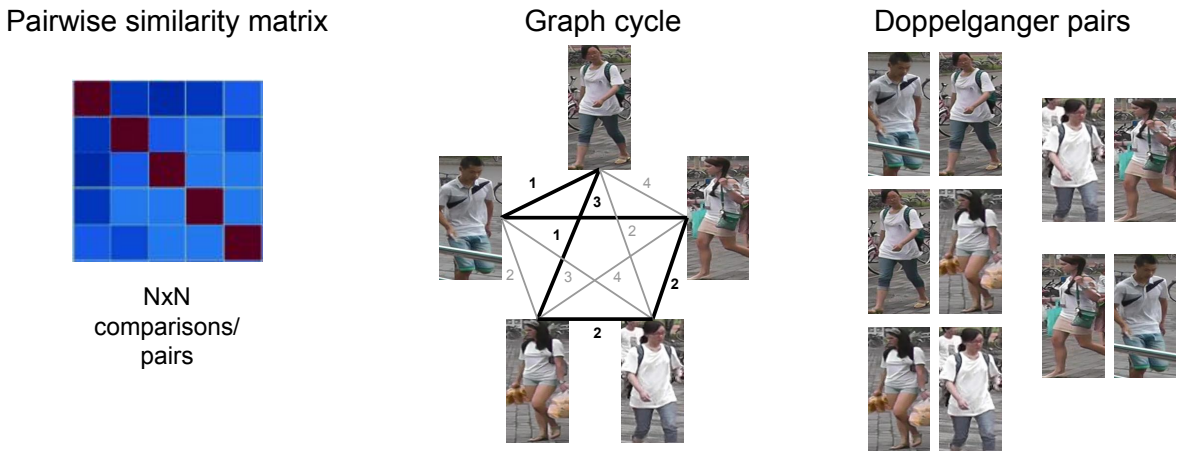

Figure 4. The representative chip for each identity is used in exactly two doppelgänger pairs through a cycle computation. A cycle is generated by treating the NxN feature similarity matrix as a fully-connected graph (N=5 here for illustration purposes). Since the edge weights of the fully-connected graph contain the distance between each possible pair, we can compute a cycle with the minimum possible sum of weights. The selection of doppelgängers is described in Sect. 4. Note, while this figure contains people with mostly white shirts, the full set of doppelgängers contains variance in clothing attire — generally, the cycle groups people with similar colored clothing or clothing style such as shorts, pants, dresses, hats, etc.

tion of the model, and are relatively fast to compute due to their requirement for only a single forward and backward pass through the model. Finally, Fong and Veldadi [16] proposed a masking procedure to find meaningful perturbations of the image that influence model predictions. They did this by casting the problem as an optimization problem, with a process to find masks constrained by size over the input image. More recently, there has also been interest in creating explanations for other image understanding tasks, such as object detection [38] and image similarity [11,13,27,48,56].

3. Methods

Consider a ReID scenario with two separate cropped video clips each containing one individual, with each represented as a sequence of chips called a tracklet. Given a representative chip from each of the two tracklets which best approximates a frontal view of the individual in the tracklet, we compare the two representative chips for visual differences which can be used to differentiate the pair of individuals. Given any model that has been trained for person ReID and that can produce person-specific features, we propose the following method to produce saliency maps which highlight these discriminative image regions in the pair of representative chips.

Tracklet Definition. A tracklet is a sequence of chips that have been cropped to an individual that was tracked across the full-frame video. A tracklet can be variable length and from any camera. In the case of doppelgängers, we assign one tracklet, A, to represent one of the two doppelgängers, and another tracklet, B, to represent the other doppelgänger. Examples of tracklets that represent doppelgänger pairs can be seen in Fig. 2. Notice that doppelgängers can differ in degree of visual similarity, and that not all pairs that are considered doppelgängers by both models human observers.

Model Definition. In the context of person ReID, we consider a model that takes as input two tracklets and predicts the probability of whether or not the identities match. In the doppelgängers case, two individuals are so visually similar that the prediction of whether the two tracklets are of the same person or not is considered somewhat unreliable. Thus, a model for doppelgänger saliency requires more than a simple binary decision, but instead a prediction of which tracklet, A or B, the selected chip belongs to. There are numerous model definitions that could work here, but three criteria must be met: 1) the end-to-end model must predict binary-classification probabilities, 2) the feature generator must be pre-trained, and 3) the binary-classification layer/step must be retrainable for each doppelgängers pair.

Before we generate saliency for two representative chips of tracklet A and tracklet B, we must re-train the classification portion of the model on tracklet A and tracklet B. One reason and motivation for the use of a video-based person ReID dataset is to have access to a sufficient number of chips for each individual to train this classifier (instead of a single chip per individual). The objective is to force the binary classifier to learn the descriminative features of *only* tracklet A and tracklet B (instead of more general features useful for the overall person ReID task) — here a slight over-fitting of the classifier is desired in order to learn pixel-level semantic differences between tracklet A and tracklet B. A pipeline of the entire model and saliency generation procedure can be seen in Fig. 3.

Saliency Definition. Assume there exists a function **F** that fits the previously given model criteria, which predicts the

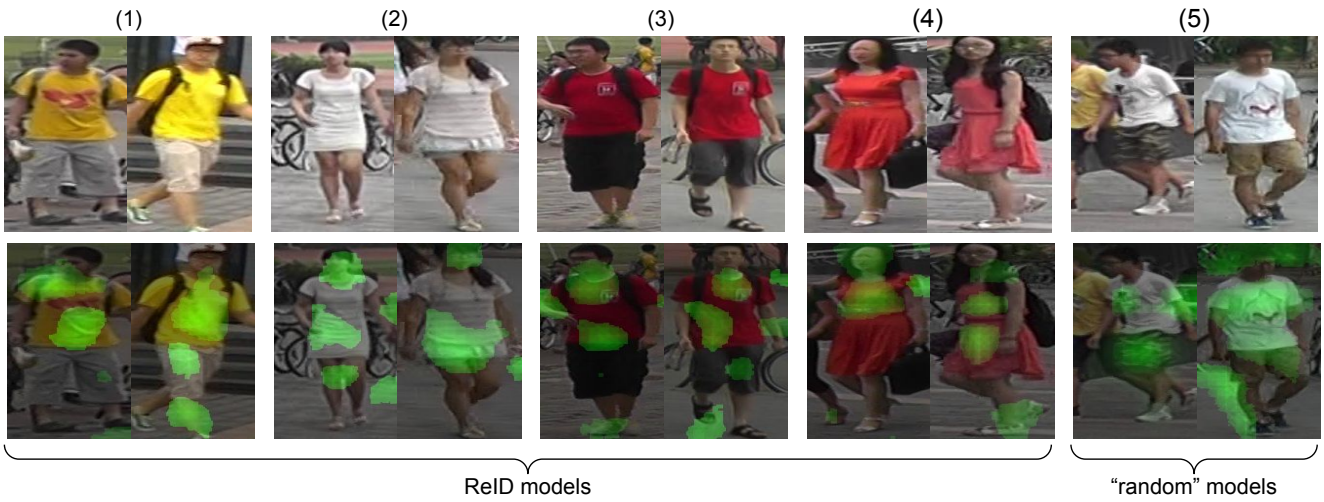

Figure 5. Example saliency results demonstrating the proposed method. The first four pairs of doppelgänger saliency are from a trained ReID model and the last pair is a sanity check with training data randomization (see Sect. 4 for details). The first four pairs highlight at least one key difference, in order from left to right: missing logo, shorts instead of dress, different shoes, and different face. The fifth pair shows more diffuse saliency due to the data randomization, showing that the saliency maps represent the model's training.

probability p_c an input chip x belongs to class c:

$$\mathbf{F}_c(x) = \begin{cases} p_{\mathrm{A}}, & if\ \mathrm{c} = \mathrm{A} \\ p_{\mathrm{B}}, & if\ \mathrm{c} = \mathrm{B} \end{cases} \quad (1)$$

where p_{A} is the probability x belongs to tracklet A and p_{B} the probability that input image x belongs to tracklet B. Because F is a binary classifier, $p_{\mathrm{A}} + p_{\mathrm{B}} \stackrel{!}{=} 1$.

Given an input image x of an individual, $\mathbf{F}_A(x)$ and $\mathbf{F}_B(x)$ are then the predicted probabilities for classes A and B by the model $\mathbf{F}$. We define x' to be a perturbed version of the original image x. This perturbation is usually done with some form of occlusion, using a pixel-wise multiplication of image x with a mask of the same size with values between 0 and 1. These occlusions remove important image features from a particular region of the image (Fig. 3). For example, x' might be the image after removing the face of the individual in the image. In this case, if the individual belonged to class A, the predicted probability for class A, $\mathbf{F}_A(x')$, might decrease as a result of the perturbation.

By repeating these perturbations many times and recording the change in the predicted probabilities for each identity A and B, we can compute a weighted average of the masked regions and obtain a saliency map corresponding to each class (Figure 3). Intuitively, image regions that are highlighted by the saliency map are critical for the model's prediction (i.e. removing them impacts the predicted class probability). In other words, the region of pixels that has the strongest signal is also the region which most strongly discriminates it from its doppelgänger counterpart: if image x initially belongs to tracklet A, the strongest signal / region of pixels is the region that differs most from tracklet B, and vice-versa. This is what we call doppelgänger saliency. In our subsequent results, we compute saliency maps for each representative chip in each doppelgänger pair.

4. Experiments

MARS dataset. The MARS (Motion Analysis and Re-identification Set) dataset is a large-scale, video-based person ReID dataset collected from six near-synchronized cameras [59]. The dataset consists of 1,261 different pedestrians, spanning more than 20,000 tracklets. Each of the videos contains significant variations in pose, color, and illumination, along with the resolution of different pedestrians, each of whom were captured by at least two cameras. Moreover, the dataset also contains 3,248 distractor images for testing the robustness of person ReID algorithms.

Saliency is difficult to apply across a sequence of images, particularly in the case of doppelganger saliency. In addition, if every pair of possible chips were used from MARS, the number of saliency maps that would need to be generated would become intractable. As such, for the purposes of demonstrating the utility of doppelganger saliency, we selected a subset of the MARS dataset to highlight the utility of the method. First, for each identity one track is manually selected to represent that identity — distractors are excluded. The track was selected to be frontalized in order to capture as much detail in a person as reasonably available. Second, since each tracklet requires a representative chip from which the saliency maps will be generated, we manually selected one representative chip for each of those tracklets. In all there were 616 tracklets representing 616 identities with a total of 616 representative chips. Frame selection was done independent of cameras. Lastly, the MARS dataset includes a small number of children in the frames —

all children were excluded in our manual selection process.

Models. We selected three ReID models which vary in model architecture, rank-1 accuracy, and publication date to demonstrate the generalizability of our proposed method. Working from oldest to newest, which is also the ordering by rank-1 accuracy, we have a ResNet-50 [20], a DenseNet-121 [28], and a PCB [50] person ReID model. At this point in time, ResNet-based models are the classic "go to" of the deep-learning models as it is known for being able to train quickly with what is still considered extremely deep neural architecture layers. ResNet introduced shortcut residual connections which reformulate layers as learning residual functions instead of unreferenced functions. Effectively, He *et al*. [20] demonstrated that ResNet was easier to optimize and could gain accuracy from the considerably increased depth while also remaining computationally less complex than its predecessor network structures.

With the innovation of residual connections changing the game in terms of network depth and optimization efficiency, Huang *et al*. [28] introduced a variation on ResNet that provided a substantial improvement in the flow of information through a network. At the cost of the great depth residual networks could achieve, DenseNet structure connected every previous layer to every future layer all the way from the input layer to the final feature layer. This network structure allows for each layer to have access to the the output information of every layer before it. A consequence of this is that the reuse of information allow for more compact models while maintaining higher performance.

Our final model, Part-based Convolutional Baseline (PCB), is a rework of neural architectures designed specifically for persons, where-as the two were originally based on object recognition. PCB can take any network without the fully-connected layers as the backbone, but in the model under consideration ResNet-50 is used as the backbone. The backbone's global pooling layer is removed and replaced with PCB to spatially down-sample the resulting feature activation tensors into column vectors which each get turned into classifiers. However, as is the case for all three of our networks, the final output layers are removed for both the purposes of training them into ReID models as well as generating doppelgänger saliency.

Adapting the ReID trained version of each of these models is simple. Each of the ReID classification layers is removed such that only a feature representation remains as output. The weights are then frozen in place so the model does not change from the original dataset that it was trained upon (i.e. Market-1501 [60]). For every new doppelgänger pair, the associated tracklets (one for each person) are passed onto the model and features are extracted. A binary support vector machine (SVM) classifier is trained to predict the probabilities that each feature came from one of the two tracklets. The SVM is retrained on every pair in order to force the learned SVM to fixate on differences in the tracklets. One might consider that more data is better because it usually allows for better generalization, however in this case too much generalization may prevent the SVM from fixating on key differences. Likewise, too little data can cause the SVM to only fixate on individual pixels. In practice, we use the entire length of the associated tracklets as no "perfect" number has been determined.

Doppelgänger pairs Even with our full set of tracklets and representative chips, not all combinations of these chips will be doppelgangers. Instinctively, one might consider taking the pairs with the minimal Euclidean distance in the person ReID model's feature space, however this runs into an issue where one representative image can be selected in many of the pairs (this is a known issue and described in Doddington's Zoo [10]). Instead, what we want is to ensure that every representative chip is used at least once and select the closest chips by Euclidean distance. In theory, a perfect selection would be to select a cycle in the graph using the minimum some of distances, unfortunately this is also known as the traveling salesman problem and NP-hard.

To get this approximated selection, we use an open-source implementation [3] of Ant Colony Optimization to get a fast approximation of a cycle. With the computed cycle, every representative chip is selected to be in exactly two pairs. Finally, since this process produces a cycle that is specific to each model, we take the average similarity matrix across our three selected models first and use then used the averaged similarity matrix to select the cycle. We call this the global set of doppelgängers. A subset of the global set of doppelgängers can be viewed in Fig. 4, which also describes the cycle computation process.

Saliency Generation. To facilitate the masking of the input doppelgängers pairs and generation of saliency masks for each selected model, we use an off-the-shelf publicly available Python package named *xaitk-saliency*, which is part of the larger Explainable AI Toolkit (XAITK) [26]. The *xaitk-saliency* package — hereby referred to as *xaitk* — implements a class of explainable AI (XAI) algorithms known as saliency algorithms. In our person ReID paradigm, a ReID model + trained SVM operates on a representative input chip to produce a binary classification probability. Saliency algorithms build on this to produce visual explanations in the form of saliency maps as shown in Fig. 5. To ensure the validity of the computed saliency methods, we also performed a data randomization test as proposed in [1], where we randomly shuffled the labels of the data used to train the classifier model (which should result in chance performance) and and re-computed saliency maps. We found that our proposed saliency algorithm passed this sanity check.

Evaluation Protocol. To quantitatively evaluate the efficacy of doppelgängers saliency, we use two automatic evaluation metrics, deletion and insertion as proposed in [37].

Figure 6. An example sequence of insertions (top) and deletions (bottom) showing the steps used in the evaluation metrics as described in Sect. 4. The plot on the right shows the corresponding curves for such insertion (red) and deletion (green), which can be quantified by an area under the curve (AUC). Our reported results are averages across all image pairs.

The deletion metric seeks to remove input pixels that will force the model to decrease its predicted probability of the predicted class. The insertion metric is the opposite to the deletion metric in that it measures the increase in the probability of the predicted class when pixels are slowly introduced to the input image. In both metrics, a sharp change in fraction of pixels will cause either a higher or lower area under the probability curve. A better score is higher AUC for insertion and lower AUC for deletion. The background image used for both metrics is gray, which is set to the mean channel-wise values of the pretrained models.

In the person ReID setting, these metrics measure how useful the image regions identified by the saliency algorithm are for distinguishing between doppelgängers. For example, if the logo on the shirt of one of the individuals is the critical image feature that helps distinguish him or her from a visually similar individual, saliency should be focused on that logo. When masking this logo, the ability of the classification model to accurately distinguish between doppelgängers should decrease, and this should be reflected in better insertion and deletion scores. A complimentary figure showing example insertion and deletion results can be seen in Fig. 6.

To compare both the quantitative and qualitative results with a control sample, we performed a second set of experiments by where we randomly shuffled the training labels for chips in tracklets A and B. This acts as a data randomization test, which effectively changes the SVM optimization and produces saliency maps that are effectively "random" but still plausible. These "random" saliency maps were then used to alter the order in which pixels would be inserted or deleted. This creates a mismatch between the SVMs that were trained with proper class labels and the saliency maps that were generated with random class labels, allowing a sanity check of the proposed saliency method [1].

Qualitative Results. Generally, the visually quality of the doppelgängers saliency for the trained models is in many cases quite good. We see that differences in logos on shirts, hand bags, clothing colors, and shoes get highlighted with a strong signal. However, we often also see substantial amounts of non-person regions being highlighted (*e.g.*, background). In a full person ReID system, this could be mitigated by using a person detector, so this is less consequential than one would initially consider. From a practical point of view within a full ReID system, the proposed "differences" saliency only needs to provide one usable difference in a pair for a user to determine if the pair is same or different. One area of highlighting that we would have liked to see is faces, but in most cases this is ignored. There are two main reasons that this could be: 1) the faces are low resolution as a result of the tracklet capture process described in the the MARS dataset, and 2) the ReID models are fixating on clothing colors, as demonstrated in Fig. 4 which contains a sub-cycle of the full global doppelgänger cycle.

We also visualize one random saliency map from the control experiments for a doppelgänger pair in Fig. 5 (5). This pair can be compared to the trained saliency presented in Fig. 1. Generally, random saliency is more diffuse and more background/non-person context is highlighted compared to the trained saliency. That said, this method of producing "randomness" in the saliency is meant to preserve some locality of regions, and you can see that occurring in the logo of the two individuals in the doppelgänger pair. When training the SVM for this pair, the model still picks up on the logo being different across the chips despite the shuffled class labels. We can see this region highlighted in the random saliency. In contrast, more course-grained features such as the shoe color highlighted in Fig. 1 are lost.

Quantitative Results. To perform a quantitative saliency analysis, we computed the insertion and deletion scores on each of the three models using positive saliency. Positive saliency by definition is a region which increases the confidence the representative chip belongs to the tracklet A or B. We would expect that when inserting pixels with strong signal strength, we would see a sharp rise in the class probabilities, thus also a higher aggregate insertion score. The reverse is also true: pixels with strong signal should have a

Model	Reference	Rank-1	Trained		Random	
			Insertion ↑	Deletion ↓	Insertion ↑	Deletion ↓
PCB	Sun *et al*. (2018) [50]	92.64%	0.532	0.468	0.501	0.499
DenseNet-121	Huang *et al*. (2017) [28]	90.17%	0.547	0.450	0.499	0.501
ResNet-50	He *et al*. (2016) [20]	88.84%	0.538	0.458	0.498	0.497

Table 1. Three selected person ReID models which vary in architecture, rank-1 accuracy, and publication date. Each model was trained and evaluated on the Market-1501 dataset [60] for the reported rank-1 accuracy. Models are ordered by rank-1 score from highest to lowest. Insertion and deletion scores, as described in Sect. 4 and Fig. 6, were computed on the MARS dataset [59]. An arrow next to the insertion and deletion column header indicates the direction of the better score (*i.e*., higher insertion scores and lower deletion scores are better).

sharp decrease in predicted probabilities when removed and thus lower aggregate deletion scores. Positive saliency also has a counterpart negative saliency, which by definition is a region that reduces the confidence for a given class. So if we were to add the negative saliency regions, we do not expect them to increase the insertion score. For this reason, insertion and deletion scores were only generated using positive saliency (mirroring the original RISE paper which proposed these metrics [37]) even though visualizations utilize both positive and negative (i.e., all saliency figures in this article display both positive and negative saliency).

The chosen models showed varying performance that did not always correlate with their insertion and deletion scores, as shown in Table 1. Although the PCB model achieved the top rank-1 performance on Market-1501, its insertion and deletion scores were slightly worse than the DenseNet model. This suggests that the highest performing model may not be the most explainable one, *i.e*., the model that best generates highlighted regions that can effectively be used by someone in a full ReID system to differentiate between doppelgängers. To confirm that our saliency maps represent the underlying models, we can also compute insertion and deletion scores under the control case of data randomization. In the control case, what we would expect to see is both the insertion and deletion scores to degrade relative to the trained model's metric; insertion should decrease, and deletion should increase. Our results show that insertion scores decrease and deletion scores increase for all models under this data randomization control, suggesting the proposed saliency method passes one of the sanity checks and is indeed sensitive to model training. This change in insertion/deletion scores between the control and our baseline is a good quantitative indicator that our trained models are successfully selecting discriminative features to be highlighted in saliency and usable for real-life ReID.

5. Discussion

A limitation of the current work is that we did not validate our approach with actual human users in a real-world deployment scenario due to time and resource constraints. One possibility would be to design an experiment that allows human users to adjudicate pairs of doppelgängers with and without saliency, and measure their overall ReID performance under these two conditions. This would allow us to better quantify the utility of the explanations in the person ReID setting and how this scales with task difficulty. With the addition of human annotations, we could also validate whether or not the regions indicated by the saliency maps are similar to those that humans would use to distinguish between similar identities instead of our current automated evaluation approach, which is only a proxy.

The current saliency algorithms also operate on a per-frame basis, ignoring potential temporal information that is readily available in video data of tracklets of people that could further be used to help disambiguate different individuals. As such, we also only use a single chip from each tracklet of individuals, making it critical to have frontal and unoccluded views of individuals as input to the algorithm. Future work could study how to leverage the rich spatio-temporal information present in tracklets of detected individuals, both in the form of better ReID algorithms as well as corresponding explanations of these algorithms.

Finally, the proposed saliency technique may also be more broadly applicable to the set of computer vision problems known as fine-grained visual recognition. In this area of research, algorithms must learn to make use of the small differences between classes of very similar objects (e.g. different species of birds or visually similar faces). Utilizing our saliency method in these paradigms may also reveal the features used by models to distinguish between visually similar classes, helping to limit any potential bias and improve the quality of models in data-limited regimes. As such, we believe the techniques proposed here may be of broad interest to computer vision communities at large.

6. Acknowledgements

This work is supported by the Defense Advanced Research Projects Activity (DARPA) via contract HR001120C0180. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018. 6, 7

[2] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K. Roy-Chowdhury, and Ziyan Wu. Spatio-temporal representation factorization for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 152–162, October 2021. 2

[3] John Berroa. Ant colony optimization. https://github.com/johnberroa/Ant-Colony-Optimization, 2019. 6

[4] Soma Biswas, Gaurav Aggarwal, Patrick J Flynn, and Kevin W Bowyer. Pose-robust recognition of low-resolution face images. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):3037–3049, 2013. 1

[5] Soma Biswas, Kevin W Bowyer, and Patrick J Flynn. Multidimensional scaling for matching low-resolution face images. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):2019–2030, 2011. 1

[6] Defense Innovation Board. Ai principles: Recommendations on the ethical use of artificial intelligence by the department of defense. *Supporting document, Defense Innovation Board*, 2019. 2, 3

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 1, 3

[8] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *Advances in neural information processing systems*, pages 8930–8941, 2019. 3

[9] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3300–3310, 2020. 2

[10] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. Technical report, National Inst of Standards and Technology Gaithersburg Md, 1998. 6

[11] Bo Dong, Roddy Collins, and Anthony Hoogs. Explainability for content-based image retrieval. In *CVPR Workshops*, pages 95–98, 2019. 4

[12] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 3

[13] Oliver Eberle, Jochen Büttner, Florian Kräutli, Klaus-Robert Müller, Matteo Valleriani, and Grégoire Montavon. Building and interpreting deep similarity models. *arXiv preprint arXiv:2003.05431*, 2020. 4

[14] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12036–12045, 2021. 2

[15] Luciano Floridi. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1(6):261–262, 2019. 1

[16] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3, 4

[17] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2, 2017. 2

[18] Manuel Günther, Andras Rozsa, and Terranee E Boult. Affact: Alignment-free facial attribute classification technique. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 90–99. IEEE, 2017. 2

[19] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, 2020. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6, 8

[21] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021. 2

[22] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 2

[23] Kashmir Hill. Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. *New York Times*, 2020. 1

[24] Kashmir Hill. Wrongfully Accused by an Algorithm. *New York Times*, 2020. 1

[25] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017. 3

[26] Brian Hu, Paul Tunison, Bhavan Vasu, Nitesh Menon, Roddy Collins, and Anthony Hoogs. Xaitk: The explainable ai toolkit. *Applied AI Letters*, 2(4):e40, 2021. 3, 6

[27] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-mir: Explainable medical image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 440–450, 2022. 4

[28] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 6, 8

[29] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *arXiv preprint arXiv:2007.04612*, 2020. 3

[30] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and

latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 384–393, 2017. 1, 2

[31] Pei Li, Loreto Prieto, Domingo Mery, and Patrick J Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019. 1

[32] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 2

[33] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018. 2

[34] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017. 2

[35] Zongyi Liu and Sudeep Sarkar. Outdoor recognition at a distance by fusing gait and face. *Image and Vision Computing*, 25(6):817–832, 2007. 2

[36] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 2

[37] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 6, 8

[38] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. *arXiv preprint arXiv:2006.03204*, 2020. 4

[39] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005. 1

[40] P Jonathon Phillips, Sudeep Sarkar, Isidro Robledo, Patrick Grother, and Kevin Bowyer. The gait identification challenge problem: Data sets and baseline algorithm. In *Object recognition supported by user interaction for service robots*, volume 1, pages 385–388. IEEE, 2002. 2

[41] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019. 1, 3

[42] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020. 1, 3

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 3

[44] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 3

[45] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017. 2

[46] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[47] Ben Shneiderman. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31, 2020. 1

[48] Abby Stylianou, Richard Souvenir, and Robert Pless. Visualizing deep similarity networks. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 2029–2037. IEEE, 2019. 4

[49] Ravichandran Subramanian, Sudeep Sarkar, Miguel Labrador, Kristina Contino, Christopher Eggert, Omar Javed, Jiejie Zhu, and Hui Cheng. Orientation invariant gait matching algorithm based on the kabsch alignment. In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2015)*, pages 1–8. IEEE, 2015. 2

[50] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 501–518, Cham, 2018. Springer International Publishing. 6, 8

[51] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9. 2021. 1

[52] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4997–5006, 2019. 2

[53] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 1, 2

[54] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12026–12035, 2021. 2

[55] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2

[56] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. Explainable face recognition. *arXiv preprint arXiv:2008.00916*, 2020. 4

[57] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3

[58] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020. 2

[59] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision*, pages 868–884. Springer, 2016. 2, 5, 8

[60] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2, 6, 8