An Empirical Study of Person Re-Identification with Attributes

Vikram Shree[1], Wei-Lun Chao[2] and Mark Campbell[1]

***Abstract*— Person re-identification aims to identify a person from an image collection, given one image of that person as the query. There is, however, a plethora of real-life scenarios where we may not have a priori library of query images and therefore must rely on information from other modalities. In this paper, an attribute-based approach is proposed where the person of interest (POI) is described by a set of visual attributes, which are used to perform the search. We compare multiple algorithms and analyze how the quality of attributes impacts the performance. While prior work mostly relies on high precision attributes annotated by experts, we conduct a human-subject study and reveal that certain visual attributes could not be consistently described by human observers, making them less reliable in real applications. A key conclusion is that the performance achieved by non-expert attributes, instead of expert-annotated ones, is a more faithful indicator of the status quo of attribute-based approaches for person re-identification.**

I. INTRODUCTION

Accurate identification of people in crowded environments plays an indispensable role in various applications such as pedestrian tracking and surveillance. The specific task of person re-identification (re-ID) is to match images of people across diverse scenes, taken from different camera views or spatial locations.

One common assumption of person re-ID is the accessibility to a set of probe (query) images, which are to be matched to a different set of gallery images. Prior work has focused on extracting discriminative features from individuals' appearances [1]–[3] and designing (or learning) appropriate distance metrics for feature matching [2], [4]–[7]. More recently, deep convolutional neural networks (CNNs) have been widely applied for the re-ID problem because of their flexible architectures which jointly learn the two stages [8]–[11], significantly improving the overall performance.

In contrast to typical re-ID problem, there are numerous scenarios in which there is no access to the query image of the person whom we want to identify. This problem is important in applications such as search and rescue, surveillance and suspect finding. Without prior images, additional information is required to initialize the search. It is hypothesized that visual attributes such as clothing, hair color, and footwear type embody rich semantic information about a person's visual appearance and could serve as appearance descriptions. For example, consider a situation where person 'A' is trapped in a building due to an unfortunate event and their friend 'B' wants to search them with a camera-mounted robot. To accomplish this task autonomously, the robot must rely upon the appearance information provided by person 'B'; this assumes that an image of person 'A' is not available at the moment. Even if an image is available, it may not match the current physical attributes of person 'A'. We abstract this task to be an attribute-to-image search problem where the robot searches for person 'A' in the gallery of person-images from the scene, as illustrated in Figure 1. Since the task is no longer an image-to-image comparison, conventional re-ID methods become inadequate.

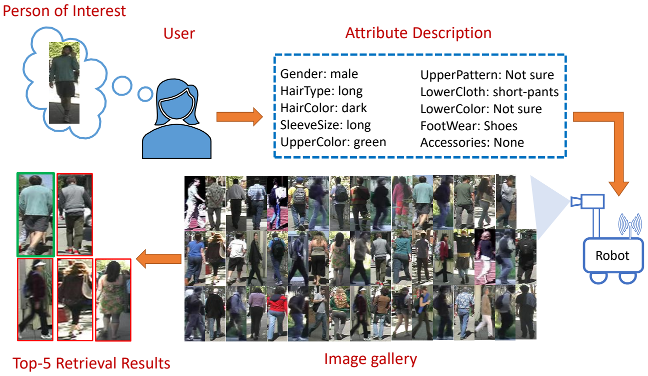

Fig. 1: Illustration of person of interest (POI) search problem. A user describes the attributes that characterize the POI. A robot, by looking at the scene, creates a gallery of person-images. Based upon the attribute input, the robot returns the *top-5* most relevant images from the gallery.

Removing the query images introduces two main challenges to the re-ID problem. First, the query-data and the gallery images lie in different domains. This multi-modality in the data complicates the learning task. To address this problem, we leverage zero-shot learning (ZSL) methods which focus on associating data from different modalities [12]–[15]; thus, the approach is referred to as *zero-shot re-identification* [16], [17]. The success of zero-shot re-ID depends highly upon the feature representation that we choose for the images. For this, we leverage a deep-CNN [8] which extracts highly discriminative features from the gallery images.

The second challenge is uncertainty. It is said that, "A picture is worth a thousand words", implying that there is inherent uncertainty when people are asked to describe the appearance of a person. The uncertainty becomes even more severe when people provide the visual attributes based upon their recollection. Consequently, not all the attributes are correctly reported by the user during real-world applications and could hamper the performance of the zero-shot re-ID

[1] Vikram Shree and Mark Campbell are with the Sibley School of Mechanical and Aerospace, Cornell University, USA {vs476, mc288}@cornell.edu

[2] Wei-Lun Chao is with the Department of Computer Science, Cornell University, USA wei-lun.chao@cornell.edu

system. For example, in Figure 1, the user is not sure about the pattern of the clothing that the POI is wearing. A key limitation of the recent works in zero-shot re-ID is their reliance on expert-labelled attributes, obtained in a laboratory setting, which do not account for such uncertainties that may arise in the wild. To address this issue, we present an exploratory human-subject study to select the most distinctive attributes for training and testing ZSL models. The major contributions of this paper can be summarized as follows:

- First, in order to perform attribute-based person re-ID, we utilize a state-of-the-art feature representation for the images and evaluate the performance of several representative ZSL models on publicly available datasets.
- Second, we design and conduct a human-subject study to identify key attributes that are consistently reported correct by various users. These are defined as *distinct* attributes.
- Finally, we leverage the *distinct* attributes for training ZSL models. The performance of zero-shot re-ID is evaluated using non-expert attributes and compared with that obtained from expert labels.

II. Related Work

A. Person Re-ID

Traditional methods for re-ID have mainly focused on searching for better hand-crafted features such as chromatic content, spatial arrangement of colors, texture etc. [18], [2], [19]. The intent is to find features that are mostly invariant to changes in pose, viewpoint and lighting conditions. Some hand-crafted features have the advantage of being human-interpreteble, for example visual attributes, and can serve as appearance description for zero-shot re-ID. This is followed by a metric learning step which maps the features into a new space where feature vectors corresponding to same person are close to each other [4], [5], [20]. Hand-crafted features, however, are usually not discriminative enough for differentiating idetities, leading to poorer performance.

Recently, supervised learning frameworks comprising of CNNs have been used for re-ID because of their ability to capture semantic and spatial information [8], [9], [21]–[23]. Broadly, the methods can be divided into two categories: deep representation learning and deep metric learning. The first aims at creating a discriminative feature representation for the images [8], [24], [25]. In [24], a robust feature embedding is learnt by training the model in multiple domains with domain guided dropout. In [8], the authors fuse multiple embeddings across layers to capture spatial details lost in the last layer. In contrast, deep metric learning intends to learn the similarity between images belonging to the same person [10], [26], [27]. In [10], a siamese CNN model is leveraged for jointly learning the features and the metric for re-ID. In [26], a multi-task method is proposed to integrate the classification and ranking task together. The state-of-the-art supervised learning approaches for re-ID have achieved formidable feature representation capability. However, there is still skepticism about the scalability because of their heavy reliance on large amount of training data, demanding careful design of learning objectives and network architectures [28], [29]. Besides, their inadequacy in handling multi-modal data paves the way for ZSL models in zero-shot re-ID problem.

B. Application of Attributes in Re-ID

Visual attributes such as clothing, hair-style, shoe-type, accessories etc, have been used as mid-level feature representation for re-ID in a number of works [16], [30]–[33]. In [31], the authors propose an attribute-recognition network which combines ID classification and attribute classification losses. Because the attributes are human-interpretable, they are useful for zero-shot re-ID [16], [17]. In [17], the authors train an attribute-classifier and use the attributes as feature representation. Further, a distance metric is used for comparing similarity of images in the attribute space. In [16], a clustering scheme is used to determine the useful attributes in order to maximize re-ID performance. Instead of just relying upon attributes, a few authors utilize them as auxiliary information, thus inhibiting their application to the zero-shot re-ID setting. In [30], features from the CNN are fine-tuned on a pedestrian attribute dataset by adding a combination attribute loss term.

In this work, we leverage visual attributes for characterizing person-appearance in the zero-shot re-ID problem. However, methods relying on attribute detection achieve inferior performance because high-quality attribute prediction is difficult when training data is sparse and images have low resolution. This is our primary motivation for taking a representation-learning route, where the gallery images are projected into a discriminative feature embedding, followed by a ZSL model which associates the query attributes to the projected gallery features.

III. Problem Formulation

Consider a gallery with K distinct person images within it. Let us assume that the query set consists of attributes for M people to be searched in the gallery. Define $\mathcal{X}_f^g = \{\mathbf{z}_f^i | i = 1, \ldots, K\}$, as the set of features for the gallery images, each being $d_f \times 1$ dimensional and $\mathbf{U}^g$ denotes the associated identity labels i.e. $\mathbf{U}^g = [u^1, \ldots, u^N]^T$, where $u^i \in \{1, \ldots, K\}$. Similarly, define $\mathcal{X}_A^q = \{\mathbf{x}_A^j | j = 1, \ldots, M\}$, as the set of attributes for query people, each being $d_a \times 1$ dimensional.

A ZSL module learns a classifier function $\mathbf{f}_k : (\mathbb{R}^{d_a} \times \mathbb{R}^{d_f} \times \ldots \times \mathbb{R}^{d_f}) \rightarrow \mathbb{R}$, which yields the likelihood of the query attributes belonging to each person-class k. This is followed by predicting labels $\hat{v}^j$ for the queries, as follows:

$$\hat{v}^j = \arg \max_{k \in \{1, \ldots, K\}} \mathbf{f}_k(\mathbf{x}_A^j, \mathbf{z}_f^1, \ldots, \mathbf{z}_f^K)$$

IV. Pedestrian Attributes

Visual attributes encode high level appearance information which enables humans to distinguish between different people e.g. clothing, shoes, hair-style etc. Since hand-labelling attributes is a time-consuming and tedious process, in this work, we leverage the attribute labels provided in Pedestrian Attribute dataset (PETA) [34]. PETA is a large public

TABLE I: Pedestrian attributes categorized into Mutually Exclusive Classes.

	Classes	Attributes or Meta-attributes	Number of Attributes
Q1	Gender	female, male	2
Q2	HairType	short, long, bald	3
Q3	HairColor	*binaryDark*, *binaryLight*	2
Q4	UpperCloth	jacket, suit, sweater, t-shirt	4
Q5	SleeveSize	long, short, noSleeve	3
Q6	UpperColor	red, blue, green, *dark*, *light*	5
Q7	UpperPattern	*stripes*, plaid, logo	3
Q8	LowerCloth	formalPants, *informalPants*, *shortPants*, skirt	4
Q9	LowerColor	red, blue, green, *dark*, *light*	5
Q10	FootwearType	sandals, *allShoes*, boots	3
Q11	FootwearColor	binaryDark, binaryLight	2
Q12	Carrying	backpack, messengerBag, plasticBag, folder	4
Q13	Accessory	headphones, sunglasses, hairband, hat	4
Total number of attributes			**44**

TABLE II: Collection of basic attributes combined to create meta-attributes.

Meta-attributes	Basic attributes
binaryDark	black, blue, purple, brown
binaryLight	white, pink, orange, green, yellow, grey, red
dark	black, purple, brown
light	white, pink, orange, yellow, grey
informalPants	jeans, trousers
shortPants	shorts, capri, hotPants
allShoes	shoes, sneakers, leatherShoes
stripes	thickStripes, thinStripes

collection of images with labelled attributes; consisting of 61 binary and 4 multi-class attribute labels.

We divide the attributes in PETA into 13 mutually exclusive classes, each one consisting of its own list of attributes. It is known that a high dimensional function is more difficult to learn than a low dimensional function in the absence of sufficient data. In order to simplify the learning problem, we prune the attributes. In particular, we remove rare attributes in the dataset and categorize multiple basic attributes under a common meta-attribute e.g. top clothing colors like black, brown, and purple have been combined into a single attribute called ‘UpperBodyDark’. Furthermore, we exclude certain attributes e.g. age, due to the difficulty in its consistent prediction based upon the low-resolution images in the dataset. The final set of attributes and meta-attributes in each class is presented in Table I. The attributes that constitute the meta-attributes are listed in Table II.

In this work, all attributes and meta-attributes belonging to a class are encoded by binary values. As a result, a 44 dimensional vector $\mathbf{x}_A^j$ is obtained in the attribute space, comprising of ‘0’ and ‘1’, where a ‘0’ signifies the absence of an attribute/meta-attribute in the image and ‘1’ signifies its presence in the image. It should be noted that considering continuous attribute values could be useful for certain classes, like UpperColor; however, we were limited by the binary nature of the labelled data in PETA.

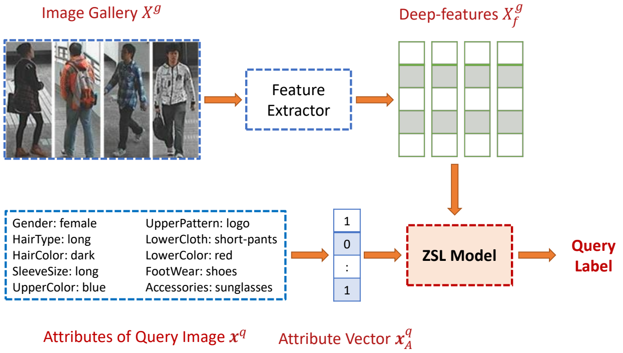

Fig. 2: Illustration of a general zero-shot re-ID pipeline. The feature extractor maps the gallery images into semantic embedding and ZSL models learns to associate the attributes with the predicted semantic embedding, thus assigning labels for the query attributes.

V. APPROACH

We describe the framework for addressing the zero-shot re-ID, where the task is to classify the query attributes into label space of the gallery images. A general approach is shown in Figure 2. There are two important steps involved in zero-shot re-ID. The first step is to learn a discriminative feature representation for the people-images, referred to as the feature extractor. The second step involves learning to associate the attributes with the deep-feature representations. In this section, we describe both the steps, while assuming that labelled attributes are available for the query images.

A. Learning Feature Representation

The goal of feature learning is to transform the images into semantic embedding space where the Euclidean distance between features belonging to the same person is smaller than those of different people. As mentioned in Section II, several deep-CNN models have been proposed in order to learn feature representations for re-ID, from images. In this work, we leverage Deep Anytime Re-ID (DaRe) [8], a state-of-the-art re-ID framework which also allows trade-off between computational resource requirement versus performance; this makes it suitable for robotics applications. The DaRe architecture consists of several sequential, convolutional stages and finally, a fusion stage. Information is fused across multiple layers to capture both coarse semantic information and fine-level details. To achieve this, the feature maps at intermediate layers are first passed through a fully connected layer, which brings them to the same dimension. For a given image $\mathbf{x}$, assume that $\phi_s(\mathbf{x})$ denotes the embedding produced at stage s. Subsequently, the deep-feature $\mathcal{X}_d$ is obtained by doing a weighted sum of the intermediate embeddings:

$$\mathcal{X}_d = \phi_{fusion}(\mathbf{x}) = \sum_{s=1}^{S} w_s \phi_s(\mathbf{x}), \tag{1}$$

where S is the total number of stages in the network and w_s are learnable parameters. The overall loss function used for training the network consists of individual loss functions for each of the layers and a final fusion loss function. Each loss function is computed based on triplet loss since [29] showed that it has superior performance as compared to conventional surrogate losses for the re-ID task.

B. Associating Attributes with Deep-Features

Numerous techniques have been proposed in ZSL literature to associate the query-attributes to the visual features of the gallery [15], [35]. We describe three broad classes of ZSL algorithms, different by how the association is achieved.

1) Learning to predict visual embedding: Given the query-attributes $\mathcal{X}^q_A$, one can learn to predict the visual embedding i.e. the deep-feature vectors $\mathcal{X}^q_d$. Consequently, the identity-label of the query can be obtained using a similarity measure between the predicted embedding $\mathcal{X}^q_d$ and the gallery-features $\mathcal{X}^g_d$. In this paper we apply EXEM [15], a state-of-the-art ZSL algorithm that uses support vector regression to predict visual embedding, followed by nearest neighbor in the Euclidean space for label assignment.

2) Learning to predict attributes: Given the deep-feature representation of the gallery images $\mathcal{X}^g_d$, one can learn to predict the attributes $\mathcal{X}^g_A$. Based upon the similarity between the predicted attributes $\mathcal{X}^g_A$ for gallery images and the given query attributes $\mathcal{X}^q_A$, the query-label can be estimated. In this paper we apply DAP [12], which learns a probabilistic classifier to predict each attribute and labels the query by MAP assignment. The main difference between EXEM and DAP is the space in which the similarity is computed: EXEM computes similarity in the visual space, while DAP does so in the attribute space.

3) Compatibility learning: In this approach, a common representation is learned, onto which both the query-attributes and deep-features of the gallery, are projected. This is followed by an optimization which maximizes the compatibility score of the projected instances in the space. Different methods under this category differ in their choice of the common space or the compatibility function. In this paper we apply ESZSL [13], which uses bi-linear mapping for the projection and has an efficient closed form solution.

VI. EXPERIMENTAL EVALUATION

A. Dataset and Implementation Details

We implement three ZSL algorithms—EXEM [15], DAP [12], and ESZSL [13]—as described in Section V-B, and evaluate the performance with labelled attributes on two public datasets: a) VIPeR [36] and b) PRID [19].

The VIPeR dataset consists of 632 pedestrian image pairs taken from two cameras with significant variation in viewpoint and illumination. All the images in the dataset are scaled to 128×48 pixels. The dataset is divided randomly into three sets: train, validation and test sets. There are 280 pedestrian image pairs in the train set, 36 in the validation set and 316 in the test set.

The PRID dataset consists of both single-shot and multi-shot images, taken from two cameras. Similar to [17], we only use the first 200 person-images for each camera, since they appear in both the cameras. All the images in the dataset are scaled to 128×64 pixels. The dataset is divided randomly into three sets: train (80), validation (20) and test (100) sets.

TABLE III: *Cumulative Matching Characteristic* accuracy for zero-shot re-ID on VIPeR and PRID dataset. Superior results are shown in **bold**.

	Method	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-25**
VIPeR	*Layne et al.* [17]	6.0	17.1	26.0	48.1
	DAP [12]	7.0	25.6	34.5	56.7
	ESZSL [13]	**8.9**	27.2	41.5	61.1
	EXEM [15]	7.9	**31.3**	**43.0**	**62.0**
PRID	*Layne et al.* [17]	8.0	29.0	47.0	73.0
	DAP [12]	12.0	38.0	50.0	**78.0**
	ESZSL [13]	**14.0**	**48.0**	**55.0**	76.0
	EXEM [15]	12.0	32.0	47.0	72.0

Following Section IV, 44 dimensional attribute vectors are created for the images in VIPeR and PRID, based on expert-labels available from PETA dataset. The ZSL models are trained with the computed attribute vectors and the deep-features produced by DaRe network. We utilize the ResNet-50 implementation of the DaRe for learning the deep-feature representation, training it with the same protocol as in [8].

B. Results

We compare the performance of different zero-shot re-ID methods, obtained with the DaRe feature representation, to the previous state-of-the-art approach by *Layne et al.* [17]. The standard Rank-N *Cumulative Matching Characteristic* accuracy is used as the performance metric.

Table III summarizes our results on both the datasets. For VIPeR dataset, we observe that all the representative ZSL approaches, DAP, ESZSL and EXEM, outperform the baseline by a significant margin. For PRID dataset, the ZSL models achieve superior accuracy at low rank values. However, at higher ranks, the accuracies are competitive.

The improvement over the state-of-the-art approach can be attributed to two main factors. First, DaRe features are more identity-discriminating than the previously used feature representations. Second, in contrast to other re-ID techniques which rely on very high dimensional features, DaRe features are fairly low dimensional (128). With scarce attribute data, learning a lower dimensional mapping from the attribute space to the deep-feature space is easier than learning a high dimensional mapping. Comparing among the three ZSL algorithms, we found that EXEM and ESZSL outperform DAP in general, which aligns well with the observations in object recognition [15], [35].

VII. ATTRIBUTE SIGNIFICANCE STUDY

Previous works in attribute-based re-ID have relied on expert attributes. This is a valid assumption in a laboratory setting; however, in a real-world application, the user may not remember all the attributes corresponding to a person of interest. Different techniques have been proposed to handle missing attributes like mean-imputation [37]; yet, having more imputed attributes can significantly impede re-ID performance. Furthermore, some of the attributes can be ambiguous and different people might perceive it differently e.g. a messenger-bag can be confused with a plastic-bag.

In this work, rather than using all the attributes (Table I), available in the PETA dataset, we propose to rely only upon *distinct* attributes. The major drawback of using all the attributes is that ambiguities in attribute-inference during test will impede the performance of the trained re-ID module. Consequently, we have defined *distinct* attributes as the ones whose predictions consistently match the expert-labels. This motivated us to perform a human-subject study that analyzes the significance of attributes from a non-expert's perspective. In this section, our experimental setup and data processing steps involved in the study are described.

A. Research Question

The goal of this study is to evaluate whether humans can perceive and recall visual attributes of people, whose images they have seen before.

Hypothesis – Given an image of a person, humans are better than a randomized algorithm at identifying and memorizing the *distinct* visual attributes of the POI.

B. Experiment Design

In order to test our hypothesis, we modelled a setup where participants were asked to look at an image of a person for a limited amount of time. This was followed by a questionnaire[3] about the attributes that are present in the image. This setup is an attempt to model a real-world situation where a person is trying to find someone whom them have seen in the immediate past. The limited time setting is important to identify the attributes that stand out to humans. In contrast, letting humans examine an image for extended period of time would allow them to memorize every detail in the image and thus defeat the purpose of the study.

Each participant repeated the task five times with different images, each taken from a different re-ID dataset i.e. VIPeR [36], CUHK [38], PRID [19], TownCentre and iLIDS [39]. The examination time for each image was decided based on trial run of the study on 6 participants, where 4 out of 6 participants reported that 15 seconds is enough time for them to identify and remember the attributes. Consequently, we fixed the examination time to be 15 seconds for the actual study.

The questionnaire consists of 13 questions, one corresponding to each attribute class listed in Table I; the options represent the attributes present in that class. There was no time constraint for answering the questionnaire. Not all classes used in the study have exhaustive set of attributes e.g. a person might wear a cloth type that is not present in the class 'UpperBody'. As a result, we have included 'None of the above' option for such classes. Broadly, the questions can be divided into two types: a) multiple correct type which have more than one correct choice e.g. accessory, and b) single correct type which have only one correct option e.g. gender. However, irrespective of the type of question, participants can always choose 'Not sure' option, if they are doubtful.

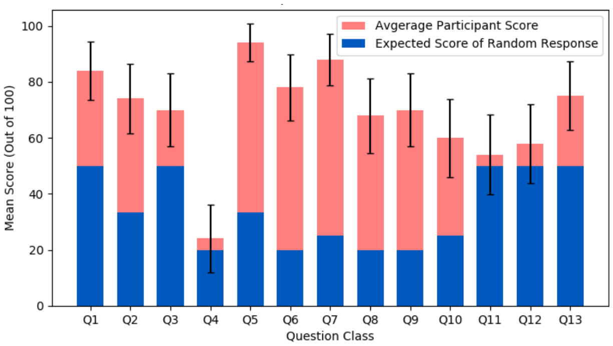

Fig. 3: *Mean* score of participants and *expected* score of random guess for each attribute class. Error bars indicate 95% confidence intervals.

C. Data Processing

To evaluate the performance of a participant in the study, we compare their response with the label attributes, available from PETA dataset, and assign scores based on the comparison. The following intuitive scoring scheme is used:

- For single correct type questions, we assign 100 points for choosing the correct option and 0 points for choosing an incorrect option.
- For multiple correct type questions, we assign $\frac{100}{n}$ points for choosing each correct option and 0 points for the incorrect ones, where 'n' is the number of correct options in the question.

D. Results

The participation for the study was voluntary and a total of 10 participants completed it. All of them are graduate students, enrolled at Cornell University and belong to the age group of 20-30 years. Each participant examined and labelled five different images, sequentially. Figure 3 shows the mean score of participants and expected score of a randomized algorithm. We observe that the error bars for classes Q4, Q11 and Q12 overlap with the expected random score, indicating that humans perform poorly at identifying the attributes in those classes; thus, they should not be considered as *distinct*.

To substantiate our claims about *distinct* attributes, we ran a one-tailed t-test for each question class, comparing participants' score with the expected score of the random guess. A *significance level* (α) of 0.05 was used, which corresponds to confidence interval of 95%. As shown in Table IV, attribute classes Q4, Q11 and Q12 have p-values higher than the threshold of 0.05, and hence, cannot be regarded as *distinct* attributes. In contrast, the 10 other classes have p-values significantly less than 0.05, implying that our hypothesis is supported. Consequently, we consider the attributes belonging to these classes as *distinct*.

E. Discussion

We believe there are two main reasons behind humans performing only marginally better than random guess for certain attributes: a) inconspicuous attributes and b) linguistic

[3] A sample of the questionnaire that was used in the study can be viewed at https://github.com/vikshree/humanRobotReid

TABLE IV: Participant Scores and *t*-Test Results for each Attribute-Class. * denotes *distinct* attribute classes.

Classes	Sample Mean	*Expected* Rand. Score	Std. Error	*t*-Stat	*p*-Value
Q1*	84.0	50.0	5.237	6.492	< 0.0001
Q2*	74.0	33.33	6.266	6.492	< 0.0001
Q3*	70.0	50.0	6.546	3.055	0.002
Q4	24.0	20.0	6.101	0.656	0.258
Q5*	94.0	33.33	3.393	17.883	< 0.0001
Q6*	78.0	20.0	5.918	9.80	< 0.0001
Q7*	88.0	25.0	4.642	13.571	< 0.0001
Q8*	68.0	20.0	6.664	7.203	< 0.0001
Q9*	70.0	20.0	6.546	7.638	< 0.0001
Q10*	60.0	25.0	6.999	5.001	< 0.0001
Q11	54.0	50.0	7.120	0.562	0.288
Q12	58.0	50.0	7.051	1.135	0.131
Q13*	75.0	50.0	6.103	4.096	< 0.0001

Number of samples (*n*) = 50

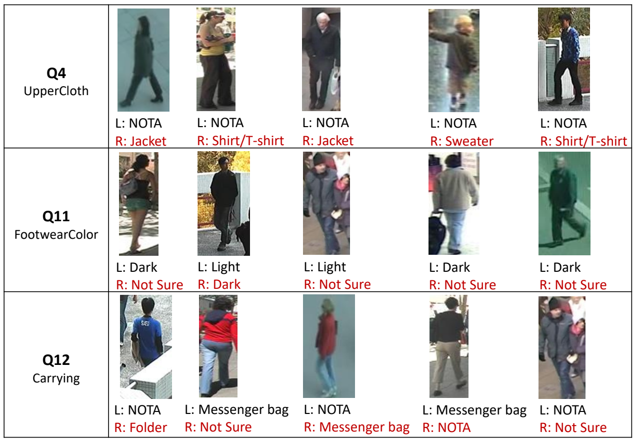

Fig. 4: Randomly chosen *incorrect* responses from the survey for Q4, Q11 and Q12 attribute class. 'L' denotes label data and 'R' denotes users' response. 'NOTA' - None of the above.

biases. First, some attributes are inherently hard to notice or remember. For example, in Figure 4, we can observe that most people are unsure about the 'footwear-color'. Second, different people can have differences in their choice of words for a certain visual attribute. This is referred to as linguistic bias [40], and may arise from cultural, regional or gender differences. For instance, in Figure 4, for the first image Q4 class, the person seems to be wearing a 'Jacket', however, the label suggests otherwise. Presence of linguistic biases in the labelled data could degrade the re-ID performance in real-world applications.

Generalization: Since the survey was conducted with a few number of participants from a particular section of the society (graduate students at Cornell University), the conclusions regarding the *distinct* attributes may vary across different population samples e.g. policemen may focus on very specific attributes for identifying a person. However, the important point that we want to emphasize upon is that the uncertainty in attribute-inference remains a challenge and depending upon the application, such a study is important to identify the key attributes that should be used for training the learning module.

F. Limitations

Although, our experimental setup is designed to simulate a realistic problem setting, there are certain limitations to this study which motivate future work. After the survey, most participants reported that their performance was obstructed due to the low resolution of the images, as the datasets were mostly obtained from surveillance videotapes. In an actual robotics application of zero-shot re-ID, the user would probably recall more detailed features about the person of interest, such as facial features, which were not used in our study. In this work, we were mainly limited by the attribute dataset. However, if a more descriptive dataset is available in the future, one could conduct the same study to identify the *distinct* attributes and use them for training and testing ZSL models.

VIII. RE-ID WITH NON-EXPERT ATTRIBUTES

Having established the distinctiveness of certain attributes in Section VII, we conclude that not all the attributes can be reliably observed by humans, assuming they do not have access to the query images. In the context of zero-shot re-ID, incorrect and missing attributes could potentially penalize the accuracy of ZSL models; thus, it is important to only rely upon the *distinct* attributes. To justify this, we experiment with attributes provided by our human observers (denoted as non-expert attributes), in comparison to the expert-annotated ones, provided in PETA. This section describes the experimental design steps and results.

A. Experiment Design

The proposed experimental setup is exactly as in a typical zero-shot re-ID setting, where attributes are used to match the query images to the gallery. The only difference here is that the attributes are now reported by humans. To fetch the attributes, we designed a survey, similar to the one in Section VII, where participants can label the attributes corresponding to the query images. The participants are allowed to examine each image for limited time (15 seconds), which is followed by the questionnaire about the attributes. However, unlike Section VII, the questionnaire only consists of the *distinct* classes i.e. Q1, Q2, Q3, Q5, Q6, Q7, Q8, Q9, Q10, and Q13. Although, using only the *distinct* classes does not guarantee no missing and incorrect attributes, we expect the use of *distinct* attributes to be more reliable.

B. Dataset and Implementation Details

We performed the experiment using VIPeR dataset [36]. The training and validation splits are the same as Section VI-A, while the test set consists of 50 pedestrian image pairs, randomly selected from the original test set. The ZSL models are always trained with expert-annotations, obtained from PETA dataset. However, the testing is performed based on both expert and non-expert annotated attributes. Non-expert annotations are retrieved from the survey, conducted with 10 participants, each responding to the questionnaire for 5 images; that covers all the 50 query images in the test set. The participation for the survey was voluntary. All

TABLE V: *Cumulative Matching Characteristic* accuracy for zero-shot re-ID on VIPeR dataset, consisting of 50 person-image pairs.

	Type of Test Data	Rank-1	Rank-5	Rank-10	Rank-25
DAP	Expert, all attributes	28.0	68.0	86.0	96.0
	Expert, *distinct*	28.0	72.0	88.0	94.0
	Non-Expert, *distinct*	22.0	44.0	72.0	88.0
ESZSL	Expert, all attributes	28.0	66.0	90.0	100
	Expert, *distinct*	30.0	76.0	90.0	100
	Non-Expert, *distinct*	22.0	56.0	80.0	100
EXEM	Expert, all attributes	32.0	60.0	78.0	98.0
	Expert, *distinct*	22.0	66.0	76.0	96.0
	Non-Expert, *distinct*	22.0	60.0	76.0	96.0

of them are graduate students, enrolled at Cornell University, belonging to the age group of 20-30 years. In addition, it was ensured that the participants of this survey are disjoint from the participants of the study in section VII. A few missing attributes can be encountered in the responses collected from participants; thus, during testing, the mean of training data is used for filling in missing attributes.

C. Results

We test the performance of ZSL models with: 1) Expert-annotated full attribute set, 2) Expert-annotated *distinct* attributes, and 3) Non-Expert annotated *distinct* attributes. The results are presented in Table V.

Comparing the performance in first and second cases, we observe that DAP and ESZSL achieve superior performance when they are trained and tested with the *distinct* labelled attributes. EXEM achieves similar accuracy in both the cases. This suggests that using less number of attributes does not hurt the performance and indeed improves it for certain models. One explanation for this is that there is positive correlation between attributes being *non-distinct* and having a lower ability to discriminate between different people. For example, footwear color is not a *distinct* attribute, based on our survey in Section VII. Also, most people tend to wear dark colored footwear, thus diminishing the discriminative power of that attribute. Therefore, removing 'footwear-color' attribute will not degrade the re-ID performance. In the contrary, it could improve the performance by simplifying the learning task, since the size of one of the inputs to the model has reduced.

The third case revealed the uncertainties that could arise during the deployment of the re-ID system for real-world applications. Comparing the participant responses[4] with expert-annotated, *distinct* attribute data shows that on an average, there are 12 (out of 34) attributes that are either missing or incorrect. Consequently, we observe that the performance is noticeably diminished with non-expert annotations as compared to the expert-annotated case, for all three ZSL methods. In contrast to the current scenario, where the participants only annotate 34 *distinct* attributes, using all 44 non-expert annotations in the ZSL models would have led to an even higher number of missing and incorrect values, thus impeding the performance even further.

Moreover, we observe that ESZSL achieves an impressive Rank-25 accuracy of 100, even with non-expert annotated attributes. For a test set with 50 images, this implies that the search space can be shrinked to half of its original size, while still being confident that the POI is in the reduced search volume.

IX. CONCLUSION

In this paper, an attribute-based zero-shot re-ID solution is developed. By combining learned deep feature representation that renders high discrminative capability across identities, and representative ZSL models, we achieve state-of-the-art performance. Our human-subject study suggests that the task of estimating attributes encompasses uncertainties in the real-world. Consequently, we demonstrate that not all attributes are reliable for the re-ID task, in a realistic scenario. Finally, we evaluate practical sensitivities of the approach in realistic robotics applications by comparing ZSL methods tested with non-expert annotations vs expert-annotated attributes. To the best of our knowledge, this is the first experimental evaluation of zero-shot re-ID methods where attributes are reported by participants, entirely on the basis of their recollection of the query images. We found that the performance of some ZSL models can be severely impaired with incorrect and missing attributes in the wild. Thus, to execute zero-shot re-ID in real-world applications, it is necessary to take the uncertainty/unreliability in attribute annotations into account, for example, by only utilizing the *distinct* attributes in our model.

References

[1] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1363–1372.

[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197–2206.

[3] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European conference on computer vision*. Springer, 2014, pp. 536–551.

[4] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European conference on computer vision*. Springer, 2014, pp. 1–16.

[5] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *European conference on computer vision*. Springer, 2012, pp. 780–793.

[6] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person re-identification*. Springer, 2014, pp. 247–267.

[7] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[8] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8042–8051.

[9] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3908–3916.

[4] The participant responses and the images used for the re-ID experiment can be viewed at `https://github.com/vikshree/humanRobotReid`

[10] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 34–39.

[11] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 791–808.

[12] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[13] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International Conference on Machine Learning*, 2015, pp. 2152–2161.

[14] S. Changpinyo, W.-L. Chao, and F. Sha, "Predicting visual exemplars of unseen classes for zero-shot learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3476–3485.

[15] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Classifier and exemplar synthesis for zero-shot learning," *arXiv preprint arXiv:1812.06423*, 2018.

[16] J. Roth and X. Liu, "On the exploration of joint attribute learning for person re-identification," in *Asian conference on Computer Vision*. Springer, 2014, pp. 673–688.

[17] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*. Springer, 2014, pp. 93–117.

[18] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2360–2367.

[19] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.

[20] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3685–3693.

[21] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[22] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

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

[24] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1249–1258.

[25] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 274–282.

[26] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[27] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 403–412.

[28] Y. Zhai, X. Guo, Y. Lu, and H. Li, "In defense of the classification loss for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[29] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[30] T. Matsukawa and E. Suzuki, "Person re-identification using cnn features learned from combination of attributes," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2428–2433.

[31] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017.

[32] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–28.

[33] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2275–2284.

[34] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 789–792.

[35] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[36] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3, no. 5. Citeseer, 2007, pp. 1–7.

[37] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[38] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.

[39] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*. Springer, 2014, pp. 688–703.

[40] C. H. Echols and C. N. Marti, "The identification of words and their meanings: From perceptual biases to language-specific," *Weaving a lexicon*, p. 41, 2004.