Mitigating Catastrophic Interference using Unsupervised Multi-Part Attention for RGB-IR Face Recognition

Kshitij Nikhal[1*], Nkiruka Uzuegbunam[2], Bridget Kennedy[2], and Benjamin S. Riggan[1*]

[1]University of Nebraska-Lincoln, [2]BlueHalo,
[*]Corresponding author: knikhal2@huskers.unl.edu, briggan2@unl.edu

Abstract

Modern algorithms for RGB-IR facial recognition—a challenging problem where infrared probe images are matched with visible gallery images—leverage precise and accurate guidance from curated (i.e., labeled) data to bridge large spectral differences. However, supervised cross-spectral face recognition methods are often extremely sensitive due to over-fitting to labels, performing well in some settings but not in others. Moreover, when fine-tuning on data from additional settings, supervised cross-spectral face recognition are prone to catastrophic forgetting. Therefore, we propose a novel unsupervised framework for RGB-IR face recognition to minimize the cost and time inefficiencies pertaining to labeling large-scale, multi-spectral data required to train supervised cross-spectral recognition methods and to alleviate the effect of forgetting by removing over dependence on hard labels to bridge such large spectral differences. The proposed framework integrates an efficient backbone network architecture with part-based attention models, which collectively enhances common information between visible and infrared faces. Then, the framework is optimized using pseudo-labels and a new cross-spectral memory bank loss. This framework is evaluated on the ARL-VTF and TUFTS datasets, achieving 98.55% and 43.28% true accept rate, respectively. Additionally, we analyze effects of forgetting and show that our framework is less prone to these effects.

1. Introduction

Facial recognition (FR) technology has been shown to have far-reaching social implications, from criminal identification to unlocking personal devices. Most FR technologies rely upon visible (RGB) spectrum (0.4–0.75μm) imagery, since conventional visible cameras are ubiquitous and cost effective. However, RGB imagery is vulnerable to illumination changes, making RGB-based FR ineffective in low-light conditions, such as nighttime. To address this limitation, researchers have shifted focus to FR beyond the visible spectrum [1], such as midwave (MWIR) and longwave infrared (LWIR) or 3–8μm, and 8–15μm, respectively. Although infrared (IR) sensors have been deployed in real world applications, most FR databases/galleries consist of RGB images leading to reduced match scores between RGB and IR images. There are three major challenges in advancing FR outside of the visible spectrum:

1. learning with limited data of specific size (number of images), quality (resolution), and complexity

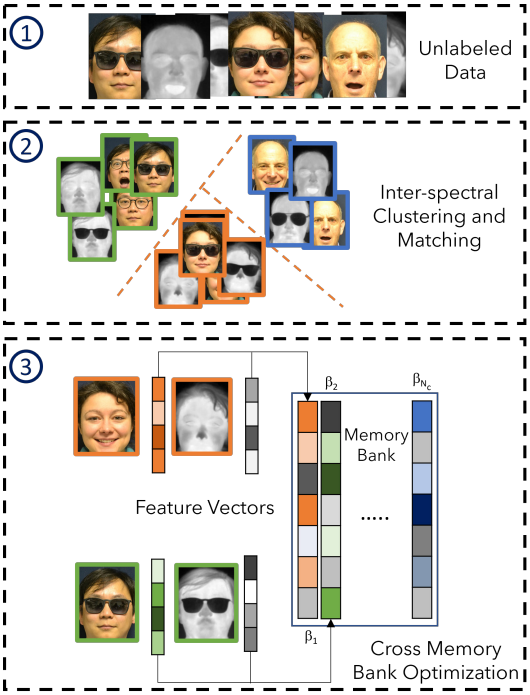


Figure 1. Overview of our proposed method. The approach uses clustering and matching to generate pseudo-labels, that enable the learning of spectral invariant information through a cross-memory bank without the need for supervision.

(co-variates), comprising multi-spectral imagery with 100+ identities,
2. the lack of annotated data and the significant time, cost, and resources required for curation and labeling (from challenge 1),
3. the large domain disparity (texture, color, resolution, etc.) between RGB and IR images, making it susceptible to over-fitting.

While these problems are applicable to many cross-domain applications, this paper addresses the three challenges in the context of RGB-IR face recognition [11, 16, 29].

Recently, methods leveraging local face parts (or patches) [20, 28], holistic face [13], synthesized face [31, 32], or attribute-enhanced [4, 5, 14] representations have been developed to mitigate the divergence between corresponding faces acquired from IR and RGB spectra. These methods extensively rely on supervised learning optimization techniques, which heavily depend on precisely curated (labeled) information for both RGB and IR face imagery. Although data annotation may be easier to semi-automate for RGB imagery due to prevalence of existing tools and data, comparable tools or data are not as widely available for IR imagery. Therefore, the ability to learn discriminative representations without needing significant amounts of labeled data can better inform multi-spectral FR data collections (challenge 1), and can help quickly shift effort/resources toward analyzing new data (challenge 2).

While supervised methods have achieved great success on datasets such as [12, 21, 24], the cost and time of annotating such datasets is very high. Additionally, supervised methods, especially highly parameterized deep neural networks, are susceptible to poor generalization and catastrophic interference (or forgetting) due to over-fitting to labels. Recently, unsupervised learning (e.g., clustering) techniques have shown promise for visible-based recognition [17–19], but often exhibit reduced discriminability compared to supervised techniques and have trouble under challenging conditions like resolution differences, occlusions, and spectral differences. Therefore, in this paper, we are among the first to enhance unsupervised learning for RGB-IR (cross-spectrum) FR (challenge 3).

In our proposed unsupervised cross-spectrum FR framework, by clustering and matching face images corresponding to the same identity from different spectra, we aim to learn spectral invariant information without identity labels or labeled image pairs (addressing challenge 2). The proposed framework is set up with an efficient backbone architecture that is augmented with lightweight part-based attention enabling it to learn on limited amount of data (addressing challenge 1), and bridging the domain gap without over-fitting (addressing challenge 3).

Cross-spectrum techniques (supervised and unsupervised) often share parameters of early layers to process inputs from both spectra [20, 25, 27], but tend to learn spectral specific information in later layers. Instead of two (or more) networks with dedicated processing per spectra, we design a singular network to learn highly localized information via multi-part attention which acts as form of regularization to limit the amount of spectral specific information. This framework is optimized using a new cross-spectrum memory bank loss that exploits pseudo-labels created by combining intra-spectral clustering and inter-spectral matching. Our approach effectively learns discriminative and spectral invariant embeddings without identity labels. This enables us to (a) learn spectral invariance based on multi-spectral clusters and (b) discriminatively cluster multi-spectral data using spectral invariant embeddings.

Overall, we propose a novel unsupervised framework (Figure 1) for RGB-IR FR with the following contributions:
(a) an efficient attention-based architecture—to focus on more generalizable information that is significantly less specific toward spectral specialized information,
(b) a new multi-part attention—to promote highly localized information that limits the amount of spectral context,
(c) a new cross-spectral memory bank clustering loss—to encourage discriminative cross-spectral clustering of identities using pseudo-labels.

This framework is extensively evaluated using both ARL-VTF [24] and TUFTS [21] datasets, where we demonstrate enhanced RGB-IR FR using unsupervised learning. Moreover, we perform an important systematic study that highlights how our framework (and in general, unsupervised learning) can mitigate the effects of catastrophic interference/forgetting that are common to supervised learning.

2. Related Work

Supervised methods include RST [20], where a residual spectral transform is learned to produce domain invariant representations. In [26], dictionary learning is used to generate a sparse feature representation that is domain-independent. But supervised methods are known to be prone to over-fitting and need significant amounts of labeled data to overcome domain disparity, whereas our method is based on an efficient backbone that can learn with limited amount of unlabeled data.

Many methods use generative adversarial networks (GANs) to synthesize an image to another spectral domain. Pix2Pix [15] utilizes conditional adversarial networks to translate images from thermal to visible using a U-Net based architecture. GANVFS [31] jointly estimates the visible features and visible image reconstruction from thermal images using identity and perceptual objectives, to retain discriminative face characteristics. SAGAN [3] uses a self-attention module to capture long-range dependency information with cycle consistency and a patch discriminator for inter-domain synthesis. While generative models visually

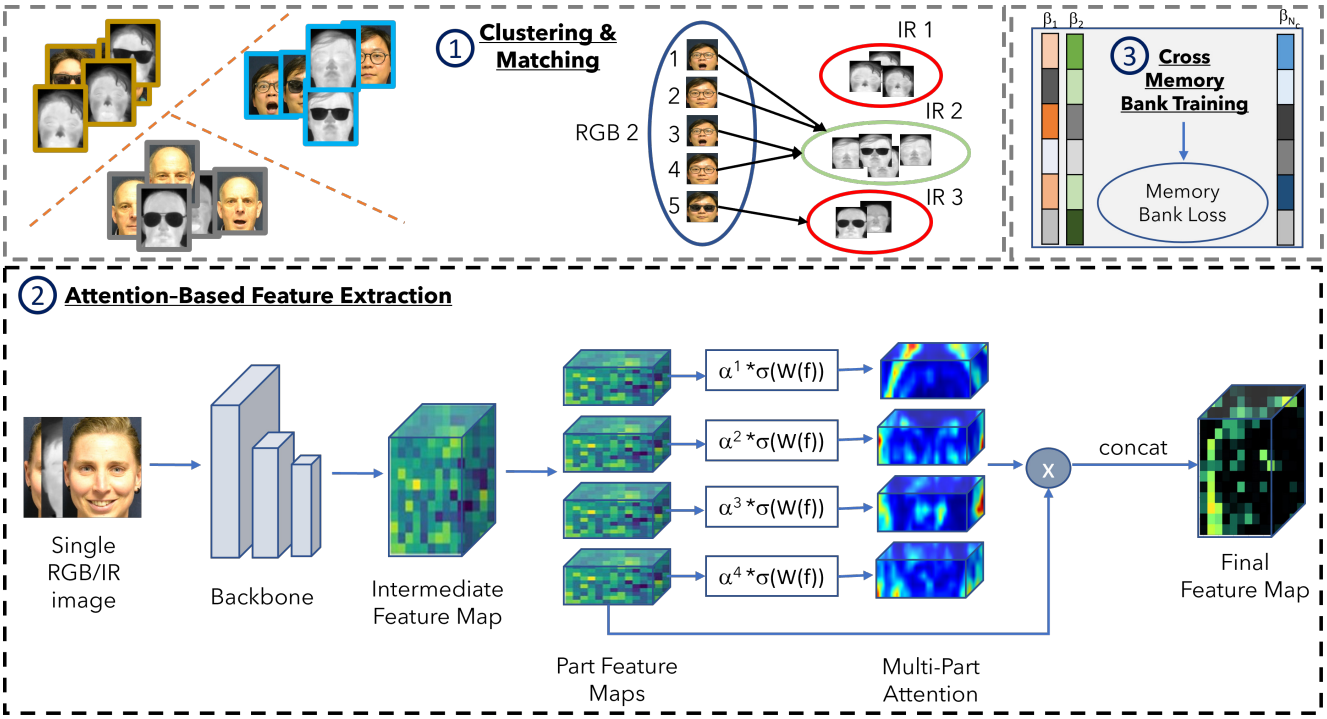


Figure 2. The framework performs inter- and intra-spectral clustering and matching to generate an accurate dataset without using labels. Intermediate features are extracted by the EfficientNet model and spatially divided into four equal parts. Attention for each part is calculated independently and the final attention-refined feature map is concatenated. This feature map is optimized using the cross-memory bank loss.

show notable domain gap reduction, there is a new intermediate domain gap created between synthesized and 'real' images leading to reduced discriminability. Our method does not synthesize images, but uses spectral-specific paths to learn generalizable information between domains.

Most work in unsupervised face recognition is adapting a labeled source dataset to an unlabeled target dataset, known as unsupervised domain adaptation (UDA). A popular idea is to perform feature space alignment where the distance between domains is minimized by learning a transformation from source to target domain [2, 7, 8]. Fully unsupervised learning has been explored in tasks such as person re-identification (ReID) where no labels are assumed. Methods such as BUC [17] and GAM [18] use bottom-up clustering to find identity matches and use a memory bank to optimize the network, where each image is considered a singleton cluster at the start. However, such methods have not studied inter-domain clustering, being prone to cluster intra-domain instances at early stages of training, thereby increasing the domain gap in the feature space. SpCL [10] uses self-paced contrastive learning framework that utilizes a hybrid memory to gradually generate reliable clusters. However, the lack of attention and high number of parameters results in inferior performance compared to our method. In addition, the cluster quality criterion also introduces noise in the cluster labels. Our method uses high confident clusters, and an efficient network with fewer parameters to learn discriminative features from less data.

3. Methodology

Figure 2 depicts our cross-spectral framework that has three distinct components: attention-based feature extraction, clustering and matching, and cross memory bank optimization. First, RGB and IR images are clustered independently to form intra-spectral clusters. Then, RGB clusters are matched with IR clusters using inter-spectral matching. Pseudo-labels are used to optimize the network using the cross-spectral memory bank loss and multi-part attention.

3.1. Preliminaries

We denote RGB and IR training samples as $X_{RGB} = \{x_1^{rgb}, x_2^{rgb}, \ldots, x_n^{rgb}\}$ and $X_{IR} = \{x_1^{ir}, x_2^{ir}, \ldots, x_n^{ir}\}$,respectively. Pseudo-labels generated by the clustering algorithm are denoted as $Y_{RGB} = \{y_1^{rgb}, y_2^{rgb}, \ldots, y_n^{rgb}\}$ and $Y_{IR} = \{y_1^{ir}, y_2^{ir}, \ldots, y_n^{ir}\}$. Initially, the labels are initialized to -1, indicating not clustered. We aim to learn a mapping $\phi(x_i; \theta)$ that is discriminative and invariant to the spectra, where θ are the learnable parameters.

3.2. Clustering and Matching

We adopt the challenging setting where no intra- or inter-spectral labels are available during the training phase. First, we utilize DBSCAN [6] (density-based clustering) to generate the initial set of cluster sample labels within spectra (i.e., RGB and IR separately). The cosine distance metric is used to quantify cluster similarity. DBSCAN performs superior to K-means as it only picks samples in spatially high density areas, while also allowing clusters to be any shape (vs K-means which assumes convex shapes). While DBSCAN overlooks many training samples in the initial phase, our lightweight attention-based network is adept at learning efficiently from the limited data present.

After obtaining the pseudo-labels for each spectra, inter-spectral matching is used to solve the association problem between RGB clusters and IR clusters. For each RGB cluster, we rank each sample in that cluster to all the IR clusters according to the cosine distance metric. The most common cluster that the RGB matches to is assigned as the matching cluster. In the first box of Figure 2, the matching process is shown where RGB cluster 2 has five samples in the cluster, and four out of five (1,2,3 and 4) have IR cluster 2 as the top match according to the distance metric. Hence, IR 2 is assigned to RGB cluster 2 (the cluster numbers are arbitrary). To ignore noise and draws, we only consider matches where majority of the votes are for the same cluster.

Finally, after some training (Section 3.4), this clustering and matching process can be repeated to help generate additional pseudo-labels for training.

3.3. Multi-Part Attention (MPA)

To match faces across spectrum, it is vital to learn a mapping that is robust to spectral changes. Intermediate features have been shown to better transfer between spectral domains for FR [20]. Given the intermediate feature maps from a backbone convolutional network, we employ a new multi-branch attention-based architecture that consists of four 1×1 convolutional layers and four weighting parameters α. This helps to further emphasize highly localized information that is robust to spectral differences. We use a truncated EfficientNet [30] backbone (only 851.81K parameters and 255.43 MMac computational complexity) to balance network depth and width. Formally, given an image, the MPA refined feature maps are given by

$$\phi(x) = concat\{F^1_{refined}(x), \ldots, F^4_{refined}(x)\} \quad (1)$$

Eq. 1 combines four local part-refined representations

$$F^p_{refined} = \alpha^p \sigma\{W * F^p_{EN}(x)\} \odot F^p_{EN}(x), \quad (2)$$

where F^p_{EN} for p $\in 1, 2, 3, 4$ denotes the four equal parts of F_{EN} pertaining to: hair (part 1), forehead (part 2), nose and ears (part 3), and chin (part 4). The $*$ operator denotes convolution and $\odot$ denotes element-wise multiplication. Both EfficientNet and locally weighted part-based attention models—parameterized by α^p and W— are also optimized.

3.4. Cross-spectral Memory Bank Optimization

To optimize our network without ground truth labels, we make use of a memory bank matrix M that has dimension size of number of clusters $\times$ feature embedding size. M stores cluster centers produced from both IR and RGB samples, i.e., centroid of all cross-spectral instances assigned to that particular cluster. We define the cross-spectral pseudo-labels (cluster labels) assigned by the clustering and matching as $Y = \{y_1, \ldots, y_n\}$ and the unique cluster labels in Y as $\beta = \{\beta_1, \ldots, \beta_{N_c}\}$ where N_c is the number of clusters. Therefore, $M = \{M_{\beta_1}, M_{\beta_2}, \ldots, M_{\beta_{N_c}}\}$ contains the cluster centers with size N_c.

Let β_i denote the cluster index in which x_i is assigned for $i = 1 \ldots n$. The probability that image x_i belongs to a cluster β_i is given by

$$P(\beta_i|x_i) = softmax\{(M_{\beta_i} \cdot \phi(x_i; \theta))/\tau\} \quad (3)$$

where τ is a temperature parameter of the distribution set to 0.1. Feature embeddings $\phi(x_i; \theta)$ for $i = 1 \ldots n$ are optimized via stochastic gradient based updates of parameters θ. Then, the optimized embeddings are used to update the memory bank M. Therefore, the objective function is to minimize the cross entropy as,

$$J_{MBL} = -\sum_{i=1}^{n} \log\{P(\beta_i|x_i)\}. \quad (4)$$

4. Experimental Results

4.1. Datasets

ARL Visible-Thermal Face Dataset (ARL-VTF) [24] contains 395 identities (295 in the training set, 100 in the testing set) with over 500,000 thermal and visible images. The dataset has two gallery sets: subjects without eyeglasses G_VB0- and subjects including glasses G_VB0+. The query/probe set has three different conditions: baseline, expression and eyeglasses. For baseline thermal imagery in the probe set, we have P_TB0, P_TB-, and P_TB+, where '0' denotes subjects without glasses, '-' denotes subjects who have glasses but are not wearing them, and '+' denotes subjects that have glasses and are wearing them. For expression thermal imagery, we have P_TE0 and P_TE- probe sets. The naive 'VGG16' baseline method simply extracts deep features from a VGG backbone network [23]. As a ground-truth baseline method (GT Vis-to-Vis), the query

Table 1. Baseline and Expression verification performance is compared with **Blue** indicating best unsupervised performance and **cyan** indicating best supervised performance.

			Gallery G_VB0-				Gallery G_VB0+			
	Probes	Method	AUC	EER	1% FAR	5% FAR	AUC	EER	1% FAR	5% FAR
BASELINE PROBE SETS	P_TB0	VGG16 [23]	61.37	43.36	3.13	11.28	62.83	42.37	4.19	13.29
		Pix2Pix [15]	71.12	33.80	6.95	21.28	75.22	30.42	8.28	27.63
		GANVFS [31]	97.94	8.14	75.00	88.93	98.58	6.94	79.09	91.04
		SAGAN [3]	99.28	3.97	87.95	96.66	99.49	3.38	90.52	97.81
		RST [20]	99.76	2.30	96.84	98.43	99.87	1.84	97.29	98.80
		DPIT [9]	**99.99**	**0.15**	**100.00**	**100.00**	**100.00**	**0.12**	**100.00**	**100.00**
		MPA (OURS)	**99.92**	**1.28**	**98.55**	**100.00**	**99.93**	**1.08**	**98.83**	**100.00**
	P_TB-	VGG16 [23]	61.14	41.64	2.77	16.11	57.61	44.73	1.38	6.11
		Pix2Pix [15]	68.77	38.02	6.69	20.28	52.11	48.88	2.22	4.66
		GANVFS [31]	99.36	3.77	84.88	97.66	87.34	18.66	7,00	29.66
		SAGAN [3]	99.63	2.66	91.55	98.88	89.24	19.49	16.33	41.22
		RST [20]	99.83	1.95	96.00	99.48	99.03	4.79	85.56	95.86
		DPIT [9]	**100.00**	**00.0**	**100.00**	**100.00**	**97.97**	**0.66**	**100.00**	**100.00**
		MPA (OURS)	**99.62**	**4.35**	**91.31**	**96.70**	**98.28**	**6.96**	**69.29**	**89.44**
	P_TB+	VGG16 [23]	59.52	42.60	4.66	6.00	78.26	29.77	3.88	21.33
		Pix2Pix [15]	59.68	41.72	3.33	3.33	67.08	36.44	2.68	11.11
		GANVFS [31]	87.61	20.16	20.55	44.66	96.82	8.66	46.77	83.00
		SAGAN [3]	91.11	17.43	22.33	55.66	97.96	7.21	60.11	88.70
		RST [20]	99.28	5.32	89.21	94.79	99.97	0.73	99.47	100.00
		DPIT [9]	**99.91**	**1.94**	**96.84**	**100.00**	**100.00**	**0.32**	**100.00**	**100.00**
		MPA (OURS)	**96.97**	**9.77**	**55.89**	**83.04**	**99.80**	**1.95**	**93.51**	**99.56**
EXPRESSION PROBE SETS	P_TE0	VGG16 [23]	61.40	41.96	3.40	12.18	62.50	41.38	4.60	13.25
		Pix2Pix [15]	69.10	35.98	7.01	16.44	73.97	31.87	7.93	19.60
		GANVFS [31]	96.81	10.51	70.41	84.00	97.73	8.90	74.20	86.80
		SAGAN [3]	98.46	6.44	81.11	92.49	98.89	5.60	84.23	93.94
		RST [20]	98.95	3.61	92.61	96.88	99.01	3.57	92.69	96.93
		DPIT [9]	**99.79**	**2.39**	**96.49**	**98.31**	**99.70**	**2.33**	**96.52**	**98.29**
		MPA (OURS)	**99.15**	**4.67**	**89.52**	**95.47**	**99.25**	**4.46**	**90.73**	**95.80**
	P_TE-	VGG16 [23]	63.26	42.34	4.66	16.28	59.33	43.17	2.04	8.00
		Pix2Pix [15]	68.78	36.24	7.75	18.06	51.05	49.11	2.26	4.95
		GANVFS [31]	98.66	5.93	73.17	92.82	83.68	22.41	6.77	22.13
		SAGAN [3]	99.30	3.84	82.55	97.44	86.12	21.68	9.88	31.62
		RST [20]	99.83	2.27	95.66	99.48	99.48	3.05	89.45	98.07
		DPIT [9]	**99.88**	**0.81**	**99.47**	**99.87**	**99.77**	**2.92**	**95.33**	**98.87**
		MPA (OURS)	**99.62**	**3.00**	**90.98**	**98.47**	**98.97**	**5.05**	**73.30**	**94.70**

images (IR) are replaced with the corresponding RGB images. While the dataset also contains different pose condition, matching clusters across varying poses is very challenging, and hence we do not address it in this work.

TUFTS [22] provides thermal-visible face images of 113 identities from more than 15 countries and a 0.52 male-to-female identity ratio. For this setting, we select baseline and expression images from the training set with 315 visible and 315 thermal images. We split the dataset with 63 identities in the training set and 50 identities in the testing set. For both the probe and gallery sets, we have 250 images each.

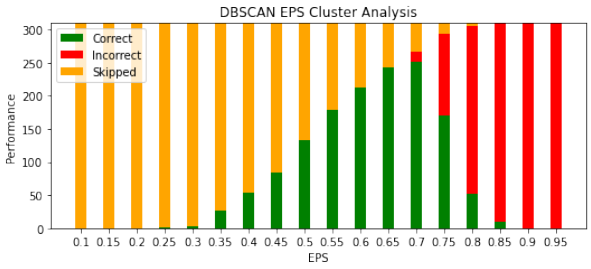
By employing both large- and small-scale datasets, we illustrate the robustness of our method with respect to dataset size, number of identities, and ethnic diversity.

4.2. Evaluation Metrics

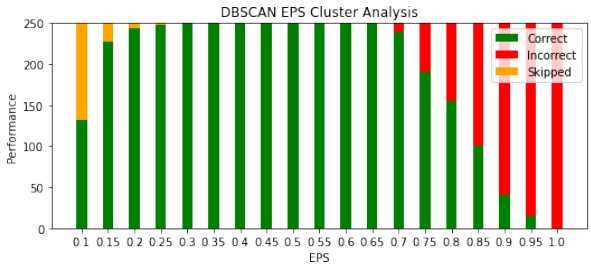
Algorithm performance is measured using Area Under the Curve (AUC), equal error rate (EER), and True Positive Rate (TPR) at False Acceptance Rate (FAR) metrics. We evaluate TPR at 1% FAR and 5% FAR.

4.3. Implementation Details

For our backbone network, we use a EfficientNet-B0 model truncated at layer 4 having only 851.81K parameters and 255.43 MMac. The multi-part attention model only contains 50.63K parameters and 2.48 MMac computational complexity. By experimentation, we found that intermediate features at layer 4 are more transferable and faster to converge. The learning rate is set to $1e^{-4}$ with the RM-

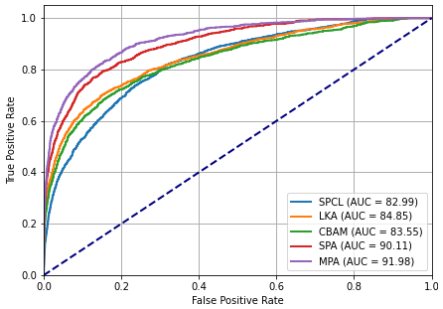


(a) Varying EPS on the TUFTS dataset.

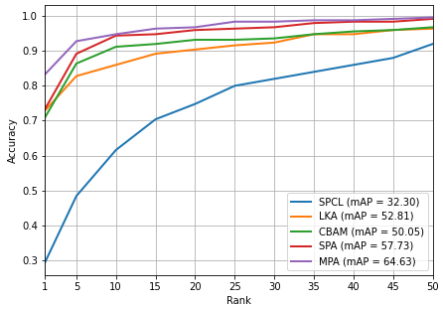


(b) Varying EPS on the ARL-VTF dataset.

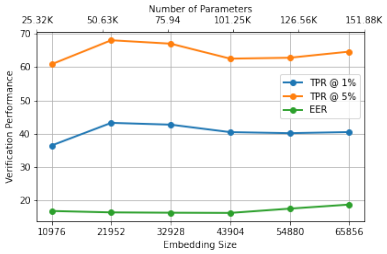
Figure 3. Analyzing the effects of EPS for spatial-based density clustering.



(a) ROC curves on the TUFTS dataset.



(b) CMC curves on the TUFTS dataset.



(c) Effect of embedding sizes.

Figure 4. Performance on the TUFTS dataset.

SProp optimizer. The images are resized to 224×224 and the batch size is set to 16. All experiments are done using the PyTorch framework and trained on one NVIDIA RTX 2080 Ti GPU with 11GB video memory.

4.4. Clustering Density Analysis for Training

Including as much correct data is essential for unsupervised training, even though DBSCAN can help address less confident clusters. Therefore, we analyze the effects of parameter epsilon (EPS) that controls the neighbourhood radius on the validation set. Figure 3 shows that if EPS is too low, all data is considered noise, while if it is too high, all data is in a single cluster. Figures 3a and 3b demonstrate that an EPS value between 0.5 and 0.7 retains correct clusters (green), avoids incorrect classifications (red), and minimizes data point skipping (orange). Therefore, we set our EPS to 0.6 for our experiments.

4.5. Quantitative Results

Table 1 presents results on the ARL-VTF dataset on the face-verification task. The first row section uses probe set P_TB0 with galleries G_VB0- and G_VB0+. All methods (except our proposed work) uses full ground truth label supervision for training. Compared to GAN methods such as Pix2Pix, GANVFS, and SAGAN, our method achieves greater than 10% improvement in TPR @ 1% FAR and substantial decrease in EER with a maximum of 9.77 across all settings. Compared to supervised state-of-the-art, our method performs comparably to DPIT [9] and outperforms RST [20] in certain conditions (e.g., P_TB0 and G_VB0+) with a 100% TPR @ 5% FAR and 1.08 EER. Similar performance trends are seen on different probe and gallery sets.

Figure 4 compares performance on TUFTS with other attention models using the proposed backbone network and objective functions in this work. Our method outperforms CBAM in Figure 4a, with a 12.3% improvement in TPR @ 1% FAR and a increase of 8.43 in AUC. Our method also surpasses the recent state-of-the-art LKA by 11.20 in TPR @ 1% FAR and 7.13 in EER. The single-path attention achieves a TPR @ 1% FAR of 39.30%, while the multi-part attention achieves the highest score of 43.28%. These trends are consistent in Figure 4b, where our method achieves the highest rank-1 accuracy of 83.60% and mAP (mean Average Precision) of 64.63.

4.6. Embedding Size

In Figure 4c, we compare embedding sizes by varying the number of kernels in the 1×1 convolutional filter. We find that increasing the dimensionality (upto 65856) does not result in any performance improvement. Moreover, increasing the number of parameters also poses the risk of over-fitting. Reducing filters to obtain a 10976 dimensional embedding results in under-parametrization and inferior performance. We determine that an embedding size of

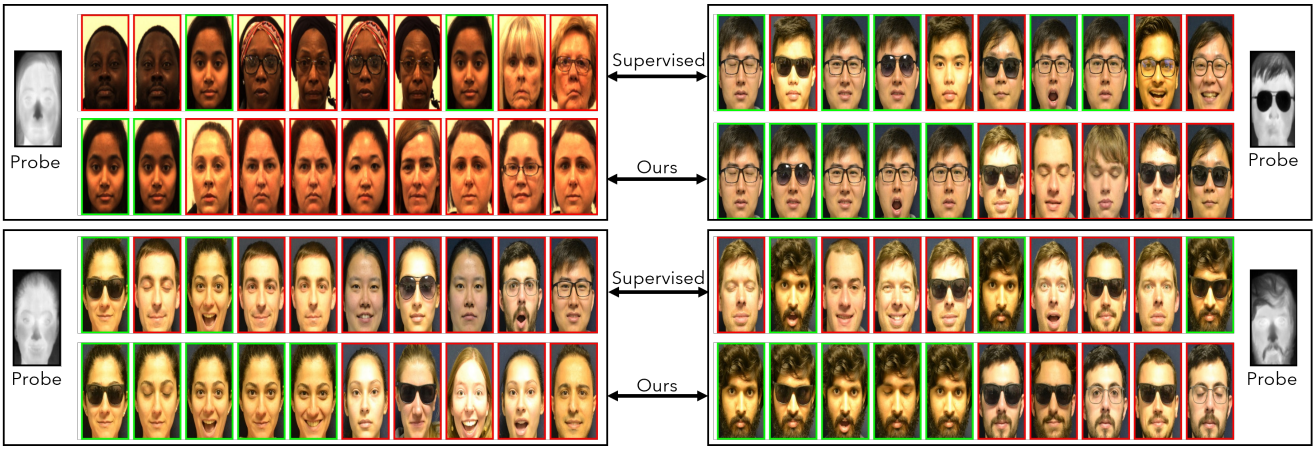
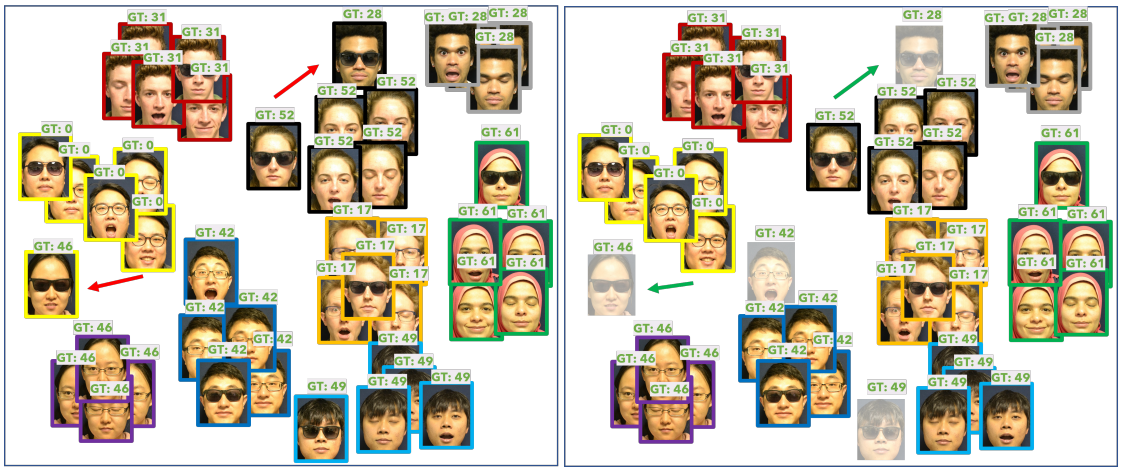
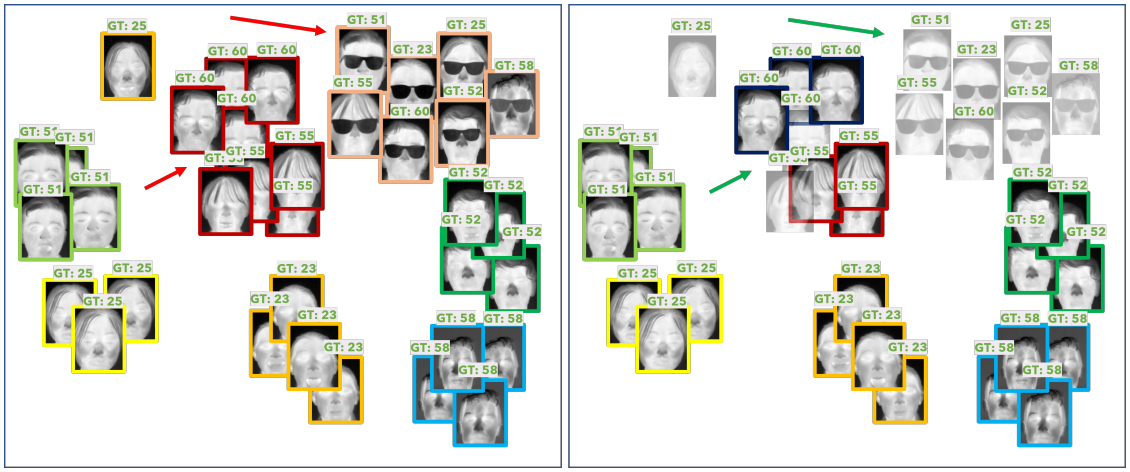


Figure 5. Ranking results (top-10) on the TUFTS and ARL-VTF dataset where green border denotes correct matches and red border denotes incorrect matches. Supervised methods (row 1 and 3) are consistently biased towards eye-wear, skin color, etc. whereas our unsupervised method (row 2 and 4) mitigates those biases by focusing on distinct features of the face.



(a) DBSCAN (right) only clusters confident clusters, essential for training on limited data, unlike K-means (left) which incorrectly classifies RGB images belonging to different identity. Less opaque images denote not belonging to any clusters.



(b) Unlike K-means (left), DBSCAN (right) prevents incorrect clustering by not grouping all images with eye-wear into one cluster.

Figure 6. T-SNE representation of the embeddings and clustering analysis. Image border is cluster labels and label is ground truth ID.

21952 produces optimal results, while also being compact enough to facilitate deployment in real-world applications. Compared to supervised RST [20], we have a 7x reduction in embedding size (21952 vs 160000).

4.7. Qualitative Results

Figure 5 showcases ranking results on the TUFTS and ARL-VTF dataset. The first and third row presents results using full supervision (supervised) and second and fourth row presents results using no ground-truth labels (Ours). In most cases, our method is capable of retrieving all images (including those with and without sunglasses), whereas supervised learning exhibits greater separability between correct matches and exhibits bias towards retrieving subjects with similar skin tone/eyewear, resulting in inferior performance. Figure 6 shows the intra-spectral clustering performance with K-means (left) and DBSCAN (right). Cluster labels are denoted by image borders, and image label is the GT (ground truth) label. As is evident, DBSCAN prevents the incorrect clustering of subjects , particularly where eyewear is present, which is crucial when learning on limited data. K-means introduces noise in the clusters, resulting in incorrect clusters that are detrimental for training.

4.8. Catastrophic Interference

Catastrophic interference (or forgetting) is the phenomena where trained models forget previously learned information on a dataset after learning new information from another dataset. While both supervised and unsupervised methods are prone to this interference, in this section, we analyze the degree of interference in our approach using unsupervised and supervised learning. In Table 2, we first perform unsupervised learning (USL) on the TUFTS dataset achieving a TPR @ 1% FAR of 43.28%. Then, we test it on the ARL-VTF dataset without any fine-tuning and achieve TPR of 76.09%. Next, we fine-tune it on the ARL dataset (again, without using any labels) and notice an improvement up to 91.80%. If we test this model back on the TUFTS dataset, we actually notice an improvement from 43.28% to 44.52% which shows that unsupervised learning of features not only helps mitigate forgetting, but also transfers better. We repeat this experiment using the ARL-VTF dataset as the starting point, and again see an improvement from 98.55% to 98.73% supporting our hypothesis.

TUFTS (USL)	ARL	ARL (USL)	TUFTS
43.28	76.09	91.80	**44.52**
ARL (USL)	TUFTS	TUFTS (USL)	ARL
98.55	38.28	47.70	**98.73**

Table 2. Unsupervised training (USL) and cross-testing.

In Table 3, we analyze forgetting in the supervised learning setting. Again, we start off with the TUFTS dataset and achieve a higher performance than unsupervised learning (which is expected with full ground truth labels) of 54.94%. However, we test it on the ARL-VTF without fine-tuning on it and observe a lower performance than USL with 74.75% TPR. Next, we fine-tune it on the ARL dataset with full supervision but is unable to achieve a high performance even with increasing number of training steps and decreasing/increasing the learning rate. Testing it back on the TUFTS dataset shows sub-optimal performance of 34.62% which supports our hypothesis that supervised learning quickly overfits on limited labeled data and suffers from catastrophic forgetting. This is corroborated by repeating the experiment on the ARL-VTF dataset as the starting point and noticing a performance dip from 98.10% to 93.59%, even though full supervision is used.

TUFTS (SL)	ARL	ARL (SL)	TUFTS
54.94	74.75	84.01	**34.62**
ARL (SL)	TUFTS	TUFTS (SL)	ARL
98.10	32.14	61.82	**93.59**

Table 3. Supervised learning (SL) and cross-testing.

5. Conclusion

In this work, we adopt the challenging setting of training an end-to-end model without using any intra- or inter-spectral annotations. We employ DBSCAN to generate initial cluster labels within spectra, and then combine inter-spectral clusters using a matching scheme. To pay attention to distinct parts on the face that is robust to both spectra, we employ a part-based attention module that is efficient enough to learn using only a few correct cluster matches. The network is optimized using a cross-spectral memory bank that gravitates the same identity's samples to a single compact cluster center. Our results showcase that utilizing unsupervised learning and encouraging the network to learn and cluster similar visual patterns has helped overcome the effect of bias, leading to better performance on different datasets with low catastrophic interference.

6. Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes not withstanding any copyright annotation therein.

References

[1] Thirimachos Bourlai and Lawrence A. Hornak. Face recognition outside the visible spectrum. *Image and Vision Computing*, 55:14–17, 2016. Recognizing future hot topics and hard problems in biometrics research. 1

[2] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. Idm: An intermediate domain module for domain adaptive person re-id, 2021. 3

[3] Xing Di, Benjamin S. Riggan, Shuowen Hu, Nathaniel J. Short, and Vishal M. Patel. Polarimetric thermal to visible face verification via self-attention guided synthesis. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019. 2, 5

[4] Xing Di, Benjamin S. Riggan, Shuowen Hu, Nathaniel J. Short, and Vishal M. Patel. Multi-scale thermal to visible face verification via attribute guided synthesis. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2):266–280, 2021. 2

[5] Xing Di, H. Zhang, and V. Patel. Polarimetric thermal to visible face verification via attribute preserved synthesis. *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10, 2018. 2

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996. 4

[7] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *2013 IEEE International Conference on Computer Vision*, pages 2960–2967, 2013. 3

[8] Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. Joint cross-domain classification and subspace learning for unsupervised adaptation, 2014. 3

[9] Cedric Nimpa Fondje, Shuowen Hu, and Benjamin S. Riggan. Learning domain and pose invariance for thermal-to-visible face recognition, 2022. 5, 6

[10] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33:11309–11321, 2020. 3

[11] Shuowen Hu, Nathaniel J. Short, Prudhvi K. Gurram, Kristan P. Gurton, and Christopher Reale. *MWIR-to-Visible and LWIR-to-Visible Face Recognition Using Partial Least Squares and Dictionary Learning*, pages 69–90. Springer International Publishing, Cham, 2016. 2

[12] Shuowen Hu, Nathaniel J. Short, Benjamin S. Riggan, Christopher Gordon, Kristan P. Gurton, Matthew Thielke, Prudhvi Gurram, and Alex L. Chan. A polarimetric thermal database for face recognition research. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 187–194, 2016. 2

[13] Seyed Mehdi Iranmanesh, Ali Dabouei, Hadi Kazemi, and Nasser M. Nasrabadi. Deep cross polarimetric thermal-to-visible face recognition. In *2018 International Conference on Biometrics (ICB)*, pages 166–173, 2018. 2

[14] S. M. Iranmanesh and N. M. Nasrabadi. Attribute-guided deep polarimetric thermal-to-visible face recognition. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019. 2

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2, 5

[16] Brendan Klare and Anil K. Jain. Heterogeneous face recognition: Matching nir to visible light images. In *2010 20th International Conference on Pattern Recognition*, pages 1513–1516, 2010. 2

[17] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 2, pages 1–8, 2019. 2, 3

[18] Kshitij Nikhal and Benjamin S. Riggan. Unsupervised attention based instance discriminative learning for person re-identification. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2421–2430, 2021. 2, 3

[19] Kshitij Nikhal and Benjamin S. Riggan. Multi-context grouped attention for unsupervised person re-identification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pages 1–1, 2022. 2

[20] Cedric Nimpa Fondje, Shuowen Hu, Nathaniel J. Short, and Benjamin S. Riggan. Cross-domain identification for thermal-to-visible face recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2020. 2, 4, 5, 6, 8

[21] Karen Panetta, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, Holly A. Taylor, Arash Samani, and Xin Yuan. A comprehensive database for benchmarking imaging systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):509–520, 2020. 2

[22] Karen Panetta, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, Holly A. Taylor, Arash Samani, and Xin Yuan. A comprehensive database for benchmarking imaging systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):509–520, 2020. 5

[23] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 4, 5

[24] Domenick Poster, Matthew Thielke, Robert Nguyen, Srinivasan Rajaraman, Xing Di, Cedric Nimpa Fondje, Vishal M. Patel, Nathaniel J. Short, Benjamin S. Riggan, Nasser M. Nasrabadi, and Shuowen Hu. A large-scale, time-synchronized visible and thermal face dataset. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1558–1567, 2021. 2, 4

[25] Domenick D. Poster, Shuowen Hu, Nathan J. Short, Benjamin S. Riggan, and Nasser M. Nasrabadi. Visible-to-thermal transfer learning for facial landmark detection. *IEEE Access*, 9:52759–52772, 2021. 2

[26] Christopher Reale, Nasser M. Nasrabadi, and Rama Chellappa. Coupled dictionaries for thermal to visible face recog-

nition. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 328–332, 2014. 2

[27] Christopher Reale, Nasser M. Nasrabadi, Heesung Kwon, and Rama Chellappa. Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 320–328, 2016. 2

[28] Benjamin S. Riggan, Nathaniel J. Short, and Shuowen Hu. Optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–7, 2016. 2

[29] M. Saquib Sarfraz and Rainer Stiefelhagen. Deep perceptual mapping for cross-modal face recognition. *International Journal of Computer Vision*, 122:426–438, 2016. 2

[30] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4

[31] He Zhang, Vishal M. Patel, Benjamin S. Riggan, and Shuowen Hu. Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 100–107, 2017. 2, 5

[32] He Zhang, Benjamin Riggan, Shuowen Hu, Nathaniel Short, and Vishal Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision*, 127:1–18, 03 2019. 2