A Large-Scale, Time-Synchronized Visible and Thermal Face Dataset

Domenick Poster[1,*] Matthew Thielke[2] Robert Nguyen[2,3] Srinivasan Rajaraman[2,3]
Xing Di[2,4] Cedric Nimpa Fondje[5] Vishal M. Patel[4] Nathaniel J. Short[2,3]
Benjamin S. Riggan[5] Nasser M. Nasrabadi[1] Shuowen Hu[2]

[1] West Virginia University, 395 Evansdale Dr., Morgantown, WV 26506
[2] DEVCOM Army Research Laboratory, 2800 Powder Mill Rd., Adelphi, MD 20783
[3] Booz Allen Hamilton, 8283 Grennsboro Dr., McLean, VA 22102
[4] Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218
[5] University of Nebraska-Lincoln, 1400 R St, Lincoln, NE 68588
[*] *Corresponding authors: dposter@mix.wvu.edu*

Abstract

Thermal face imagery, which captures the naturally emitted heat from the face, is limited in availability compared to face imagery in the visible spectrum. To help address this scarcity of thermal face imagery for research and algorithm development, we present the DEVCOM Army Research Laboratory Visible-Thermal Face Dataset (ARL-VTF). With over 500,000 images from 395 subjects, the ARL-VTF dataset represents, to the best of our knowledge, the largest collection of paired visible and thermal face images to date. The data was captured using a modern long wave infrared (LWIR) camera mounted alongside a stereo setup of three visible spectrum cameras. Variability in expressions, pose, and eyewear has been systematically recorded. The dataset has been curated with extensive annotations, metadata, and standardized protocols for evaluation. Furthermore, this paper presents extensive benchmark results and analysis on thermal face landmark detection and thermal-to-visible face verification by evaluating state-of-the-art models on the ARL-VTF dataset.

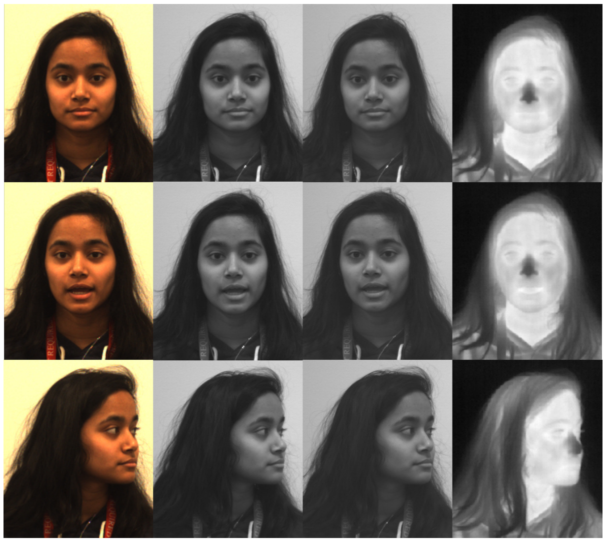


Figure 1: A set of images from the RGB (left), stereo monochrome (middle), and LWIR (right) cameras from the baseline (top), expression (middle), and off-pose (bottom) sequences.

1. Introduction

The use of thermal imaging has grown steadily over the past several decades, aided by improvements in sensor technology as well as reductions in cost. Thermal infrared sensors capture heat emissions, such as those radiated by the human body, in the 3 µm-5 µm medium wave infrared (MWIR) band and 7 µm-14 µm longwave infrared (LWIR) band. Thermal imaging of faces have applications in the military and law enforcement for face recognition in low-light and nighttime environments [14][19][33][8] and healthcare [11][28][35], which require robust recognition models in challenging unconstrained operational conditions. However, the majority of MWIR and LWIR face datasets available at the time of this paper's writing consist of lower resolution images from older thermal sensors.

While good rank-1 face recognition rates (around 90%) have been reported using 64×64 cropped face images captured by these older thermal cameras [24], there is still a large gap in meeting the aforementioned requirements for military, law enforcement, and healthcare applications. To help address these requirements, face datasets containing high resolution thermal imagery under various conditions, such as variable pose, expression, occlusion, and resolutions are needed. Furthermore, it is oftentimes desirable to synchronize and co-register the data being collected across multiple sensors to support algorithm development of fusion, domain adaptation, and cross-modal image synthesis approaches.

To this end, we present the Army Research Laboratory Visible-Thermal Face (ARL-VTF) dataset. This dataset is, to the best of our knowledge, the largest thermal face dataset publicly available for scientific research to date. The main contributions of the ARL-VTF dataset are:

- A multi-modal, time synchronized acquisition of 395 subjects and over 500,000 face images captured using multiple visible cameras for stereo 3D vision and one LWIR sensor (sample images shown in Figure 1).
- Three image sequences capturing baseline, expression, and pose conditions for each subject. A fourth condition, eye glasses, is captured if a subject wears glasses.
- Annotations for head pose, eyewear, face bounding box, and 6 face landmarks locations.
- Standardized protocols for model training and evaluation.

Results and analysis on the tasks of thermal face landmark detection and thermal-to-visible face verification using state-of-the-art deep learning models are presented as a benchmark.

2. Literature Review

In this section we provide a thorough comparison of several publicly released MWIR or LWIR face datasets and briefly highlight some notable characteristics of each. Table 1 presents a high-level comparison of the key statistics of different datasets, including the ARL Visible-Thermal Face Dataset (ARL-VTF) presented in this paper.

Collected primarily in 2002 with visible and LWIR cameras, the University of Notre Dame (UND) [5] dataset remains as one of the largest datasets in terms of unique identities (with 241 subjects), but has only four images per subject, and used what is now considered a very low resolution and low sensitivity uncooled microbolometer.

The IRIS [1] dataset has simultaneous recordings of 30 subjects in variable poses and expressions in both LWIR and visible, however no annotations included besides the subject id. The IRIS-M3 [4] dataset, however, contains 88 subjects simultaneously captured under a variety of indoor and outdoor lighting conditions with not only LWIR and visible cameras but also a multi-spectral imaging module.

Two different datasets have both been referred to as the University of Houston (UH) dataset. The more recent version [3] contains 7,590 MWIR images from 138 subjects. A slightly older version [18] contains 88 subjects and simultaneous acquisition of visible, thermal, and range data for 3d model generation. The thermal IR camera is not specified though presumably it is the same as [3].

The Natural Visible and Infrared Expression Database (NVIE) [34] captures subjects displaying a wide range of emotions. Sequences of unposed expressions were elicited by having subjects, some of whom wore glasses, observe video clips. Sequences of posed expressions were also captured with all subjects with and without glasses. The LWIR and grayscale visible image streams were simultaneously recorded and manually time-synchronized. While 238 subjects participated in the collection, [34] notes that there is only data for 105-112 subjects in the majority of scenarios.

Similar to [34], the KTFE dataset [25] elicited natural displays of emotion from 26 subjects through the use of video clips. Instrumental music was used between sequences to promote a neutral emotional state. Subjects were allowed to wear glasses during the collection. The data was captured simultaneously with an InfRec R300 camera.

The Carl dataset [9] contains time-lapse data of 41 subjects captured in four separate sessions spaced two days apart in which subjects were allowed uncontrolled natural variations in their expressions. The data was simultaneously recorded using a combined visible/LWIR camera and a separate NIR camera.

The Université Laval Face Motion and Time Lapse (ULFMT) database [12] contains 238 subjects recorded in multiple sequences under variable conditions, including significant time-lapse on the order of two to four years. Although the data was collected from the near, short, medium and long wave infrared bands, only the MWIR data has been released to date.

The ARL Multi-Modal Face Database (MMFD) dataset is composed of two separate collections, first presented in [15] and then extended in [40], both with simultaneously acquired visible, LWIR, and Polarimetric LWIR data. It has a combined total of 111 subjects. Unique to this dataset is the variable distances at which subjects are captured.

The Eurocom dataset[22], with 50 subjects captured using a combined visible/LWIR camera, notably contains a wide variety of acquisition scenarios, including sequences during which the eye and mouth regions are occluded by the subject's hand.

The RWTH-Aachen [20] dataset contains high resolution LWIR images of 94 subjects. Each subject is captured with variable expressions and head poses (both pitch and yaw), in controlled and uncontrolled sequences. The dataset is well annotated for emotions, discrete facial actions, and face landmarks. It cannot be used on its own for thermal-to-visible face recognition due to an absence of visible data, however it can still be employed to develop thermal landmark detection algorithms.

Table 1: Summary statistics of datasets containing MWIR or LWIR face data ordered (approximately) from least to most recent. Whether controlled or uncontrolled, the presence of the following variable conditions is noted: (P)ose, (I)llumination, (E)xpression, (T)ime-lapse, (G)lasses, and (O)cclusion. Image resolution is written as (w×h).

Dataset	Modalities	Subjects	Variability	IR Resolution	Range (m)
UND [5]	LWIR, RGB	241	I,E,T	320×240	Unspecified
IRIS [1]	LWIR, RGB	30	P,I,E	320×240	Unspecified
IRIS-M3 [4]	LWIR, RGB	82	I	320×240	1.2
Terravic [23]	LWIR	20	P,G	320×240	Unspecified
UH [3]	MWIR	138	P,E	640×512	Unspecified
NVIE [34]	LWIR, Mono	215	I,E,G	320×240	0.75
KTFE [25]	LWIR, RGB	26	E,G	320×240	0.85
Carl [9]	N/LWIR, RGB	41	I,E,T	160×120 (LW)	1.35
ULFMT [12]	MWIR, RGB	238	P,E,T,G	640×512	1.0
ARL-MMFD [15][40]	P-L/LWIR, RGB	111	E	640×480 (LW)	2.5, 5.0, 7.5
Eurocom [22]	LWIR, RGB	50	P,I,E,G,O	160×120	1.5
RWTH [20]	LWIR	94	P,E	1024×768	0.9
Tufts [26]	N/LWIR, RGB	100	P,E	336×256	1.5
ARL-VTF	LWIR, RGB, Mono	395	P,E,G	640×512	2.1

The Tufts Face Database [26] is a multi-modal dataset with several image acquisition devices and scenarios. The scenarios involve the simultaneous capture of visible and LWIR frontal images as well as visible, NIR, LWIR images acquired with a mobile, multi-camera sensor platform being rotated in front of the subject in an arc. In both scenarios, subjects were asked to pose with a variety of expressions and also sunglasses. Also included in the dataset are images from a 3D light-field camera, 3D point cloud reconstructed facial images, and computer-generated face sketches. The dataset contains 100 subjects.

Compared to the ARL-VTF dataset with 395 subjects, the next largest high-resolution thermal face dataset, ULFMT, contains 238 subjects and features MWIR and RGB video recordings under a comprehensive set of variable conditions but lacks synchronized data. For the RWTH dataset, although it utilized a higher resolution thermal camera and provides annotations for variable expressions, it contains no visible imagery counterpart. In contrast, ARL-VTF's synchronized acquisition and stereo arrangement supports algorithm development for 3D model learning [6], multi-modal fusion [18], domain adaptation [30], and cross-domain image synthesis [13]. Three such synthesis approaches [7][16][39] for thermal-to-visible face verification are showcased in Section 4.2.

In summary, the ARL-VTF dataset is the only dataset which has all of the following characteristics: a) time-synchronized visible and thermal imagery, b) data collected using a current commercially available uncooled LWIR camera, c) variable expression, pose, and eyewear, d) facial landmark annotations, and e) the largest number of subjects and images to-date.

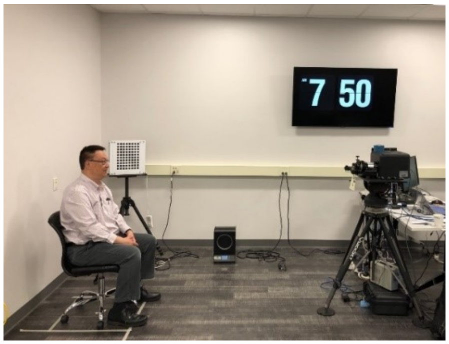


Figure 2: The collection area showing the sensor array as it collects the baseline (frontal) image sequence.

3. Database Collection

The data collection occurred over the course of 9 days in November 2019. The released dataset contains 395 subjects, each of whom completed an Institutional Review Board (IRB) approved consent form prior to image acquisition. The subjects were seated in front of a thermally neutral background 2.1 meters from the sensor array with their heads at approximately the same height as the sensors. Illumination was provided by the standard fixed overhead room lighting. The collection area setup is pictured in Figure 2.

Subjects' faces were recorded for approximately 10 seconds under each of the following conditions:

1. A *baseline* sequence of frontal images with the subject maintaining a neutral expression. If subjects were wearing glasses, they were asked to remove them.
2. An *expression* sequence of frontal images of the subject counting out loud incrementally starting from one.

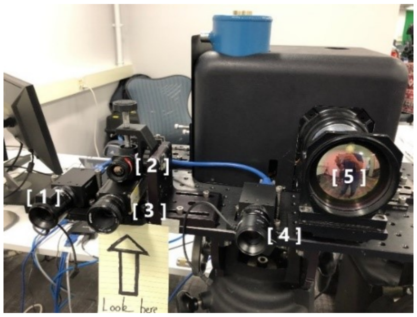


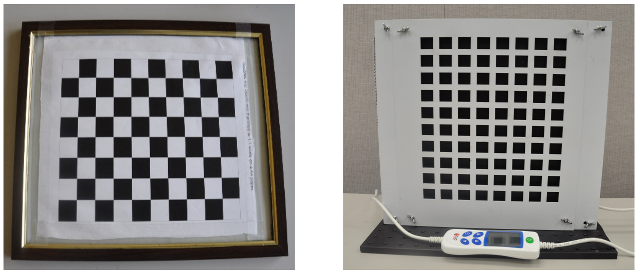
Figure 3: Sensor array with two FLIR Grasshopper3 cameras {**1, 4**}, the FLIR Boson LWIR sensor {**2**}, and the Basler Scout camera {**3**}. Polarimetric LWIR sensor {**5**} data not included.

3. A *pose* sequence of images where subjects were asked to slowly turn their heads from left to right. However, a small number of subjects rotated their entire bodies from left to right using the swiveling chair.
4. If subjects naturally wear *glasses* (removed for sequences 1-3), they were asked to put them back on for an additional sequence of *baseline* images.

Sensors: This dataset was collected with an array of three visible cameras and one LWIR thermal sensor. The visible imagery was recorded using two monochrome FLIR Grasshopper3 CMOS cameras and one RGB Basler Scout CCD camera. The LWIR data is captured by a FLIR Boson uncooled VOx microbolometer with a spectral band of 7.5 µm to 13.5 µm and thermal sensitivity of <50 mk. Table 2 lists the camera specifications. The sensors were mounted onto a single optical plate as shown in Figure 3. Data from a fifth sensor (a LWIR polarimeter) is omitted from this dataset as it was not time-synchronized with the other cameras.

Sensor Calibration and Synchronization: Sensor calibrations were conducted each day of the data collection to enable post-processing for 2D image registration and 3D geometric calibrations of the multiple visible and infrared sensors. An 8×10 checkerboard pattern with 20mm squares is mounted in front of a black body source which provides contrast for both visible and thermal images. For the thermal camera, a custom designed thermal/visible pattern using 20mm square holes with 10mm spacing was used. The visible and thermal sensor checkerboard calibration patterns are presented in Figure 4. In order to facilitate the development of 3d-based algorithms, the intrinsic and extrinsic camera parameters are provided with this dataset.

Using custom software to interface with each camera vendor's respective SDK software, the images were captured in a time-synchronized fashion via multithreaded software triggers at 15 frames per second due to bandwith limitations regarding data transfer.



(a) Visible Pattern (b) Thermal Pattern

Figure 4: Calibration patterns for the visible and thermal sensors.

3.1. Dataset Details and Usage

In total, the dataset contains 395 subjects and 549,712 images. To provide a sense of face resolution, the average inter-pupil distances (IPDs) of frontal baseline images are tabulated in Table 3. IPDs are calculated as the pixel distance between the left and right eye centers. To facilitate reproducibility and evaluation, the dataset is divided into subject-disjoint development (training and validation) and test sets with 295 subjects in the development set and the remaining 100 subjects in the test set. The subjects within the development set are sub-divided into training and validation sets using a 5-fold cross-validation scheme for hyper-parameter tuning and model selection. Of the 395 total subjects, 60 subjects were recorded both with and without glasses. These subjects have been evenly divided between the development and test sets, and proportionally divided between the training and validation sets (24 for training and 6 for validation).

3.1.1 Thermal-to-Visible Face Verification Protocols

We use the following grammar to describe the type of images in each gallery and probe set. In order to facilitate detailed analysis, the temporally-disjoint sets of gallery and probe images are defined in terms of a sequence category and an eyewear category. Gallery and Probe protocols are designated "**G**" and "**P**" respectively. "**V**" and "**T**" refer to the visible and thermal spectrum data. The sequence categories "**B**", "**E**", and "**P**" signify the baseline, expression, and pose sequences, respectively. The "$*$" symbol represents any or all sequence categories. For the purposes of the evaluation protocol, **B** also includes the glasses image sequence. There are three eyewear categories which describe if a subject possesses glasses and if the glasses are being worn in the image. Images of subjects who do not possess glasses use the tag **0**, whereas subjects who have their glasses removed or worn are notated **-** and **+**, respectively. The eyewear category is omitted when no filtering has been done on the basis of eyewear. In extended Backus–Naur

Table 2: Visible and LWIR camera information. The $\{\cdot\}$ enumeration corresponds to the camera labeling in Figure 3. The Mean (M) and Standard Deviation (SD) of inter-pupil distances (IPDs) are calculated using the baseline image sequence.

Camera	Modality	Resolution (w×h)	IPD	
			M	SD
FLIR Grasshopper3 $\{\mathbf{1, 4}\}$	Mono visible	2048×2048	89.3	6.6
FLIR Boson $\{\mathbf{2}\}$	LWIR $7.5 - 13.5\,\mu\text{m}$	640×512	45.2	3.3
Basler Scout $\{\mathbf{3}\}$	RGB color	658×492	66.7	5.0

form, the rules for producing descriptive protocol labels are:

$$\begin{aligned}
\langle\text{set}\rangle \quad &::= \quad \text{“}\mathbf{G}\text{”} \ \mid \ \text{“}\mathbf{P}\text{”};\\
\langle\text{modality}\rangle \quad &::= \quad \text{“}\mathbf{V}\text{”} \ \mid \ \text{“}\mathbf{T}\text{”};\\
\langle\text{sequence}\rangle \quad &::= \quad \text{“}\mathbf{B}\text{”} \ \mid \ \text{“}\mathbf{E}\text{”} \ \mid \ \text{“}\mathbf{P}\text{”} \ \mid \ \text{“}*\text{”};\\
\langle\text{eyewear}\rangle \quad &::= \quad \text{“}\mathbf{0}\text{”} \ \mid \ \text{“}\mathbf{-}\text{”} \ \mid \ \text{“}\mathbf{+}\text{”};\\
\langle\text{protocol}\rangle \quad &::= \quad \langle\text{set}\rangle, \text{“\_”}, \langle\text{modality}\rangle,\\
&\qquad\ \langle\text{sequence}\rangle, [\langle\text{eyewear}\rangle+];
\end{aligned}$$

Specific protocols have been developed for the evaluation of thermal-to-visible face verification algorithms. As the collection process yielded a different number of images for each subject, the test data has been selectively sampled to provide an equal number of images per subject and sequence. Additionally, specific images have been further designated as either probe or gallery images in order to standardize evaluation. Gallery images are composed solely of baseline images from the visible cameras. Probes are thermal images from all three sequences. Two distinct galleries are specified: 1) **G_VB0-** in which no subjects are wearing glasses, and 2) **G_VB0+** wherein glasses are worn by the subjects who have them.

The gallery and probe sets were constructed as follows. Seven evenly-spaced timestamps were selected from each subject's baseline sequence, starting from the first timestamp and ending with the last. The images from each of the three visible cameras corresponding to the first and last timestamp in the sequence are placed into **G_VB0-**. The images from the LWIR camera corresponding to the remaining five timestamps are designated as probes (**P_TB0-**). If a glasses sequence was recorded for that subject, then this process is repeated for the images in that sequence, with the resulting images becoming associated with the **G_VB0+** and **P_TB+** protocols. Next, 25 timestamps for the expression sequence are selected, spaced evenly to cover the span of the sequence. The images corresponding to those timestamps from all four cameras are added to the subject's set of probe images (**P_TE0-**). The same is done for the pose sequence (**P_TP0-**).

In summary, each subject has 6 gallery images (2 timestamps × 3 visible cameras) and 5 baseline probe images (5 timestamps × 1 thermal camera) without any eyewear. The subjects with glasses have an additional set of gallery and baseline probe images where the glasses are worn. This protocol can easily be extended to visible-to-visible or visible-to-thermal face verification by including the remaining images from the other cameras.

However, it should be noted that the development set has not been similarly balanced. All available images of a subject are by default included in the development set. Subsampling the development data is left to the user's discretion.

Annotations: Face bounding box and face landmark coordinates were generated using a commercial off-the-shelf face and landmark detector (Neurotechnology Verilook SDK) applied independently to the two high-resolution FLIR Grasshopper3 images assisted by manual supervision and correction of annotations. Face landmarks are in a 6-point annotation scheme corresponding to the left eye center, right eye center, base of nose, left mouth corner, right mouth corner, and center of mouth. The stereo arrangement of the Grasshopper3 cameras enabled the annotated points to be projected into the coordinate spaces of the LWIR and Scout RGB cameras using 3D geometry.

The stereo setup also allowed for the automatic estimation of head pose achieved using OpenCV's [2] implementation of the Perspective-n-Point with RANSAC algorithm. Figure 5 displays the distribution of estimated yaw angles captured during the pose sequence across all subjects. There is some slight asymmetry in the distribution about 0°, partially due to the fact that subjects oftentimes did not complete the full 180° head rotation. Metadata for each image includes the subject ID, camera, timestamp, image sequence, detected face bounding box, detected 6-point face landmarks, and estimated yaw angle.

Requesting the Database: Requests for the database can be made by contacting Matthew Thielke (matthew.d.thielke.civ@mail.mil). Requestors will be asked to sign a database release agreement and each request will be vetted for valid scientific research.

4. Performance Benchmarks

Benchmark results for landmark detection and thermal-to-visible face verification are provided in this section.

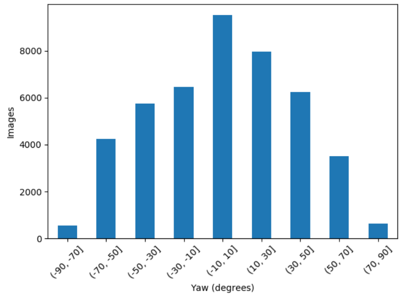


Figure 5: Distribution of head poses in terms of estimated yaw angles from the pose image sequence.

4.1. Face Landmark Detection

The Deep Alignment Network (DAN) [21] is a multi-stage convolutional neural network (CNN) designed to iteratively update the predicted landmark locations given an initial shape estimate. It has shown promising results for face landmark detection on both visible [37] and thermal [29][20] imagery. The model was trained with thermal face images from all of the recording sequences. The detected face bounding boxes are used to crop the images. The output of the model is the predicted face shape $\hat{\mathbf{y}} \in \mathbb{R}^{L\times 2}$, where L is the number of face landmark locations.

For these benchmarks, we set $L = 5$ and detect the left and right eye centers, the base of the nose, and the left and right mouth corners. Landmark detection performance is evaluated using the Normalized Root Mean Square Error (NRMSE),

$$E(\hat{\mathbf{y}}, \mathbf{y}^*) = \frac{1}{N}\sum_{i=1}^{N} \frac{\frac{1}{L}\sum_{j=1}^{L} \|\hat{\mathbf{y}}_{i,j} - \mathbf{y}_{i,j}\|_2}{\|tl(\mathbf{y}_i) - br(\mathbf{y}_i)\|_2}, \qquad (1)$$

where N is the number of samples in the test set and $\hat{\mathbf{y}}$ and $\mathbf{y}$ are the predicted and ground-truth landmark coordinates, respectively. The error is normalized by the Euclidean distance between the top left point, tl, and bottom right point, br, of the ground-truth shape's rectangular bounds. The face diagonal is used to normalize the error, rather than the IPD, as it is more stable in off-pose conditions [36]. As per [37], in addition to the mean and standard deviation (Std), the median, Median Absolute Deviation (MAD), and maximum NRMSE statistics are tabulated in Table 3. We set a threshold of 0.08 NRMSE for the Failure Rate and Area Under the Curve (AUC) of the Cumulative Error Distribution (CED).

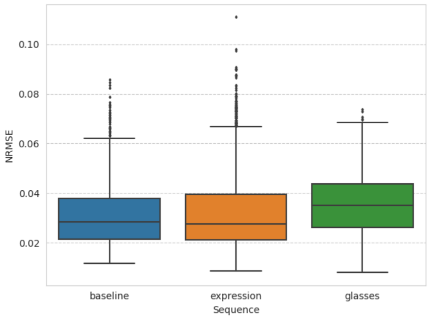


Figure 6: NRMSE of baseline, expression, and glasses sequences.

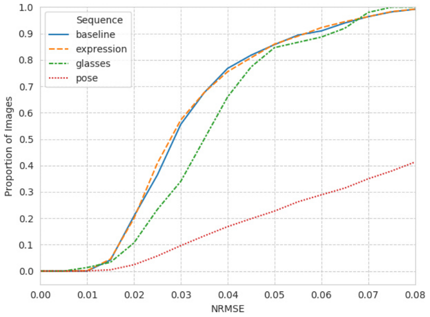


Figure 7: CED for the baseline, expression, glasses, and pose sequences.

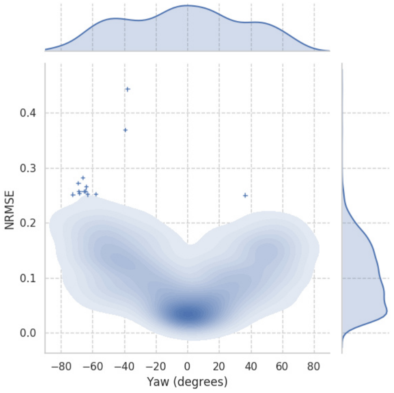


Figure 8: Bivariate distribution generated using Gaussian kernel density estimator of NRMSE across head yaw for the pose sequence. '$+$' indicates outliers with NRMSE ≥ 0.24.

As seen from Figures 6 and 7, the DAN achieves good performance on all frontal images, including images with expressions or glasses. The model fails on the head pose sequence, where performance significantly degrades with yaw angles beyond ±20°, as illustrated in Figure 8. Interestingly, while images with glasses have a slightly higher NRMSE on average compared to the other frontal images, they also have tighter performance bounds and a 0% Failure Rate, as shown in Figure 6. This may be due to the distinct visual cues granted by glasses (which absorb heat emissions and appear black in thermal images), or simply by virtue of a small sample size of subjects with glasses.

4.2. Thermal-to-Visible Face Verification

One domain-invariant feature learning approach and three thermal-to-visible synthesis approaches are benchmarked against the ARL-VTF dataset. The verification per-

Table 3: Landmark detection performance statistics in terms of the NRMSE.

Sequence	Mean	Std	Median	MAD	Max Error	$\text{AUC}_{0.08}$	Failure $\text{Rate}_{0.08}$
baseline	0.032581	0.015483	0.0283	0.0119	0.0857	0.5798	0.0080
expression	0.032445	0.015679	0.0276	0.0122	0.1109	0.5946	0.0076
glasses	0.036763	0.014687	0.0350	0.0114	0.0737	0.4649	0.0000
pose	0.101184	0.056227	0.0949	0.0472	0.4431	0.1692	0.5868

formance is measured by the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) metrics, as well as the True Accept Rate (TAR) at False Accept Rates (FAR) equaling 1% and 5%.

The first method matches thermal and visible face images by learning a domain adaptive feature extractor as proposed in [10]. This framework exhibits four main parts: (1) a truncated version of VGG16 and Resnet to extract common features, (2) a "Residual Spectral Transform" subnetwork that learns a mapping between the visible and thermal features, (3) a cross-domain identification loss to optimize task-level discrimination, and (4) a domain invariance loss which ensures domain unpredictability. The extracted probe and gallery image features are compared using the cosine similarity measure. The results reported in Figure 9 and Table 4 corresponding to this baseline were yielded by the VGG16 version of the framework. The images are preprocessed similarly to [31][15] (with bandpass filtering omitted) by first aligning images to a 5-point canonical coordinate scheme via similarity transformation and then loosely cropping the aligned face images to 360 × 280 pixels in order to provide enhanced contextual information.

The remaining three methods employ Generative Adversarial Networks (GANs) to learn a mapping from thermal face images to visible face images. Once the visible image is synthesized from the input probe thermal image, a pretrained VGG-Face model [27] is used to extract deep features (i.e. output from relu5_3 layer) from the synthesized visible probe image as well as the visible gallery image to perform thermal to visible face verification. The cosine similarity between the two feature vectors is calculated to produce the verification score. The inputs into these synthesis models are 128 × 128 face images cropped according to the annotated bounding boxes. Images from all four sequences are used to train the models. The following GAN-based methods are used for evaluation:

- Pix2Pix [16]: Conditioned on thermal images, Pix2Pix model synthesizes visible images using a U-net based architecture [16][32].
- GANVFS [39]: GANVFS uses identity loss and perceptual loss [17] to train a synthesis network.
- Self-attention based CycleGAN (SAGAN) [7]: A self-attention module [38] is adapted with CycleGAN [41] for thermal to visible synthesis.

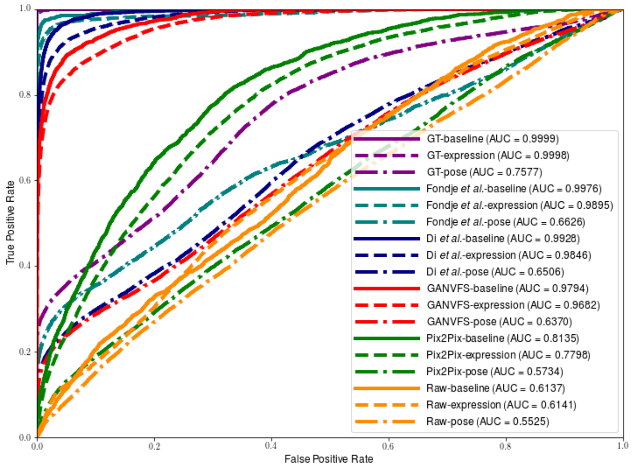


Figure 9: The ROC curves corresponding to the different methods for gallery **G_VB0-** and protocols **P_T*0-**.

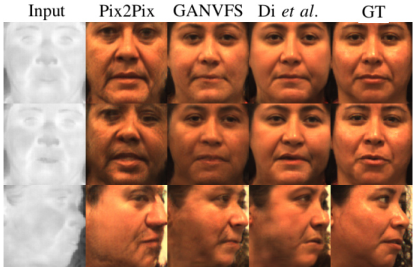


Figure 10: Sample synthesized images corresponding to different methods. First, second, and third rows correspond to baseline, expression, and profile faces.

Additionally, two baseline methods are established to gauge the performance of the GAN-based approaches. As a naive baseline method (labelled "Raw"), the thermal probes and visible gallery images are input directly to the VGG-Face model. In this scenario, no synthesis is performed on the thermal probes, nor is the VGG-Face model trained on the thermal data. As a ground-truth baseline method (labelled "GT"), the thermal probe images are replaced with the corresponding "ground-truth" visible images captured synchronously by the Basler Scout RGB camera.

The cross-modal face verification and synthesis results are shown in Figure 9 and Figure 10, respectively. As can be seen from Figure 9, simply extracting deep features

Table 4: Verification performance comparisons among the baseline methods, state-of-the-art methods for various settings.

		Gallery G_VB0-				Gallery G_VB0+			
Probes	Method	AUC	EER	FAR=1%	FAR=5%	AUC	EER	FAR=1%	FAR=5%
P_TB0	Raw	61.37	43.36	3.13	11.28	62.83	42.37	4.19	13.29
	Pix2Pix [16]	71.12	33.80	6.95	21.28	75.22	30.42	8.28	27.63
	GANVFS[39]	97.94	8.14	75.00	88.93	98.58	6.94	79.09	91.04
	Di *et al.* [7]	99.28	3.97	87.95	96.66	99.49	3.38	90.52	97.81
	Nimpa *et al.* [10]	99.76	2.30	96.84	98.43	99.87	1.84	97.29	98.80
P_VB0	GT Vis-to-Vis	99.99	0.23	99.79	99.95	99.99	0.24	99.86	100.00
P_TB-	Raw	61.14	41.64	2.77	16.11	57.61	44.73	1.38	6.11
	Pix2Pix[16]	68.77	38.02	6.69	20.28	52.11	48.88	2.22	4.66
	GANVFS[39]	99.36	3.77	84.88	97.66	87.34	18.66	7, 00	29.66
	Di *et al.* [7]	99.63	2.66	91.55	98.88	89.24	19.49	16.33	41.22
	Nimpa *et al.* [10]	99.83	1.95	96.00	99.48	99.03	4.79	85.56	95.86
P_VB-	GT Vis-to-Vis	100.00	0.00	100.00	100.00	99.06	4.33	89.66	96.22

from the raw images does not produce good verification results. This is mainly due to fact that both thermal and visible images have significantly different characteristics. The AUC corresponding to this method is only 61.37%. Pix2Pix which is a conditional GAN-based method provides slightly better results than the simple baseline of extracting features from raw data producing AUC of 71.12%. Both GANVFS and SAGAN methods are more advanced synthesis approaches and perform much better on this dataset, producing AUC of 97.94% and 99.28%, respectively. The Equal Error Rates (EER) of the Pix2Pix, GANVFS, and SAGAN models are 33.8%, 8.14%, and 3.97%, respectively. The synthesis results shown in Figure 10 are also consistent with the verification results shown in Figure 9 and Table 4.

In addition to the baseline comparisons, we analyze how different variations (baseline, expression, pose, eyewear) influence the cross-spectrum matching performance of different methods. As can be seen from Figure 10, expression slightly degrades the performance of the baseline methods. For instance, the AUC performance of SAGAN method reduces from 99.28% to 98.46%. We see similar degradation for GANVFS and Pix2Pix methods on expressive face images as well. From Figures 9 and 10, we can also see that pose affects the performance of different synthesis methods the most. The performance of the synthesis-based methods is constrained by the VGG-Face model's performance. This is evidenced by a reduction of the AUC from 99.99% in the baseline sequence to 75.76% for the pose sequence when using the ground-truth visible probe images as input. The EER of the Pix2Pix, GANVFS, and SAGAN models are 47.22%, 41.66%, and 40.24%, respectively. This experiment clearly shows that there is much that need to be done to deal with pose, expression and occlusion variations for cross-modal synthesis and verification. More advanced methods that specifically address these issues for heterogeneous face synthesis and verification are needed. A complete set of performance metrics for all the models, probe sets, and galleries are included in supplementary material.

5. Conclusion

A new, large-scale face dataset of time-synchronized visible and LWIR thermal imagery is presented. In order to emulate real-world conditions, variations of expressions, head pose, and eyeglasses have been systematically captured. Furthermore, the dataset is evaluated on the tasks of thermal face landmark detection and thermal-to-visible face verification using multiple state-of-the-art algorithms. Analysis of the results indicates two challenging scenarios. First, the performance of the thermal landmark detection and thermal-to-visible face verification models were severely degraded on off-pose images. Secondly, the thermal-to-visible face verification models encountered an additional challenge when a subject was wearing glasses in one image but not the other. This effect is further exacerbated in the thermal domain due to the occlusion induced by heat absorption in the lenses.

Acknowledgements: The authors would like to acknowledge sponsorship provided by the Defense Forensics & Biometrics Agency to conduct this research, and thank Michelle Giorgilli and Tom Cantwell for the discussions and their guidance. The authors would also like to thank Lars Ericson at IARPA and Chris Nardone, Marcia Patchan, and Stergios Papadakis at the JHU Applied Physics Laboratory for enabling ARL's participation in the 2019 IARPA ODIN data collection.

References

[1] Besma Abidi. IRIS Thermal/Visible Face Database, IEEE OTCBVS WS Series Bench. `http://vcipl-okstate.org/pbvs/bench/`, accessed 2020-06-09.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] Pradeep Buddharaju, Ioannis T. Pavlidis, Panagiotis Tsiamyrtzis, and Mike Bazakos. Physiology-based face recognition in the thermal infrared spectrum. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(4):613–626, apr 2007.

[4] Hong Chang, H Harishwaran, Mingzhong Yi, A Koschan, B Abidi, and M Abidi. An indoor and outdoor, multimodal, multispectral and multi-illuminant database for face recognition. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, volume 2006, 2006.

[5] Xin Chen, Patrick J. Flynn, and Kevin W. Bowyer. Visible-light and infrared face recognition. In *ACM Work. Multimodal User Authentication*, pages 48–55, 2003.

[6] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7093–7102, 2018.

[7] 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 *International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

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

[9] Virginia Espinosa-Duró, Marcos Faundez-Zanuy, and Jiří Mekyska. A New Face Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums. *Cognit. Comput.*, 5(1):119–135, mar 2013.

[10] Cedric Nimpa Fondje, Shuowen Hu, Nathaniel J Short, and Benjamin S Riggan. Cross-domain identification for thermal-to-visible face recognition. *arXiv preprint arXiv:2008.08473*, 2020.

[11] Travis Gault and Aly Farag. A Fully Automatic Method to Extract the Heart Rate from Thermal Video. In *2013 IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, pages 336–341, 2013.

[12] Reza Shoja Ghiass, Hakim Bendada, and Xavier Maldague. Université Laval Face Motion and Time-Lapse Video Database (UL-FMTV). Technical report, Université Laval, 2018.

[13] Ran He, Yi Li, Xiang Wu, Lingxiao Song, Zhenhua Chai, and Xiaolin Wei. Coupled adversarial learning for semi-supervised heterogeneous face recognition. *Pattern Recognition*, page 107618, 2020.

[14] Shuowen Hu, Nathaniel Short, Benjamin S. Riggan, Matthew Chasse, and M. Saquib Sarfraz. Heterogeneous Face Recognition: Recent Advances in Infrared-to-Visible Matching. In *2017 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG 2017)*, pages 883–890. IEEE, may 2017.

[15] 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 *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pages 119–126, 2016.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016.

[18] Ioannis A Kakadiaris, Georgios Passalis, Theoharis Theoharis, George Toderici, Ioannis Konstantinidis, and Najam Murtuza. Multimodal face recognition: Combination of geometry with physiological information. In *Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005*, volume II, pages 1022–1029. IEEE Computer Society, 2005.

[19] Brendan F. Klare and Anil K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(6):1410–1422, 2013.

[20] Marcin Kopaczka, Raphael Kolk, Justus Schock, Felix Burkhard, and Dorit Merhof. A Thermal Infrared Face Database with Facial Landmarks and Emotion Labels. *IEEE Trans. Instrum. Meas.*, 68(5):1389–1401, may 2019.

[21] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep Alignment Network: A Convolutional Neural Network for Robust Face Alignment. In *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, volume 2017-July, pages 88—-97, 2017.

[22] Khawla Mallat and Jean Luc Dugelay. A benchmark database of visible and thermal paired face images across multiple variations. In *2018 Int. Conf. Biometrics Spec. Interes. Group, BIOSIG 2018*. Institute of Electrical and Electronics Engineers Inc., oct 2018.

[23] Roland Miezianko. Terravic Facial IR Database. `http://vcipl-okstate.org/pbvs/bench/`, accessed 2020-06-09.

[24] Eslam Mostafa, Riad Hammoud, Asem Ali, and Aly Farag. Face recognition in low resolution thermal images. *Comput. Vis. Image Underst.*, 117(12):1689–1694, dec 2013.

[25] Hung Nguyen, Kazunori Kotani, Fan Chen, and Bac Le. A thermal facial emotion database and its analysis. In *Pacific-Rim Symp. Image Video Technol.*, volume 8333 LNCS, pages 397–408. Springer Verlag, oct 2013.

[26] Karen Panetta, Arash Samani, Xin Yuan, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, and Holly A. Taylor. A Comprehensive Database for Benchmarking Imaging Systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(3):509–520, mar 2020.

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

[28] Ioannis Pavlidis, James Levine, and Paulette Baukol. Thermal image analysis for anxiety detection. In *IEEE Int. Conf. Image Process.*, volume 2, pages 315–318, 2001.

[29] Domenick Poster, Shuowen Hu, Nasser Nasrabadi, and Benjamin Riggan. An Examination of Deep-Learning Based Landmark Detection Methods on Thermal Face Imagery. In *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, 2019.

[30] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 54–62, 2016.

[31] B. S. Riggan, N. J. Short, and S. 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.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[33] Panagiotis Tsiamyrtzis, Jonathan Dowdall, Dvijesh Shastri, Ioannis T Pavlidis, MG Frank, and P Ekman. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, 71(2):197–214, 2007.

[34] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. Multimed.*, 12(7):682–691, nov 2010.

[35] Zhi-Hao Wang, Gwo-Jiun Horng, Tz-Heng Hsu, Chao-Chun Chen, and Gwo-Jia Jong. A Novel Facial Thermal Feature Extraction Method for Non-Contact Healthcare System. *IEEE Access*, 8:86545 – 86553, may 2020.

[36] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019.

[37] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The Menpo Facial Landmark Localisation Challenge: A step towards the solution. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, pages 170—-179, 2017.

[38] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7354–7363, 2019.

[39] He Zhang, Vishal M. Patel, Benjamin S. Riggan, and Shuowen Hu. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 100–107. Institute of Electrical and Electronics Engineers Inc., jan 2017.

[40] He Zhang, Benjamin S. Riggan, Shuowen Hu, Nathaniel J. Short, and Vishal M. Patel. Synthesis of High-Quality Visible Faces from Polarimetric Thermal Faces using Generative Adversarial Networks. *Int. J. Comput. Vis.*, 127(6-7):845–862, jun 2019.

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.