A-LINK: Recognizing Disguised Faces via Active Learning based Inter-Domain Knowledge

Anshuman Suri, Mayank Vatsa, Richa Singh
IIIT-Delhi
{anshuman14021, mayank, rsingh}@iiitd.ac.in

Abstract

Recent advancements in deep learning have significantly increased the capabilities of face recognition. However, face recognition in an unconstrained environment is still an active research challenge. Covariates such as pose and low resolution have received significant attention, but "disguise" is considered an onerous covariate of face recognition. One primary reason for this is the unavailability of large and representative databases. To address the problem of recognizing disguised faces, we propose an active learning framework A-LINK, that intelligently selects training samples from the target domain data, such that the decision boundary does not overfit to a particular set of variations, and better generalizes to encode variability. The framework further applies domain adaptation with the actively selected training samples to fine-tune the network. We demonstrate the effectiveness of the proposed framework on DFW and Multi-PIE datasets with state-of-the-art models such as LC-SSE and DenseNet.*

1. Introduction

State-of-the-art face recognition models have demonstrated high performance on datasets such as CMU MultiPIE [5] and Labeled Faces in the Wild [8]. These models face challenges due to several covariates such as disguise and low resolution. For example, a person may get a photo clicked while wearing sunglasses, a wig or with makeup. Ideally, such disguised appearances should not confuse face-recognition models. At the same time, as shown in Figure 1, an impostor wearing a disguise to impersonate another user should not be able to fool an ideal recognition model. The covariate of face disguise is not uncommon; Dhamecha *et al*. explore the feasibility of face verification under disguise variations using multi-spectrum face images [2]. This covariate has significant implications for real-world face recognition systems used by government agencies, online social media networks and surveillance systems.

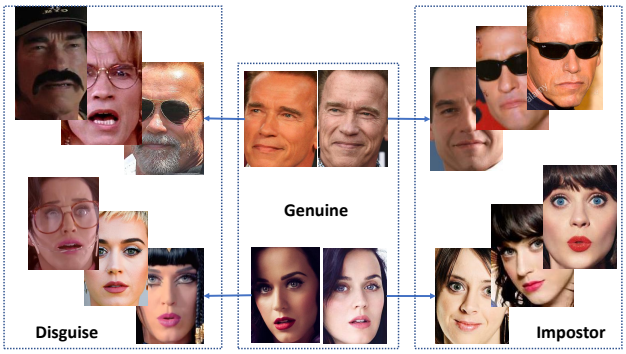

Figure 1: Subject image samples from DFW [3], [23] database, along with their corresponding impostor and disguised images.

Matching faces with disguised variations can be modelled as a domain adaptation problem; here, the source domain (source data distribution) contains undisguised faces, whereas the target domain (different, but related to source domain) may include disguised images. Recent work by Kan *et al*. [10] uses such an approach to work with unlabeled data in the target domain. However, their work cannot be used on top of already trained models, or convolutional neural networks/deep learning models. Yao *et al*. utilize a similar approach: they consider low resolution and high resolution as two distinct domains and propose a projection technique that utilizes domain adaptation [30]. The scarcity of labelled data from the target domain makes this an even harder challenge. Models can assign pseudo-labels, but the level of disguise in most cases is too much for any model trained on the source domain to work well.

Active learning, a technique that intelligently selects examples while training or queries the annotator for labels, can be useful in cases where data is scarce. Several techniques have been explored in the literature for active learning. Query by committee [13] selects points for which an ensemble of models disagrees the most. Uncertainty sam-

*https://github.com/iamgroot42/A-LINK

978-1-7281-1522-1/19/$31.00

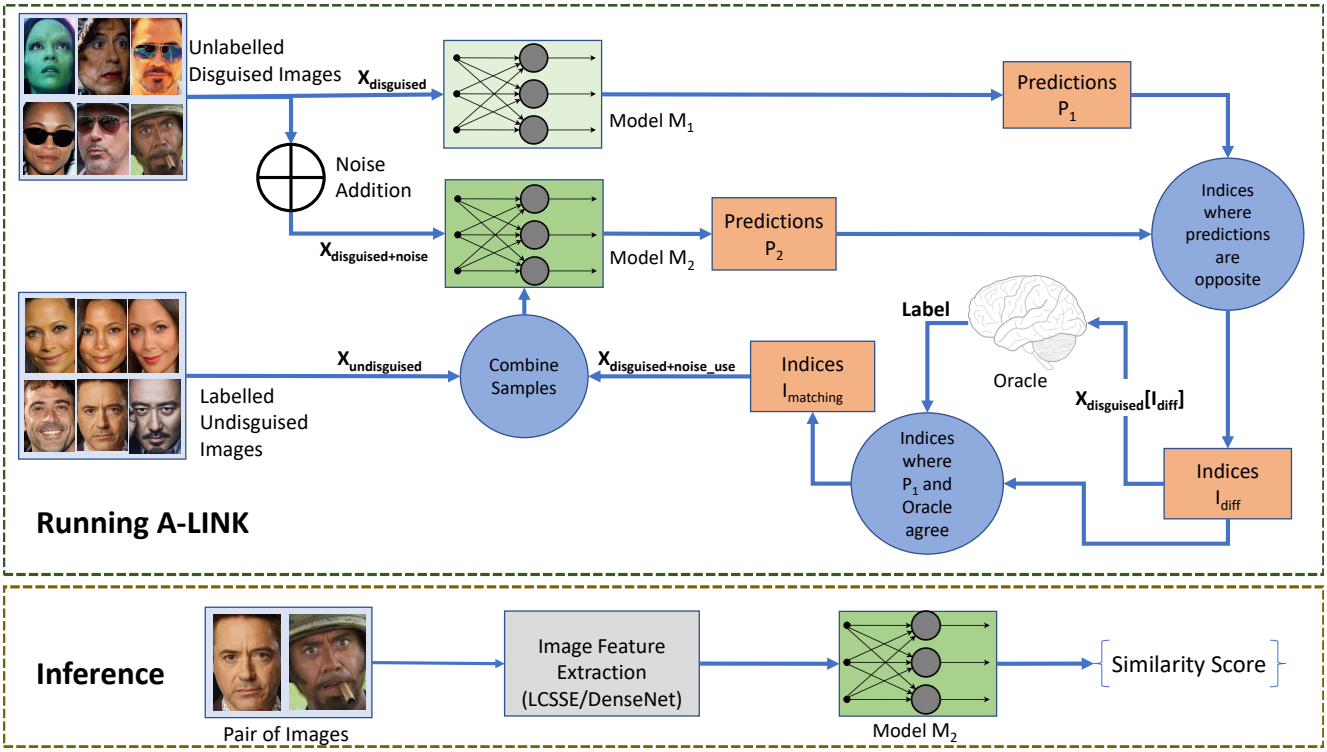

Figure 2: A data-flow diagram of A-LINK for the case of disguised faces. A-LINK starts with a batch of unlabelled data from the target domain and ends up with a batch of data from both source and target domain, on which M_2 is finetuned.

pling [12] selects points for which a given model is least confident about its predictions. Recent advancements extend existing methods in active learning to deep neural networks. Work on estimating distances of points from decision boundaries using adversarial perturbations [4] is inspired by work that queries points in an SVM [27]. Lin *et al*. create a teacher-student framework to mix predictions from a teacher model with annotated labels to fine-tune a given student CNN model [14]. Recent work on partial feedback [6] has shown significant improvements for deep-neural networks. Huang *et al*. describe a novel query strategy for active learning on surveillance images [9]. Tarvainen *et al*. propose a model-weight based ensemble technique which trains student and teacher models together, yielding near state-of-the-art performance with a fraction of labelled data [26]. Dictionary learning has also been explored in literature: Suri *et al*. propose a method that combines task-dependent and task-independent features [25]. All of the above work considers data to be from the same domain, and thus may not work well in the presence of varying covariates.

Combining active learning with domain adaptation can, potentially, lead to a system that works better than either of the techniques applied individually. These two techniques have been combined in the form of active-supervised domain adaptation [18] for improved performance. This algorithm trains an additional classifier while segregating examples into the source or target domain. However, such a clear distinction may not always be available in an instance such as facial disguises: the covariates may not be distinctly separated across the two domains. The need to train such a classifier also makes the overall framework time-consuming.

Inspired by the recent success of domain adaptation and active learning, we propose A-LINK (***A**ctive-**L**earning based **I**nter-domai**N** **K**nowledge*): a framework that utilizes active learning to adapt to a given target domain with a covariate, in specific, faces with disguises. The proposed framework does not require re-training the model from scratch, and therefore, it is fast. Further, the model does not require any pre/post processing at inference. A-LINK is model agnostic; we evaluate the effectiveness of the proposed framework on multiple state-of-the-art deep learning models including DenseNet [7] and Local Class Sparsity Supervised Autoencoder (L-CSSE) [15] to ascertain our claim. Additionally, we evaluate the generalizability of the proposed algorithm on another covariate: Multi-PIE [5] dataset, treating low-resolution as the target domain.

2. Proposed A-LINK Algorithm

In this research, we propose Active-Learning based Inter-domain Knowledge (A-LINK) algorithm which allows a model access to an intelligently crafted subset of im-

ages from the target domain. In effect, the chosen subset of points would allow the best possible transfer of weights from the source to the target domain. The algorithm uses two models M_1 and M_2: M_1 plays the role of a teacher model which has been trained well on the source domain, whereas M_2 plays the role of a student model which has been trained on a relatively smaller set of data from the target domain. Data points from the target domain for which M_2 yields wildly fluctuating scores and M_1 does not are filtered and used for fine-tuning M_2. The hypothesis behind selecting these particular points is that if M_2 performs well on them, it should perform well on the entire target domain. Figure 2 illustrates the concept of A-LINK.

The algorithm assumes a pair of deep architectures/models: M_1 and M_2. Here, M_1 plays the role of a **teacher model**: having knowledge of the source domain. The model M_2 plays the role of a **student model**, with some information about the target domain. The ultimate goal of the algorithm is to use M_1's knowledge to refine M_2's performance on the target domain in a semi-supervised manner while maintaining performance on the source domain. We also assume the presence of an *Oracle*, an annotator which yields the corresponding ground truth for given inputs. This entity is found in most active learning settings; its use in this particular problem is elaborated upon in Section 2.3.

The presence of two models helps achieve an ensemble-like effect, transferring knowledge from M_1 to M_2. The presence of an Oracle ensures that the data used to fine-tune M_2 does not have incorrect or noisy labels, thus ensuring there are no outliers. These four sub-protocols are described in detail in the following sections.

2.1. Generating Predictions

The main objective is to fine-tune M_2 with limited amounts of training data, by utilizing domain information available from M_1. Since annotating data is an expensive process, we consider a setting where a sufficient amount of unlabeled data are available as our starting premise.

We sample a batch of **unlabeled** target-domain image-pairs (called X_{target}) and pass it to M_1 to get a set of predictions, P_1. These predictions vary in $[0, 1]$: 0 implies that the images in a pair are of distinct people, while 1 corresponds a perfect match. Next, we take a copy of this batch and add some noise to each image in an image-pair, for all pairs. There is no restriction on the type of noise which can be incorporated into an image; we have experimented with Gaussian, Salt and Pepper, Perlin, and Poisson noise; as well as combinations of these four. This set of now-noisy samples (called $\mathrm{X}_{target+noise}$) is fed as input to M_2 and the set of predictions obtained are taken to be P_2.

The intuition here for adding noise is: a model that has overfitted on the given data distribution or is not confident enough about its predictions is most likely to yield incorrect predictions for such perturbed inputs. However, if the model predicts a score that is close to the actual label for a pair of images perturbed with noise, it is expected to perform well for unperturbed images as well. Thus, adding noise to images and considering only such cases helps identify data points which, when used to fine-tune M_2, can potentially improve its performance.

Algorithm 1: A-LINK

Input: mix_ratio, ϵ

Train model M_1 on image-pairs without the covariate (source domain);
Train model M_2 on the limited-size set of image-pairs with the covariate (target domain);
while X_{target} *contains data* **do**
  Get next batch of unlabeled image-pairs with covariate (from X_{target});
  Get predictions P_1 for X_{target} from M_1;
  Add noise to a copy of X_{target} to get batch $X_{target+noise}$;
  Pass $X_{target+noise}$ to M_2, get predictions P_2;
  $I_{diff} = \{i \mid (P_1[i] \geq 0.5) \neq (P_2[i] \geq 0.5)\}$;
  $I_{diff} = \{i \in I_{diff} \mid i \geq 0.5 + \epsilon \text{ or } i \leq 0.5 - \epsilon\}$;
  Query Oracle for $X_{target}[I_{diff}]$ to get $Label$;
  $I_{matching} = \{i \in I_{diff} \mid Label[i] == P_1[i]\}$;
  $X_{target+noise\_use} = X_{target+noise}[I_{matching}]$;
  Get X_{source} = $mix\_ratio - 1$ more batches of source-domain image-pairs;
  Fine-tune M_2 with $concatenate(X_{source}, X_{target+noise\_use})$ and their corresponding labels;
end

2.2. Data Filtering

Once we have P_1 and P_2, we compute the set of indices I_{diff}; as the indices where P_1 and P_2 differ in predictions. By difference in prediction (disagreement), we mean that if an index i $\in \mathrm{I}_{diff}$,

1. P_1[i] $\in$ [0, 0.5) and P_2[i] $\in$ [0.5, 1] or,
2. P_1[i] $\in$ [0.5, 1] and P_2[i] $\in$ [0, 0.5)

Similarly, P_1 and P_2 agree on some prediction, or some index j $\notin \mathrm{I}_{diff}$ if,

1. P_1[j] $\in$ [0, 0.5) and P_2[j] $\in$ [0, 0.5) or,
2. P_1[j] $\in$ [0.5, 1] and P_2[j] $\in$ [0.5, 1]

The intuition here is that data points, for which P_1 and P_2 agree, would not provide any additional information required to fine-tune M_2. Thus, using such data-points would not be very beneficial in improving M_2's performance.

To further reduce the number of queries made to the oracle, we further filter I_{diff} in the following way: cases for

which the predictions from M_1 lie in [0.5-ϵ, 0.5+ϵ] are discarded. These cases most likely correspond to the scenario where M_1 is not confident enough about its predictions and thus have a higher probability of being incorrect. This way, we end up making fewer queries to the oracle.

2.3. Using the Oracle

After the first filtering step, we utilize the oracle to further prune data points. The Oracle provides the ground truth for any input it is supplied with. In our case, the oracle ascertains whether a pair of images belong to the same class/identity or not. Once we compute I_{diff}, we take the corresponding image-pairs, *i.e.*, $\mathrm{X}_{target}[\mathrm{I}_{diff}]$ and query for the ground truth corresponding to each pair from the oracle. In the same way, as explained in Section 2.2, we choose points where the oracle is in sync with M_1's predictions to generate a new set of indices, namely $\mathrm{I}_{matching}$.

Since queries to the oracle are expensive, we also intend to minimize their number. This entire segment is the essential active-learning portion of the proposed algorithm since we do not query the oracle for all possible image-pair ground truth values.

2.4. Fine-tuning M_2

After computing $I_{matching}$ from the previous step, image pairs from X_{target} corresponding to these indices are selected to be used by M_2. These image-pairs correspond to the cases where M_1 is right about its predictions but M_2 is wrong about their labels when a perturbed version of the same data-points is used.

Since this batch consists of only data from the target domain (with covariate), it is possible that, over time, M_2 may start to overfit to these image-pairs. To circumvent such a possibility, some data from the source domain (with associated labels) is added to the batch as well. In our implementation, we add $mix\_ratio - 1$ more batches of data-points from the source domain to prepare one final batch on which M_2 is fine-tuned. This ensures that M_2 has good performance both on the target domain and the source domain.

3. Datasets

3.1. Disguised Faces in the Wild (DFW)

The Disguised Faces in the Wild (DFW) dataset [11, 23] dataset contains more than 11,155 images of 1000 subjects with different kinds of disguise variations. As per the predefined protocol, 400 subjects comprise the training set, and 600 subjects comprise the testing set. Face coordinates for these images are included in the dataset and were generated using Faster RCNN [17]. Using these coordinates, the face region is extracted from each image. The dataset contains images with their person-identifiers. However, for training the predefined protocol, we are interested in a format where

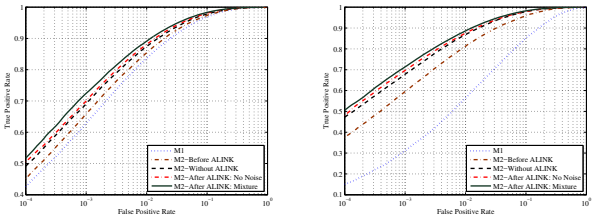

Figure 3: ROC curves on the test-set of DFW for the overall scenario with Densenet (left) and L-CSSE (right)

a pair of images, along with an indicator of them being the same, is made available. Thus, we construct image-pairs by combining inter-class and intra-class images (all possible combinations) for use in the algorithm. We maintain subject exclusivity between the training and test set while creating these image-pairs. All further references to *data point* refer to these image-pairs, not individual images from the original structure of the dataset. The three cases considered for evaluation are:

1. **Impersonation**: considering genuine validation (595 cases) vs. impostor impersonators (24,451 cases).
2. **Obfuscation**: considering genuine, disguised (13,302 cases) vs. cross-subject impostors (9,027,981 cases).
3. **Overall**: considering genuine (disguised and undisguised both; 13,897 cases) vs. impostors (impersonators and cross-subject; 9,052,432 cases).

3.2. CMU Multi-PIE Dataset

The CMU Multi-PIE dataset [5] contains images of 337 subjects. We follow a setup similar to that used by Singh *et al*. [22]: out of all images, we select images with frontal pose, uniform illumination, and neutral expression. For running the proposed algorithm, 100 subjects are randomly chosen for training while the remaining 237 compose the test set. This step is repeated five times, and the results are averaged across these runs. Through this dataset, we wish to showcase the generalization of our algorithm to covariates other than disguise. We artificially downsize the images from their original size of 150×150, considering the original dataset as our source domain, and the distribution of downsized images as our target domain. We experiment with multiple resolutions for the target domain, referred to as *target resolution* throughout the paper, for the downsized images: 16×16, 24×24, 32×32, and 48×48.

4. Implementation Details

The algorithmic description of A-LINK is shown in Algorithm 1. For feature extraction, we use a Densenet [7] model for feature extraction, trained on Labeled Faces in the Wild Dataset (LFW) [8] (which is a standard dataset used in the literature for training face-verification models). To assess the generalisation of our approach to an-

other feature extraction model, experiments are also shown using a Local Class Sparsity Supervised Autoencoder (L-CSSE) [15] (Section 5.1). This L-CSSE Autoencoder is used with a Siamese network built on top of it with three fully-connected layers: the absolute difference in feature vectors is passed as input to the fully-connected layers. L-CSSE autoencoder is trained on LFW before being used as a featurization model for A-LINK. All feature extraction layers are frozen while training these Siamese networks.

A machine with an Nvidia K-80 GPU, with 128GB RAM and a Xeon E5-2630v2 CPU is used to perform all the experiments. For both the datasets, we varied Disparity-ratio in the range $\{0.25, 0.5, 1, 2, 4\}$. Additionally, we also varied the kind of noise added: {Gaussian, Poisson, Speckle, Salt & Pepper, Perlin}, along with a combination of all of these noises. An oracle is simulated artificially by accessing ground-truth for labelled data from the target domain, which is otherwise not used by the algorithm, and keeping track of the number of such unique accesses.

DFW: Architectures of models M_1 and M_2 are the same: it consists of an absolute difference layer over the two inputs (features extracted from images), followed by two layers with ReLU [16] activation of 512 and 64 neurons. These layers are followed by a one-neuron layer and Sigmoid activation, thus predicting a score in the range [0,1]. The AdaDelta optimizer with its default learning rate of 1 was used while training all the models, with a batch size of 64. M_1 is trained using *labeled, undisguised* face-image pairs, while M_2 is trained using 50% of the *labelled, disguised* face-image pairs available. We use the remaining 50% disguised face-image pairs for A-LINK. All cropped face-images are resized to 224×224.

Multi-PIE: The architecture for M_1 is the same as described above. For the source domain, we resize all images from their original 150×150 resolution to 224×224 resolution (linear interpolation) to pass via the featurization model. We do not use a pre-trained featurization model for M_2: up-scaling an image from a resolution like 16×16 to 224×224 is bound to be lossy. Thus, instead of using an explicitly separate featurization model, we modify M_2 to take as inputs raw images: two 3×3 convolutional layers with 32 filters, followed by 2×2 max-pooling, followed by a similar block with 64 filters. The rest of the architecture is the same as M_1. The optimizer, learning rate, and other hyper-parameters are kept the same as in the case of DFW. M_1 is trained using labeled face-image pairs, while M_2 is trained using only 50% of the labelled images with *target resolution* (48×48, 32×32, etc). Similar to DFW, we use the remaining 50% target-resolution image-pairs for A-LINK. Unlike DFW, there are no separate datasets for source-domain and target-domains: the resolution covariate is defined independently of the dataset. Thus, for pre-training M_1 and M_2, we use the same data: M_1 sees the original image-pairs, while M_2 sees downsized image-pairs.

We used Tensorflow (https://www.tensorflow.org/) and Keras (https://keras.io/) for implementing the algorithm and training models.

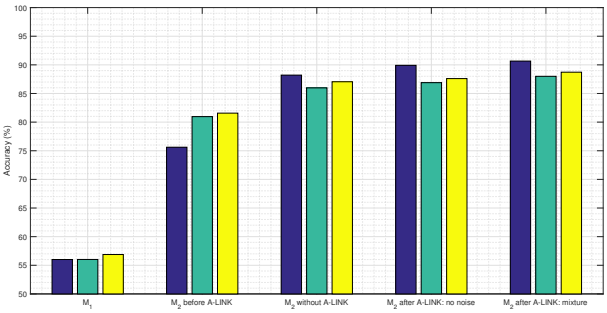

Figure 4: GAR at 1% FAR for impersonation, obfuscation and overall performance on DFW, for M_1 and variations of M_2. M_1, M_2 use L-CSSE for feature extraction.

5. Experimental Results and Analysis

Experiments are performed for all possible configurations, using the parameters specified in Algorithm 1. The configuration of parameters that performed best for us is: ϵ=0.5, disparity-ratio=none, mix_ratio: 2, 50% of labelled disguised-face data used in Step 2, noise used: a mixture of Gaussian, salt-pepper, Poisson, speckle.

Receiver operating characteristic (ROC) curves pertaining to the overall cases for DFW are shown in Figure 3, using DenseNet and L-CSSE. The curves belong to the best configurations in the overall cases. Genuine accept rates (GAR) at 1% false accept rate (FAR) and 0.1% FAR for DFW are given in Table 1.

In addition to the models trained with A-LINK, we also include results for model M_1 and M_2 (M_2 **before A-LINK**) after the step where it has been trained on a limited-size set of disguised face-image-pairs (Line 2, Algorithm 1). Varying this percentage to $\{30, 40, 50\}$% of data yields a GAR (at 1% FAR) of 88.35 ± 0.32%; although using 50% of the available data gives the best results (88.72%), varying this ratio does not alter the results significantly. Further, we also performed experiments with the variation:

$$I_{diff} = argsort(- \mid P_1 - P_2 \mid)[: sample\_size] \quad (1)$$

where *sample_size* is a specified percentage (disparity-ratio) of the size of $\mid P_1 - P_2 \mid$.

For the best configuration of A-LINK (using a model with the same architecture trained on all data without A-LINK, as benchmark) that uses a DenseNet model, absolute improvements of 4.35%, 1.98%, and 1.74% are observed in GAR at 1%FAR for the cases of impersonation, obfuscation, and overall case of DFW. For GAR at 0.1% FAR, absolute improvements are 1.65%, 3.18% and 3.19% for

Table 1: GAR at 1% FAR and 0.1% FAR for impersonation, obfuscation and overall performance of DFW, for M_1, M_2 before A-LINK, M_2 after A-LINK using no noise, M_2 after A-LINK using mixture of noise, and M_2 without A-LINK.

Model	$GAR_{1\%}$	$GAR_{0.1\%}$
Impersonation		
Baseline (VGG-Face)	52.77	27.05
Baseline (VGG-Face2)	73.94	38.48
AEFRL [24]	<u>96.80</u>	<u>57.64</u>
MiRA-Face [11]	95.46	51.09
UMDNets [1]	94.28	53.27
M_1 (DenseNet)	89.68	65.60
Uncertainty Sampling [12]	89.71	65.78
Margin Sampling [19]	91.09	73.12
Entropy Sampling [21]	91.27	73.53
M_2 before A-LINK	89.15	69.41
M_2 without A-LINK	91.38	71.93
M_2 after A-LINK: no noise	92.84	73.27
M_2 after A-LINK: mixture	**95.73**	**75.38**
Obfuscation		
Baseline (VGG-Face)	31.52	15.72
Baseline (VGG-Face2)	54.86	31.55
MiRA-Face [11]	<u>90.65</u>	<u>80.56</u>
AEFRL [24]	87.82	77.06
UMDNets [1]	86.62	74.69
M_1 (DenseNet)	83.11	63.01
Uncertainty Sampling [12]	83.44	63.28
Margin Sampling [19]	85.01	68.92
Entropy Sampling [21]	85.07	68.99
M_2 before A-LINK	84.23	65.15
M_2 without A-LINK	86.99	68.95
M_2 after A-LINK: no noise	87.52	69.28
M_2 after A-LINK: mixture	**88.97**	**72.13**
Overall		
Baseline (VGG-Face)	33.76	17.73
Baseline (VGG-Face2)	56.22	32.68
MiRA-Face [11]	<u>90.62</u>	<u>79.26</u>
AEFRL [24]	87.90	75.54
UMDNets [1]	86.75	72.90
M_1 (DenseNet)	83.74	63.18
Uncertainty Sampling [12]	83.89	63.71
Margin Sampling [19]	85.50	65.97
Entropy Sampling [21]	86.08	69.04
M_2 before A-LINK	85.41	65.99
M_2 without A-LINK	87.56	69.53
M_2 after A-LINK: no noise	88.14	70.15
M_2 after A-LINK: mixture	**89.30**	**72.72**

the cases of impersonation, obfuscation and overall case of DFW.

For comparison, we have also included results from the DFW competition organized with CVPR 2018. According to the competition, MiRA-Face [11], AEFRL [24] and UMDNets [1] are the current state-of-the-art for this dataset. It can be observed that in addition to outperforming base models, our models trained with A-LINK perform comparably to the current state-of-the-art (Table 1).

Some of the steps in Algorithm 1 can be replaced with other variations, thus making the approach generic to the presence of any covariates in the data. A few of them that are explored in experiments are:

- We performed experiments with model agnostic noises (Step 6, Algorithm 1) including Gaussian noise, Salt and Pepper noise, Speckle noise, Poisson noise, Perlin noise, and a mixture of these. A mixture of these noises was observed to perform the best on the DFW dataset. It yields GAR of 87.05% and 88.01% (at 1% FAR) for the Overall case with no noise and a mixture of Gaussian noise, Salt and Pepper noise, and Speckle noise, respectively.

- Steps 8 and 11 (Algorithm 1) check for equality by considering the two outputs as part of a binary classification problem. This criterion can be replaced with another rule (Equation 1); in Equation 1, $sample\_size$ is a hyper-parameter and can be set as a specific percentage of $\mid P_1 - P_2 \mid$ (**disparity-ratio**). Varying this ratio in $\{$12.5%, 25%, 50%, normal criteria (defined in Section 2.2)$\}$ yields a GAR (at 1% FAR) of $88.24 \pm 0.32\%$.

- We tried varying the value of ϵ in Step 9 (Algorithm 1). A higher value of ϵ corresponds to fewer queries to the oracle, but at the same time results in lesser data for fine-tuning M_2. Varying ϵ in $\{0, 0.05, 0.10\}$ yields a GAR (at 1% FAR) of $88.39 \pm 0.29\%$, with $\epsilon = 0.05$ giving the best results.

- We experiment with existing active-learning techniques. **Uncertainty Sampling** selects data instances based on usefulness; 1 - maximum posterior probability (across classes). Yang *et al*. utilize this in their proposed algorithm for diversity maximization [29]. **Margin Sampling**: selects data instances based on the difference between the highest and second-highest posterior probabilities (across classes). Wu *et al*. propose a weighted sampling method for deep embeddings [28]. **Entropy Sampling**: selects data instances based on entropy; Shanon entropy over all classes [20]. We ran experiments varying the upper cap on the number of queries that were made by the active learning algorithm. Across all ratios, our algorithm outperforms all of these three active learning techniques, for both datasets (Tables 1, 2).

We used modAL (https://modal-python.github.io/) for implementing these active-learning algorithms.

Table 2: Rank-1 classification accuracies for target resolutions: 16×16, 24×24, 32×32, 48×48, for Multi-PIE.

Resolution	Singh *et al*. [22]	M_1 (Dense--Net)	Uncertainty Sampling	Margin Sampling	Entropy Sampling	Proposed A-LINK
16×16	91.1	89.9	80.5	81.1	81.9	**92.4**
24×24	91.8	90.1	86.2	86.7	86.8	**92.6**
32×32	91.9	90.2	87.5	87.9	88.0	**92.8**
48×48	91.5	90.2	89.4	89.5	89.5	**92.9**

5.1. Generalization Over CNN Models

As a proof-of-concept of the proposed approach and the boost in performance it yields, we have also conducted experiments using an L-CSSE model [15] for feature extraction for DFW. The same boost in performance is observed in this model as well (Figure 4). Absolute improvements of 2.45%, 2.02%, and 1.67% are observed in GAR at 1% FAR for the cases of impersonation, obfuscation and overall case of DFW, with and without using A-LINK, respectively. For GAR at 0.1% FAR, absolute improvements are 3.12%, 3.74% and 3.57% for the cases of impersonation, obfuscation and overall case of DFW, with and without using A-LINK, respectively. The core of this algorithm lies in the performance gain it yields: for a model that has mediocre performance, A-LINK boosts its performance significantly and for a model which already performs well, A-LINK helps further boost its performance.

5.2. Analysis

It is observed that A-LINK required 30-35% less labelled disguised face images while training the algorithm. For all three cases of impersonation, obfuscation, and overall performance of DFW, a proportionate mixture of Gaussian, Speckle, Perlin and Salt-Pepper noise outperforms individual variations for the noises it uses. To assess the importance of noise addition in A-LINK, we also compare the proposed model with the variation that does not add any noise (M_2 **after A-LINK: no noise**). As expected, the variation of A-LINK with no noise is outperformed by the one with multiple types of noise. Tuning $mixture\_ratio$ significantly alters performance: a 1:1 ratio of noisy disguised images and clean undisguised images work best. Increasing the ratio of noisy images tends to make M_2 overfit on noisy images, whereas increasing the proportion of unperturbed images tends to dilute the effect of images with added noise.

To study the effect of actively selecting samples while fine-tuning M_2, we performed experiments on the variation in which M_2 is trained without running A-LINK; using all available labelled disguised-faces data (M_2 **without A-LINK**). This variation outperforms the version of M_2 trained on only partially available labelled samples and not fine-tuned by A-LINK. However, the proposed model outperforms this variation in all cases by a significant margin.

For an estimate of the performance boost M_2 gains after executing A-LINK, we also report metrics for M_2 before the algorithm is run (M_2 before A-LINK). As visible in ROC curves (Figure 3), all of the models get a significant increase in their performance.

5.3. Performance on Multi-PIE Database

To further reinforce the boost in performance provided by A-LINK, we observe results on the Multi-PIE dataset as well. Similar to the evaluation setup described by Singh *et al*. [22], we report Rank-1 identification accuracies for several target resolutions: 16×16, 24×24, 32×32, 48×48. While Singh *et al*. focus on synthesizing high-resolution images, the proposed algorithm focuses on the feature extraction level. It can be observed that in addition to outperforming base models, our models trained with A-LINK outperforms current state-of-the-art (Table 2) on Multi-PIE.

6. Conclusion

The proposed A-LINK algorithm combines concepts from active learning and domain adaptation to successfully boost the performance of a given model. Further, A-LINK is faster than training a model on all the data points, has low auxiliary storage requirements, and reduces the number of labelled examples required significantly, without compromising on the model's performance. Experimental results show that A-LINK leads to significant improvements while fine-tuning a model, for all three cases of the DFW dataset; impersonation, obfuscation and overall. It even comes close to outperforming the current state-of-the-art. Our algorithmic framework shows good generalization: generalizing across featurization models (L-CSSE, Densenet) as well as different covariates: disguise in DFW, low resolution in Multi-PIE. Since A-LINK is generic, it can be used to incorporate more sophisticated active-learning criteria, along with variations of noise, while fine-tuning the model under consideration.

7. Acknowledgements

M. Vatsa and R. Singh are partly supported by the Infosys Center for AI at IIIT Delhi. M. Vatsa is also supported through the Swarnajayanti Fellowship by Government of India.

References

[1] A. Bansal, R. Ranjan, C. D. Castillo, and R. Chellappa. Deep features for recognizing disguised faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–16, 2018.

[2] T. I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa. Disguise detection and face recognition in visible and thermal spectrums. In *2013 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2013.

[3] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar. Recognizing disguised faces: Human and machine evaluation. *PLOS ONE*, 9(7):1–16, 07 2014.

[4] M. Ducoffe and F. Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[6] P. Hu, Z. C. Lipton, A. Anandkumar, and D. Ramanan. Active learning with partial feedback. *arXiv preprint arXiv:1802.07427*, 2018.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[8] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[9] Y. Huang, Z. Liu, M. Jiang, X. Yu, and X. Ding. Cost-effective vehicle type recognition in surveillance images with deep active learning and web data. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–8, 2019.

[10] M. Kan, J. Wu, S. Shan, and X. Chen. Domain adaptation for face recognition: Targetize source domain bridged by common subspace. *International Journal of Computer Vision*, 109(1-2):94–109, 2014.

[11] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa. Disguised faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2018.

[12] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[13] R. Liere and P. Tadepalli. Active learning with committees for text categorization. In *AAAI/IAAI*, pages 591–596, 1997.

[14] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang. Active self-paced learning for cost-effective and progressive face identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):7–19, 2018.

[15] A. Majumdar, R. Singh, and M. Vatsa. Face verification via class sparsity based supervised encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1273–1280, 2017.

[16] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing systems*, pages 91–99, 2015.

[18] A. Saha, P. Rai, H. Daumé, S. Venkatasubramanian, and S. L. DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112, 2011.

[19] T. Scheffer, C. Decomain, and S. Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[20] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[21] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[22] M. Singh, S. Nagpal, M. Vatsa, R. Singh, and A. Majumdar. Identity aware synthesis for cross resolution face recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–59209, June 2018.

[23] M. Singh, R. Singh, M. Vatsa, N. K. Ratha, and R. Chellappa. Recognizing disguised faces in the wild. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):97–108, 2019.

[24] E. Smirnov, A. Melnikov, A. Oleinik, E. Ivanova, I. Kalinovskiy, and E. Luckyanets. Hard example mining with auxiliary embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 37–46, 2018.

[25] S. Suri, A. Sankaran, M. Vatsa, and R. Singh. On matching faces with alterations due to plastic surgery and disguise. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2018.

[26] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[27] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(Nov):45–66, 2001.

[28] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl. Sampling matters in deep embedding learning. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[29] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, Jun 2015.

[30] Y. Yao, X. Li, Y. Ye, F. Liu, M. K. Ng, Z. Huang, and Y. Zhang. Low-resolution image categorization via heterogeneous domain adaptation. *Knowledge-Based Systems*, 163:656–665, 2019.