A Suspect Identification Framework using Contrastive Relevance Feedback

Devansh Gupta, Aditya Saini, Sarthak Bhagat, Shagun Uppal, Rishi Raj Jain
MIDAS Lab, IIIT Delhi
{devansh19160, aditya18125, sarthak16189, shagun16088, rishi18304}@iiitd.ac.in

Drishti Bhasin
IIT Roorkee
drishti_b@me.iitr.ac.in

Ponnurangam Kumaraguru
Precog Lab, IIIT Hyderabad
pk.guru.iiit.ac.in

Rajiv Ratn Shah
MIDAS Lab, IIIT Delhi
rajivratn@iiitd.ac.in

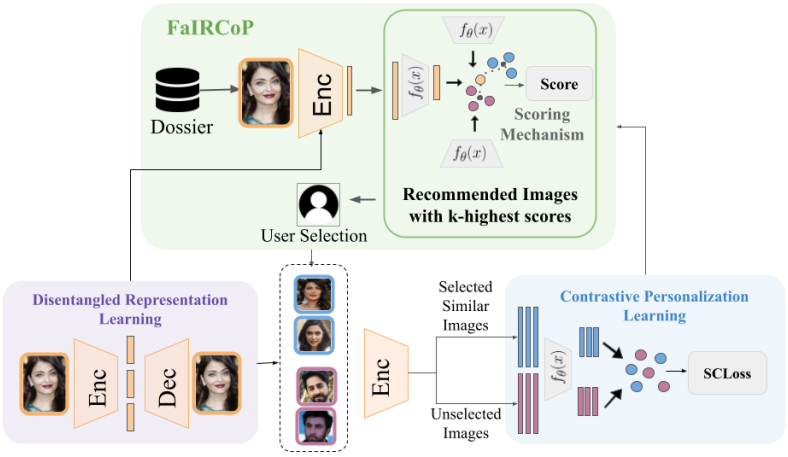
Abstract

Suspect Identification is one of the most pivotal aspects of a forensic and criminal investigation. A significant amount of time and skill is devoted to creating sketches for it and requires a fair amount of recollections from the witness to provide a useful sketch. We devise a method that aims to automate the process of suspect identification and model this problem by iteratively retrieving images from feedback provided by the user. Compared to standard image retrieval tasks, interactive facial image retrieval is specifically more challenging due to the high subjectivity involved in describing a person's facial attributes and appropriately evolving with the preferences put forward by the user. Our method uses a relatively simpler form of supervision by utilizing the user's feedback to label images as either similar or dissimilar to their mental image of the suspect based on which we propose a loss function using the contrastive learning paradigm that is optimized in an online fashion. We validate the efficacy of our proposed approach using a carefully designed testbed to simulate user feedback and a large-scale user study. We empirically show that our method iteratively improves personalization, leading to faster convergence and enhanced recommendation relevance, thereby, improving user satisfaction. Our proposed framework is being designed for real-time use in the metropolitan crime investigation department, and thus is also equipped with a user-friendly web interface with a real-time experience for suspect retrieval.

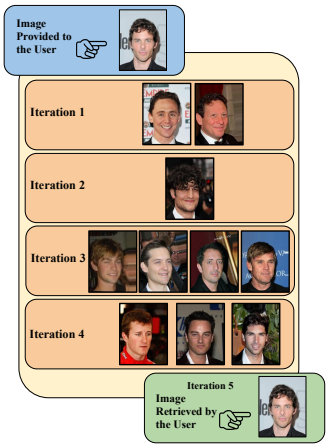
1. Introduction

Interactive facial image retrieval shows great potential in the domain of digital forensics for several tasks such as facial recognition [13] and suspect identification [14, 8]. These systems aim to retrieve the images most similar to the query image by narrowing down the search space using image attribute descriptions. In real time, there is no deterministic knowledge about this query image, but recollections about certain aspects of the image in the user's visual memory. Such supervision can be provided in the form of detailed natural language descriptions [6] or progressive attributes [25] which can be either expensive to annotate well or error prone, especially in tasks such as suspect retrieval where the witness often relies only on their visual memory. Moreover, the high degrees of variation in attributes including pose, illumination, expressions, and occlusions present in different facial images adds to the challenge of developing such systems. In this work, we address these challenges by developing a weakly-supervised facial image retrieval system. For this, we utilize high-level categorical feature attributes as a weaker form of supervision capturing the user's notion of similarity and propose a relevance feedback mechanism by incorporating these cues.

Prior work in the area of facial image retrieval has focused on the utilization of predefined annotated features to retrieve images from a database [26, 25]. This approach limits the user's expressibility to a limited number of tangible attributes and is expensive to train due to the requirement of feature annotations. To alleviate this issue, user feedback has been used to obtain relevant images in an online manner. In such cases, facial features become subject to the user's interpretation, making it crucial for the system to appropriately model user preferences. Some approaches [25, 26, 18] utilized user feedback which explicitly mentions the changes as a query and suggested images. These approaches tend to impose a higher cognitive load on the user since they require them to recall the image's fine details from their visual memory. More user-friendly approaches such as [9, 2, 17] successfully diminish the cognitive load by requiring the user to classify the mental image based on certain predefined parameters. SeekSuspect[8] exploited the similarity-based user feedback mechanism to learn the notion of divergence between similar and dissimilar images



(a) Illustration of our proposed framework FaIRCoP for facial image retrieval. We define the SCLoss for contrastive personalization learning in Equation 2.



(b) A search result from user study on our system showing similar images selected at each iteration.

Figure 1: FaIRCoP - Facial Image Retrieval using Contrastive Personalization system. For security purposes, we do not show the images from the Criminal dataset and only use the CelebA dataset for demonstration.

with respect to the user's mental image model. Despite attempting to learn representations aiming to personalise user preferences, these approaches failed to learn a proper distance metric that encapsulates the variability among various factors of variation within the image.

In this work, we propose a contrastive learning framework for suspect identification that adapts to each witness' personalized notion of similarity. Our loss function, called the Separating Cluster loss, clusters the images selected by the witness while establishing the dichotomy between the selected and non-selected images. We also utilize unsupervised disentangled representation learning to obtain robust image embeddings which separate multiple facial attributes into partitioned latent spaces. This makes the representations more interpretable and also aid in efficiently narrowing the search space without the explicit dependency on labels, which can be expensive as well as noisy. Since our approach requires human interaction as part of our pipeline, our proposed algorithm referred to as the *Facial Image Retrieval using Contrastive Personalization* (FaIRCoP), is equipped with a user-friendly web-based interface for retrieving images in real-time[1]. We also designed a custom user simulator with simulated human feedback to compare our method against various baselines and design choices before performing an extensive user study on two extensive facial image databases. This method performs superior to other methods in retrieving suspects from the criminal database, efficiently handling the high degree of noise in images of the dataset. This method has also been tested for use in metropolitan crime investigation department and is going to be deployed for use in real-time. Figure 1a illustrates the pipeline of our proposed framework.

We summarize our contributions as follows:
(1) A relevance feedback framework, referred to as FaIRCoP, designed for automating suspect identification.
(2) A contrastive learning-based loss function called the *Separating Cluster Loss* for iteratively modifying the search space by clustering selected and non-selected images.
(3) A custom simulator to automate the user feedback to compare the proposed suspect retrieval method with other algorithms.
(4) A responsive web-based interface for real-time suspect retrieval equipped with our proposed algorithm.

2. Related Works

Disentangled Representation Learning. Disentangled representation learning is an approach for encoding high-dimensional data into independent low-dimensional latent space partitions, each capturing a distinct factor of variation. Several works [10, 20] followed this by exploiting limited supervision in order to extract the specified attribute from the rest of the underlying factors of variations. The resulting embeddings provide the model with enhanced interpretability and downstream task performance. However, they are subjected to inherent biases due to their dependence on specified feature annotations of single or multiple factors. Due to these reasons, unsupervised disentangled representation learning has gained traction in the community. Various prior works [7, 15] focus on learning factored representations in a completely unsupervised manner. Such representations that capture each tangible feature into a dis-

[1]We provide more details of the web interface in the supplementary.

crete chunk within the latent space compactly represent data as low-dimensional embeddings that can be used as effective initialization for several underlying downstream tasks.

Contrastive Learning. The contrastive learning paradigm is popularly used to learn representations by comparing different samples in the dataset using distance metrics for structuring the latent space into similar and dissimilar embeddings [1, 21]. SimCLR[3] utilized this idea to maximize the similarity between two views of the same input besides minimizing the similarity between the representations obtained from other images in a batch, leading to a stronger form of self-supervision.

Contrastive learning aims at learning meaningful representations based on positive and negative pairs of embeddings. Hence, in the case of iterative image retrieval we can model the positive and negative pairs through the selections made by the user, where all the pairs of images selected by a user can be modelled as positive pairs. On the other hand, the images not selected by the user act as the negative pair with all the selected images. Thus, using this as a weak supervision, we apply this concept to learn representations that map the notion of similarity specific to the user to a known similarity metric in the latent space.

Image Retrieval. The task of image retrieval using cues from users is a challenging task due to the high level of subjectivity and personalization in the user's conception of different visual features. [25] used a higher form of supervision through natural language in order to retrieve images. Other approaches [26, 18] relaxed these constraints by retrieving images correlated with the current query image based on a set of attributes that are either decided during system formulation or specified by the user. On similar lines, [26] found the nearest set of orthogonal vectors as representatives for independent attributes in the latent space and weighted them by the attribute preferences to obtain modified query vectors. [2] explored the concept of learning similarity metrics in task-dependent projection spaces based on user feedback. Even though these systems vividly exploit some essential characteristics of facial images, they do not consider the information obtained from the previously chosen images selected by the users.

In this work, we utilize the notion of similarity (or dissimilarity) as a form of user feedback, analogous to [9, 2]. Through our framework, we highlight the significance of mapping the similarity notion of a particular user as distances in the latent space, enabling image recommendations closest to the user's mental image. [14] adopted a similar approach for user personalization, however, it relied on labelled supervision.

3. Proposed Approach

We propose a method in which the images selected and not selected by the users can be viewed as positive and negative samples respectively to capture the notion of similarity in the user's mind. We attempt to associate this notion using a certain isotropic metric in the projected latent space of image representations. Thus, we put a constraint that the embeddings of images selected by the user are closer than the ones not selected, in this low-dimensional space. We incorporate such a framework by projecting the pretrained base representations onto a lower-dimensional space using a fully connected neural network and formulate `SCLoss` to train the projection network in such a way that it learns to separate the projections of images relevant and irrelevant to the current query image. We also introduce the concept of anchoring during online training to conserve the notion of similarity and to ensure that only one cluster corresponding to the similar images is being formed and all the images which had not been selected are ideally far from the cluster with a constant number of images required for training. We describe the our proposed relevance feedback algorithm `FaIRCoP` in the supplementary and highlight its essential components in the coming sections.

3.1. Disentangled Representation Learning

Our relevance feedback framework requires good base representations to ensure that the encoder and the projection network do not have an additional overhead of jointly learning good representations, thus, leading to faster convergence and improved latency. Disentangled representations act as an effective initialization due to their ability to encapsulate disjoint factors of variation within specific fixed-sized chunks, resulting in enhanced downstream task performance. This step becomes extremely important as we have designed our system for real-time use, as mentioned in Section 1, which involves images that are often distorted and noisy, even after adequate preprocessing. Shukla et Al[16] empirically show with extensive experiments, that even though disentangled representations exhibit a geometry farther from an Euclidean space as compared to non-disentangled representations for distorted and noisy images, they encode more information about the images which are robust to noise and in turn, give benefits in terms of generalization, fairness, and interpretability leading to improved performance. We discuss their analysis specific to our databases in Section 6. This sets a favorable base representation for our framework, but invites a requirement for a projection network to encode these representations into a space which model the user's notion of similarity. Since our framework is optimized online, it becomes essential that the representations are not inherently biased by labels so that the higher level notion of similarity expressed by the witness is captured faster. Hence, an unsupervised disentangled representation learning method is ideal for extracting representations for our task. We utilize [7] (`MIX`) to extract representations for the database of im-

ages. In our case, this method was also very computationally efficient as standard ResNet-18 representations require around 18 layers along with skip-connections to represent the images while we are able to extract good representations with 5 simple convolution layers.

3.2. Separating Cluster Loss

We propose a separating cluster loss (`SCLoss`) to create a cluster for the images selected by the user in the projected space and ensure that all the non-selected images are farther from the images selected by the user. It is based on the notion of N-pair loss objective [19] which quantifies the loss for the objective of maximizing the similarity between a given pair of embeddings, known as a positive pair, and minimizing similarity with all the other embeddings. The loss equation for a positive pair e and e' along with a set U consisting of all vectors, when paired up with e form negative pairs, with a scaling factor τ[3] is given by Equation 1.

$$l_U(e, e') = -log\frac{e^{sim(e,e')/\tau}}{\sum_{k\in U} e^{sim(e,k)/\tau}} \tag{1}$$

Hence, we can extend this notion into our setting, where, we maximize pairwise similarity between all the projected embeddings of similar images and ensure that all the projected embeddings of dissimilar images are farther from those of the similar images. It can be observed that our loss does not require an equal number of similar and dissimilar images and hence, can be flexibly used during online training. The set S in the below equation represents images selected by the user and D as the set of images not selected by the user.

$$\texttt{SCLoss}(S, D) = \frac{1}{|S|(|S|-1)} \sum_{x\in S} \sum_{y\in S-\{x\}} l_D(x, y) \tag{2}$$

3.3. Online Training and Inference

The objective of the `SCLoss` is to train the projection network in such a way that the images which are similar to the mental image of the user would be nearer to the projected similar images cluster. Hence, we use a scoring function that has been specified in Equation 3. The set S_a represents all the similar images selected by the user.

$$score(u) = sim\left(u, \frac{1}{|S_a|} \sum_{x\in S_a} x\right) \tag{3}$$

It can be observed that as the number of iterations increase, there will be an increase in either the number of similar images, dissimilar images or both. In such cases, the computation required for the loss would become higher and may be computationally infeasible after a certain limit. Conversely, suppose only the similar and dissimilar images of the current iteration are used to create the clusters. In that case, clusters may be created for the two sets independent of similar previous images because there is no way to associate the new images with the previous selections. To circumvent this issue, we propose a training trick called *anchoring* which ensure that the projected representations of new similar images are trained to be a part of the previously formed clusters in the projected space. During training, to ensure we have sufficient images to update our representations, we add a certain number of previously chosen similar and dissimilar images to the corresponding set obtained at the current iteration. This particular trick *anchors* the new embeddings to the previously formed clusters as the network has already learnt to project the previous embeddings to the respective clusters avoiding the formation of multiple clusters in the latent space.

4. Experiments

We evaluate our approach against state-of-the-art baselines for the task of facial image retrieval using a set of qualitative and quantitative experiments.

4.1. Dataset

We utilize a set of two datasets, namely, Criminal Dataset and CelebA Dataset [11], to evaluate the efficacy of our approach against state-of-the-art in facial image retrieval approaches.

4.1.1 Criminal Dataset:

The proposed framework was formulated with a goal to optimize the task of suspect image retrieval to assist criminal investigations based on the witness' mental image of the criminal. For the same, we utilize a criminal data dossier provided to us by the metropolitan crime investigation department unique mugshots of criminals who were previously accused of any kind of crime or are currently involved with the department for an ongoing crime. Each data point in the dossier is associated with attributes describing the criminal's physical attributes that include `faceshape`, `facecomplexion`, etc. However, due to data privacy and confidentiality, we cannot release the dataset. This dataset had considerable amount of noise in the form of alignment and blurriness which were correctly aligned using a pretrained VGGNet [13] finetuned on our dataset while the blurred images beyond recognition had to be discarded bringing the final dataset count to $39,196$ facial images. We then extracted the facial region from the mugshots using Haar Cascades [24].

4.1.2 CelebA Dataset:

We employ the CelebA [11] facial dataset in order to portray the generalizability and qualitative efficiency of our pro-

posed framework. The CelebA dataset contains $202,599$ facial images from $10,177$ identities. Each facial image is labeled with 40 binary attributes, such as `pointynose` and `wavyhair`. However, we chose to discard all the recurring images of the same individual to maintain consistency with the Criminal Dataset which has a unique image for every individual. The CelebA dataset comprises of an exhaustive coverage of various ethnicities and genders, lending it wide acceptance among researchers.

4.2. User Simulator

We utilize a user simulator to make comparisons between different relevance feedback algorithms. The user simulator mimics a human user who has a target image in mind and provides feedback at each round. Each user simulation takes 30 minutes to complete on average. We perform 10 simulations for each algorithm and compare the average of the metrics over these simulations. We design a user simulator to replicate our user-in-the-loop framework.

To mimic the notion of similarity between two images, we design a metric comparing the mean of cosine similarities between the different representations of each of the images. If the similarity between the target image and the image under consideration is greater than a certain threshold, the simulator marks it as similar while the rest of the images are deemed dissimilar. We used a combinations of three image embeddings, namely, Histogram of Oriented Gradients (`HOG`) [4], `FaceNet` [13], and `MIX` [7] to feed the representations in order to calculate the averaged similarity.

The threshold for similarity is determined at the start of each simulation by randomly sampling a constant number of images from the database and computing the similarity from the selected target image. These similarities are then averaged to get the initial threshold for similarity. After every constant number of iterations, the threshold is updated. The exact algorithm for the user simulator can be found in the supplementary.

4.3. Metrics

Average Convergent Iterations (`ACI`)**.** Our use-case has a single unique solution to each query posed by different users. This solution is the target image which presents in a heavily filtered interval from a huge pool of data. Due to the presence of a human-in-the-loop during relevance feedback, it is important to retrieve the target image in the lowest possible number of iterations. We calculate the average number of iterations it took to reach the target image for each simulation to obtain the average convergent iterations. The magnitude of which is directly proportional to the accuracy of the model.

Average Relevance (`AR`)**.** We also use the Mean Average Relevance to quantify the relevance of the images suggested by the model in comparison to the user's mental image. It is indicative of how well the model is able to personalize according to the user's needs. We calculate the relevance for each simulation which is the fraction of similar images chosen by the user out of the total images displayed throughout the process. The mean of this over all the iterations gives us the average relevance.

4.4. Results on the User Simulator

We compare the results on the simulator with FaceFetch [14] and Rocchio Algorithm [17], similar to an intermediate process mentioned in [14] with the `MIX` embeddings. We calculate the metrics mentioned in Section 4.3 for all the aforementioned combinations of image embeddings. These results are displayed in Table 3. In Figure 3, we use t-SNE [22] to visualize the clusters of the similar images selected by the user simulator for four different simulations. Well defined clusters indicate that the notion of similarity for each user simulation was captured differently, thus, qualitatively expressing the level of personalization. Figure 2 indicates that both the `FaIRCoP` is successful at contrasting between the similar and dissimilar images selected by the user simulator as the iterations proceed.

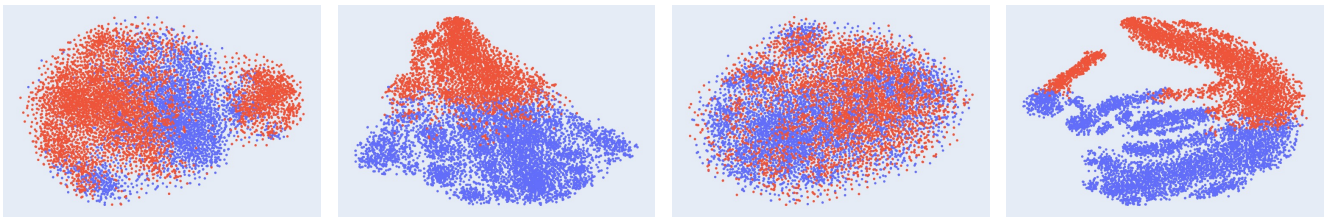
The simulated target images are consistent across algorithms for each simulation. For each experiment, we consider mean of results from 10 simulations. Thus, for each algorithm, 10 target images were used. We ensured that for each simulation, the images recommended in the initial iteration are exactly the same across algorithms. These initial images were chosen using k-means clustering on the whole dataset and sampled two images from each cluster. The simulator is provided with 16 images appropriately recommended by each algorithm in each iteration and the simulator can choose any amount of similar images of those 16 images in the iteration.

5. User Study

We conducted a user study to test the efficacy of our method in real-time scenarios with actual human in the loop

Algorithm	`PREF`	`REL`	`RESP`	`CONV`
Criminal Dataset				
`Rocchio`	0.40	0.41	0.41	0.08
`FaceFetch`	0.40	0.59	0.57	0.28
`FaIRCoP`	**0.70**	**0.72**	**0.81**	**0.44**
CelebA Dataset				
`Rocchio`	0.26	0.43	0.42	0.14
`FaceFetch`	0.5	0.63	0.58	0.29
`FaIRCoP`	**0.63**	**0.71**	**0.67**	**0.36**

Table 1: Cumulative metrics obtained from the User Study conducted on Criminal and CelebA dataset.

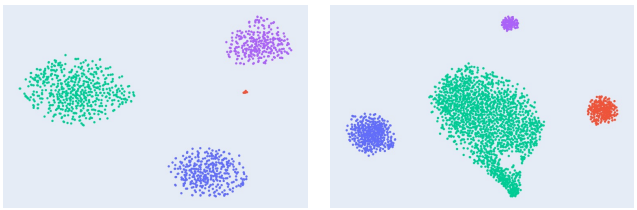


(a) CelebA - Initial (b) CelebA - Final (c) Criminal Data - Initial (d) Criminal Data - Final

Figure 2: Visualization of projected embeddings of all similar (blue) and dissimilar (red) images of the initial (top) and the trained (bottom) projection network for a simulation with `FaIRCoP` for both the datasets.

Metric	Criminal Dataset		CelebA Dataset	
	`ResNet`	`FaIRCoP`	`ResNet`	`FaIRCoP`
$\mathcal{D}$	0.23	**0.30**	0.27	**0.36**
$\mathcal{C}$	0.12	**0.15**	0.21	**0.27**
$\mathcal{I}$	**0.89**	**0.89**	0.88	**0.90**

Table 2: Interpretability metric score comparison for `ResNet` and `FaIRCoP` embedddings.



(a) CelebA Dataset (b) Criminal Dataset

Figure 3: Visualization for user-wise similarity preference clusters in the projected space using `FaIRCoP` for retrieving images on the simulator. In these plots, each color depicts a distinct user.

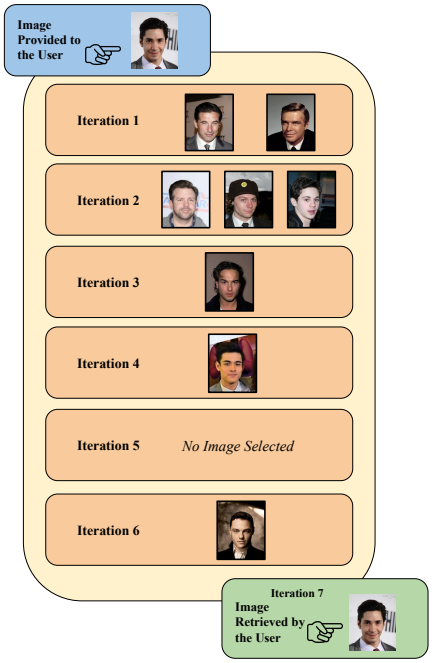


Figure 4: Another run obatined from the user study using `FaIRCoP`. Each box contains similar images selected by the user at each iteration from recommended images until convergence.

and compare the performance with baseline methods for iterative image retrieval using relevance feedback. The study involved 20 participants, with each of them assigned an image from the database. Each user was displayed their image for 40 seconds to generate a suitable visual memory of the image assigned. Based on their visual memory, they searched the image using four separate systems with the `FaIRCoP`, `Rocchio` [17], and `FaceFetch` [14] running at the respective backends. For each search, the users were initially required to select the attributes used to initialize the search with a random set of images that had suitable similarities to the attributes provided. The algorithms showed users 16 images at each iteration, from which they selected similar images and got a recommendation for the new images through the methods mentioned above. This process was repeated until the user reported an image that matched heavily with the user's visual memory. The searches were clipped to a maximum of 30 iterations in case of the criminal dataset [8]. In contrast, they were clipped for a maximum of 25 iterations in the case of CelebA [11]. At the end of each search, the users were asked to fill a questionnaire based on which some performance measures were computed as mentioned in sub-section 5.1. The averaged results for all users are depicted in Table 1. The user study results correlated significantly with the simulation results we obtained regarding the metrics employed.

5.1. Performance Metrics

Due to the presence of a human-in-the-loop during the relevance feedback mechanism, we conduct a user study and evaluate the performance of our model compared

Representation			ACI			AR		
Criminal Dataset								
FaceNet	MIX	HOG	Rocchio	FaceFetch	FaIRCoP	Rocchio	FaceFetch	FaIRCoP
✓	✓	✓	804.22	691.00	**57.25**	0.29	0.15	**0.82**
✓	✓		450.80	506.50	**68.33**	0.45	0.27	**0.83**
	✓	✓	550.95	152.50	**41.66**	0.52	0.38	**0.79**
✓		✓	565.60	457.75	**98.33**	0.34	0.17	**0.79**
	✓		441.30	380.75	**89.00**	0.59	0.36	**0.88**
CelebA Dataset								
FaceNet	MIX	HOG	Rocchio	FaceFetch	FaIRCoP	Rocchio	FaceFetch	FaIRCoP
✓	✓	✓	351.2	263.00	**40.5**	0.37	0.40	**0.61**
✓	✓		358.8	222.8	**27.4**	0.36	0.40	**0.70**
	✓	✓	299.8	255.20	**50.0**	0.54	0.37	**0.87**
✓		✓	309.00	249.00	**98.2**	0.36	0.38	**0.54**
	✓		158.4	227.00	**20.2**	0.53	0.43	**0.82**

Table 3: Quantitative metrics obtained from user simulation using different methods on the Criminal and CelebA dataset.

to other baselines based on the post-study questionnaire, which covers the user feedback on the metrics discussed below.

Relevance (REL)**.** Due to the iterative nature of the mechanism, it is essential that there must be an increasing similarity between the user's visual image and the images recommended as the iterations proceed. The relevance of a system measures the change in similarity perceived by the user between the set of images recommended at each iteration and the visual memory of the user. We asked the users to quantify their ease of selecting similar images as iterations proceeded on a scale of 1 to 5, where 1 denoted high-level mental stress in selection whereas 5 denoted increasing ease as iterations proceeded. We normalized the score to lie between 0 and 1.

Responsiveness (RESP)**.** For iterative image retrieval, it is also essential that the users observe that their previous responses are being used effectively and the images are not randomly recommended. We asked the users to quantify their perceived randomness of recommendations on a scale of 1 to 5, where 1 denoted a high amount of randomness whereas 5 denoted an effective use of previous queries. We normalized the score to lie between 0 and 1.

Convergence (CONV)**.** As a system for image retrieval, we must ensure that the system can converge in fewer iterations. To measure this, we calculate convergence (C) for a search converged in N iterations with a maximum limit of max_iter allowed as given in Equation 4.

$$C = \begin{cases} 1 - \frac{N}{\texttt{max\_iter}+5} & \text{if user reports image} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

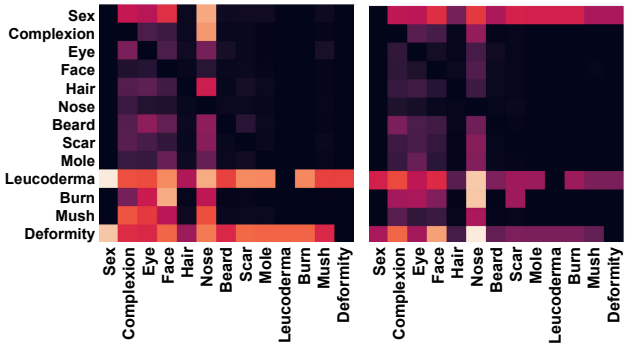


Figure 5: Modified Group Demographic Parity metric for comparing the fairness of FaIRCoP (left, $\mathcal{F} = 0.04$) and ResNet (right, $\mathcal{F} = 0.05$).

Preferability (PREF)**.** Since each user performed the image retrieval on all the algorithms, we asked them to report their willingness to continue using the system if the retrieved images did not match the exact target image. To capture the user's preferablity to use the system. We assigned a preferability score of 0 if the user was unwilling to continue, 0.75 if the user wanted to continue, and 1 if the user retrieved the exact image.

6. Interpretability and Fairness

For any given dataset, semantically meaningful feature learning requires the learnt representations to be interpretable. Due to the interactive nature of the problem, the learnt representations should be fair to avoid introducing any feedback bias among users interacting with the system.

6.1. Interpretability

We evaluated the interpretability of the representations for all the datasets using the Disentanglement ($\mathcal{D}$), Com-

pleteness ($\mathcal{C}$), and Informativeness ($\mathcal{I}$) (DCI) metric [5]. A high value for each of these metrics depicts a high semantic meaning correlated with the tangible features in the dataset [23, 27]. Considering $\mathbb{F}$ as the total number of factors of variation in the dataset, we trained $\mathbb{F}$ gradient boosting regressors for each representation in the dataset as the feature set and generated an importance matrix R such that for a given latent factor j, $R_{i,j}$ represents the ith feature importance weight of the linear regressor trained on the set of representations with jth factor of variation in the output.

Disentanglement ($\mathcal{D}$). The disentanglement score of the metric represents the degree to which a given representation disentangles the underlying factors of variation. The total disentanglement score for the factor of variations was calculated as follows:

$$\mathcal{D} = \sum_i \Big(1 - H(P_i)\Big) \frac{\sum_j R_{i,j}}{\sum_{i,j} R_{i,j}}, \tag{5}$$

where $H(P_i)$ represents the entropy of the P_i distribution where, P_i is a $j \times 1$ vector such that $P_{i,j} = \frac{R_{i,j}}{\sum_k R_{i,k}}$. The score directly represents disentanglement as it is equal to 1 only when each representation is deemed important for predicting only 1 out of the different factors of variations.

Completeness ($\mathcal{C}$). The completeness score measures the degree to which a single factor of variation j is captured by the representations and is calculated as follows:

$$\mathcal{C}_j = 1 - H(P_j), \tag{6}$$

where $H(P_j)$ is calculated in the same manner as described in the previous section. The score C_j is equal to 1 if only one representation is important for predicting the jth factor of variation and is equal to 0 if all representations contribute equally. The final score is calculated as follows for all the $\mathbb{F}$ factors of variations:

$$\mathcal{C} = \sum_{j=1}^{\mathbb{F}} C_j. \tag{7}$$

Informativeness ($\mathcal{I}$). The information score represents the degree of information captured by a representation for all the underlying factors of variations and is calculated as follows:

$$\mathcal{I} = \mathbb{E}_{j \in (\mathbb{Z} \cap [1 \ldots \mathbb{F}])} \Big[1 - \|z_j - z'_j\|\Big], \tag{8}$$

where z_j represents the true distribution of the j^{th} factor of variation and z'_j represents the distribution predicted by the j^{th} linear regressor. Ignoring the dependence of this metric on the capacity of the used regressors, this metric is equal to 1 when the representations are perfectly able to predict all the factors of variations.

Table 2 illustrates that the `FaIRCoP` embeddings outperform the embeddings of the pretrained $\texttt{ResNet} - 18$ model across all three metrics for both the datasets.

6.2. Fairness

Since we do not release the criminal dataset to maintain confidentiality, we illustrate an extensive fairness study on the dataset to provide an idea about the label distribution in the dataset. To evaluate the efficacy of the representation generators in terms of fairness, we evaluated them based on two metrics – *group fairness* [12] and *label distribution similarity*. We used a modified group demographic parity metric [12] and depict the results that we obtained in Figure 5. For group fairness, we used a custom demographic parity based measure, where, a low metric value (which is essentially an average of the pairwise differences for each pair in the sample set spanned by the joint distribution of t and s) indicates equal characterization for each minority group (represented by s) with respect to each target group (represented by t). The steps involve dividing the generated representations into training and testing sets and generating each possible pair of factor of variations as sensitive variable s and target variable t consecutively. For each pair, a K-Nearest Neighbors (KNN) clustering model was fit onto the training set with t as the output and the conditional probability $p(t_i|s_j)$ for all $t_i \in t$ and $s_j \in s$ was calculated to get the final heatmap. To summarize the heatmap in a single quantitative metric, we evaluated the average value using the following formulation.

$$\mathcal{F} = \mathbb{E}_{[t_i, s_j] \in [t,s]} \big[p(t_i|s_j)\big]. \tag{9}$$

The results can be found in Fig. 5, where, we evaluated FaIRCoP embeddings against a pretrained $\texttt{ResNet} - 18$ on our entire training set. We also evaluated the distribution similarity between the dissimilar images selected by the simulator with the entire training set to evaluated any biases in the clustering procedure. The results for distribution similarity can be found in the supplementary material.

7. Conclusion

In this work, we tackled the problem of suspect identification using contrastive learning over user feedback which serves as a weaker form of supervision for the system. For this, we propose the $SCLoss$, along with an online inference strategy. Our system caters to the personalized notion of features that each user has due to high subjectivity in the mental visual memory of a witness. Equipped with a user-friendly web interface, our proposed algorithm outperforms other state-of-the-art baselines qualitatively and quantitatively, as verified through the user study.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[2] Binod Bhattarai, Gaurav Sharma, and Frederic Jurie. Cp-mtml: Coupled projection multi-task metric learning for large scale face retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1597–1607. PMLR, 13–18 Jul 2020.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR 2005*, pages 886–893, 2005.

[5] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *ICLR*, 2018.

[6] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Schmidt Feris. Dialog-based interactive image retrieval. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 676–686, Red Hook, NY, USA, 2018. Curran Associates Inc.

[7] Qiyang Hu, Attila Szabó, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Disentangling factors of variation by mixing them. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3399–3407, 2018.

[8] Aayush Jain, Meet Shah, Suraj Pandey, Mansi Agarwal, Rajiv Ratn Shah, and Yifang Yin. Seeksuspect. *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, 2021.

[9] Young Kyun Jang and Nam Ik Cho. Similarity guided deep face image retrieval. 2021.

[10] Ananya Harsh Jha, Saket Anand, Maneesh Kumar Singh, and V. S. Rao Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. *ArXiv*, abs/1804.10469, 2018.

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[12] Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *NeurIPS*, 2019.

[13] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[14] Harsh Shrivastava, S RamaKrishnaP.V.N., Karmanya Aggarwal, Meghna P. Ayyar, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. Facefetch: An efficient and scalable face retrieval system that uses your visual memory. *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 338–347, 2019.

[15] Ankita Shukla, Sarthak Bhagat, Shagun Uppal, Saket Anand, and Pavan K. Turaga. Prose: Product of orthogonal spheres parameterization for disentangled representation learning. *ArXiv*, abs/1907.09554, 2019.

[16] Ankita Shukla, Shagun Uppal, Sarthak Bhagat, Saket Anand, and Pavan Turaga. Geometry of deep generative models for disentangled representations. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP 2018, New York, NY, USA, 2018. Association for Computing Machinery.

[17] Indah Agustien Siradjuddin, Aryandi Triyanto, and Mochammad Kautsar S. Content based image retrieval with rocchio algorithm for relevance feedback using 2d image feature representation. In *Proceedings of the 2019 2nd International Conference on Machine Learning and Machine Intelligence*, MLMI 2019, page 16–20, New York, NY, USA, 2019. Association for Computing Machinery.

[18] Brandon M. Smith, Shengqi Zhu, and Li Zhang. Face image retrieval by shape manipulation. In *CVPR*, 2011.

[19] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, 2016.

[20] Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. *ArXiv*, abs/1711.02245, 2018.

[21] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[22] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[23] Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *NeurIPS*, 2019.

[24] Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, 2001.

[25] Xinru Yang, Haozhi Qi, Mingyang Li, and Alexander G. Hauptmann. From A glance to "gotcha": Interactive facial image retrieval with progressive relevance feedback. 2020.

[26] Alireza Zaeemzadeh, Shabnam Ghadar, Baldo Faieta, Zhe Lin, Nazanin Rahnavard, Mubarak Shah, and Ratheesh Kalarot. Face image retrieval with attribute manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12116–12125, October 2021.

[27] Julian Zaidi, Jonathan Boilard, Ghyslain Gagnon, and Marc-André Carbonneau. Measuring disentanglement: A review of metrics. *ArXiv*, abs/2012.09276, 2020.